

ABSTRACT

ANDERSEN, LINNEA KATHRYN. Machine Learning Approaches to Identify Genes Underlying Aquaculture Production Traits in Striped Bass and their Hybrid. (Under the direction of Dr. Benjamin J. Reading).

Aquaculture is the fastest growing sector of food production globally, however, the United States (US) is not among the top countries in production and is lagging far behind in global and national economic contribution. In fact, the US imports roughly 80.0 % of seafood products, which has led to a trade deficit of over \$17 billion USD. A major motivation of aquaculture research is to determine breeding and rearing techniques that ensure reliable (hardy, uniform size and quality) and sustainable (optimized resource use) aquaculture products, and this is a critical gap for researchers to address towards the expansion US and global aquaculture industries. Cellular mechanisms determinant of skeletal muscle growth is one such important area of research, as the muscle constitutes the major edible component of the fillet and therefore is the primary product of many aquaculture operations. Further, cohorts of fish (i.e., full- or half-sibling offspring) often show size variability such that there may be distinct groups of fish exhibiting superior growth phenotypes relative to others. A wide variation in growth performance presents a limitation in production, as these fish are less efficient in terms of resource use and subsequently decrease the amount of animal protein available for consumers as well as the potential of economic benefits that are a result of greater production yield.

Hybrid striped bass (HSB; female white bass, WB, *Morone chrysops* x *M. saxatilis*, striped bass, SB, male) are the fourth largest finfish aquaculture industry in the US (\$50 Million USD). The WB and SB crosses were first produced in the 1960s and these fish exhibited hybrid vigor, or heterosis, whereby HSB had superior tolerance to culture conditions and resulting phenotypes (growth) compared to that of both parental species. Decades of research have been

conducted on the parent SB and WB fish since and both are considered priority species for the United States Department of Agriculture (USDA) National Animal Genome Research Program (under the NRSP-8). Work completed through this program has led to the generation of numerous genomic data, including recently updated reference genomes and transcriptomes, for SB, WB, and their hybrid offspring. Domesticated lines for both SB and WB parental fish have now been established over several generations and these fish demonstrate greater affinity for culture conditions (e.g., stress tolerance, growth rate) compared to their wild counterparts. Further, the expansion of US aquaculture has led to an increased demand for a larger, white-fleshed marine fish that cannot be met by currently available commercial aquaculture species including HSB, and as such the establishment of a standalone SB industry is underway.

Despite the known differences in performance of the HSB cross, a thorough understanding of the genetic contribution of either parental species is lacking. Without this understanding, it is not possible to truly selectively breed for superior traits. Gaining an understanding of heritability would allow for a deeper understanding of heterosis in HSB as well as the parental contribution of SB sires in the purebred fish. To promote the expansion of the aquaculture industry, as well as provide novel scientific insight into both heterosis and applications of machine learning (ML), genome-wide transcriptomics analyses of HSB and SB white skeletal muscle tissue were conducted to reveal genes most determinant of growth phenotype in these fish. The HSB study described herein is the only study that has been conducted to date that explores gene expression at the allele-level in these fish and in the context of growth and parental strain. Both the HSB and SB studies revealed metabolic pathways that may underlie growth rate and subsequently identifies potential targets for future selective breeding and genomic targeting efforts.

© Copyright 2022 by Linnea Kathryn Andersen

All Rights Reserved

Machine Learning Approaches to Identify Genes Underlying Aquaculture Production Traits in
Striped Bass and their Hybrid

by
Linnea Kathryn Andersen

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Biology

Raleigh, North Carolina
2022

APPROVED BY:

Dr. Benjamin J. Reading
Committee Chair

Dr. Russell J. Borski

Dr. Jeffrey A. Buckel

Dr. John R. Godwin

DEDICATION

To my living and late grandmothers, Patricia and Joan (1939–2010), for their unwavering encouragement while I overcame having to attend primary through secondary school, and to the friends and teachers whose presence meant doing so was not as much of an imposition.

To my dear friend Patrick (1992–2019), for everything you were and continue to be.

BIOGRAPHY

Linnea Kathryn Andersen was born 8 August 1994 in Baltimore, Maryland to Joseph M. Andersen and Kathleen M. Mullen. Linnea graduated from the John Carroll Catholic School in 2012 and pursued a Bachelor of Arts degree in Neuroscience from Franklin and Marshall College in Lancaster, Pennsylvania, where she was able to conduct research under the advisement of Dr. Robert N. Jinks throughout her freshman year and first summer as a college student. Despite enjoying her time spent in the laboratory very much, it was over those summer months that Linnea realized the pursuit of neuroscience was limiting her ability to do something critical to her happiness: spend time outside. After a handful of what she would eventually realize were quite literally life-altering conversations with two of her professors, Drs. James E. Strick and Daniel R. Ardia, Linnea decided to take some time away from academics in search of *the* path, the one that would allow for her academic interests and personal passions to intersect. Linnea moved to Toms River, New Jersey and volunteered at an animal sanctuary run by Dr. Laura Pople, a close family friend and, conveniently for nineteen-year-old Linnea in search of purpose, holder of a doctorate in psychology. After her time caring for three-hundred plus cats and numerous dogs came to an end, Linnea went on with renewed gumption to the University of Maryland (UMD) in College Park to pursue her Bachelor of Science degree. At the suggestion of her transfer advisor, Linnea enrolled in ENST 314: Fisheries Management and Sustainability taught by Dr. Reginal M. Harrell. By the end of the first week of classes, Linnea's eyes had been opened to the diverse world of fishes and the importance of aquaculture in global food systems then and projected for the future. Linnea spent her next years at UMD in excellent, engaging courses, studied abroad to Copenhagen, Denmark, worked with wonderful (and fascinating) people at an environmental contracting company that predominantly held remediation contracts

with the United States Department of Defense, served as an intern at the United States Environmental Protection Agency Office of Environmental Education, worked nights at the UMD library, ran a handful of official half and full marathons and many more miles in training (her favorite run being an early morning 9.78 miles one-way down Route 1 from her college apartment to the Washington monument, much to the dislike of Joseph and Kathleen), and had many, many exciting adventures making friends and going around to places familiar and new. At some point with a few semesters as an undergraduate remaining, Linnea realized she would need legitimate aquaculture experience in order to pursue *the* path she had come to find. Thus, Linnea sought out Dr. L. Curry Woods III who accepted her on first as a volunteer and later an employee at the UMD Crane Aquaculture Facility. Learning from Dr. Woods, Dan Theisen, and (now) Dr. Tyler Frankel is among Linnea's favorite memories and most formative experiences as an undergraduate student. In a conversation about attending graduate school, Dr. Woods mentioned being aware of someone looking for graduate students and suggested Linnea speak with this person, who turned out to be none other than Dr. Benjamin J. Reading, Linnea's major professor. Linnea began her graduate studies at the Pamlico Aquaculture Field Laboratory in Aurora, North Carolina on 4 May 2017 and moved to Raleigh later that summer to begin classes that fall.

As a graduate student, Linnea has had the privilege to participate in the 2018 Borlaug Summer Institute on Global Food Security at Purdue University, serve as an Association for International Agricultural and Rural Development and Center for Environmental Farming Systems fellow, and receive the Coastal Conservation Association of North Carolina David and Ann Speaks Coastal Conservation Association Scholarship, among other award, honors, and professional service activities. Linnea's favorite part of graduate school has undoubtedly been her role as a mentor and instructor, guiding students as they identify what *the* path is to them.

ACKNOWLEDGMENTS

I would first like to thank Dr. L. Curry Woods III for connecting me to my advisor, Dr. Benjamin J. Reading, whom I have had a truly extraordinary experience learning from.

My deepest gratitude and appreciation are extended to Dr. Benjamin J. Reading for the myriad of opportunities he has provided me with to learn and to challenge myself throughout the duration of my time as his student. Thank you for the time and energy that you have invested in my training as a scientist and in the skills necessary to be effective in the many roles a person comes to hold outside of the laboratory walls. I could not properly thank you for your time and energy without also thanking your wonderful family, Heather, Aidan, Elodie, and Marigold, for the same. Thank you all for your warmth, humor, and authenticity.

Thank you to Drs. Russell Borski, Jeffrey Buckle, and John Godwin for serving on my advisory committee. I value your perspectives greatly and am grateful for the engaging and fruitful discussions we have had about research and topics far beyond it.

Thank you to Dr. Ronald G. Hodson and his wife Kay for being something of a North Star to me throughout the trials and tribulations of graduate school and life...for listening when I needed to talk, and putting me to chores when I did not.

Thank you to Robert W. Clark, Michael S. Hopper, and Dr. Andrew S. McGinty for the infinite list of things you have taught me, including the hobbies and interests I have developed as a result of conversations and/or experiences we have shared. The time I have spent in Aurora at the Pamlico Aquaculture Field Laboratory is among the best I have spent thus far. I extend my sincere thank you to the many others that have made the lab what it is, to James and Rhonda Stallings, and those that came before me, including Dr. Valerie N. Williams and the above.

I have found immense fulfillment in the work that I have done and activities I have participated in throughout my time at NC State University, and this is in no small part due to the many exceptional people I have worked with, learned from, and look forward to seeing again:

Thank you to the striped bass and hybrid striped bass producers, for being what “it” is all about, and to some key aquaculture community members, Dr. Jason Abernathy, Pete Anderson, Dr. David L. Berlinsky, Michael Frinsko, Dr. Adam S. Fuller, Steve Gabel, Randy Gray (vroom vroom!), Eric Herbst, Dr. Linas W. Kenter, Frank Lopez, Barry Nash, and Katie Mosher. To John Davis and Joseph Bursey for keeping fish swimming at the Raleigh aquaculture facilities.

Thank you to the coordinators of the U.S. Borlaug Summer Institute on Global Food Security at Purdue University, Dr. Gebisa Ejeta, Gary Burniske, & Pamela McClure for organizing one of my most formative experiences, forging both my passion for agriculture research and lifelong friendships with Brooke Blessington and Dr. Charles Hunt Walne.

Thank you to the Center for Environmental Farming Systems, and in particular Dr. Angel Cruz, Dallas Goodnight, and my fellow-fellows. Both the opportunity to participate in the graduate fellowship program and the welcoming community have been wonderful to be a part of.

Thank you to Dr. D. Andy Baltzegar for readily answering my numerous questions and offering invaluable guidance on the sequencing components of my project, for being a comedic foil to Ben, making the effort to keep “the band” together during the pandemic, and much more.

Thank you to Drs. Carlos Goller, Leigh Ann Samsa, Arnab Sengupta, and Melissa Srougi of Biotechnology (BIT) program fame, having the opportunity to learn from you as an instructor (and student too for Carlos and Leigh Ann), developing course materials, and learning to teach CRISPR content throughout the pandemic was truly an honor and I am very grateful for the support you have all provided, including simply knowing I could reach out for guidance.

Thank you to Dr. Vanessa Doriott Anderson for her candor, sense of humor, and work in coordinating training programs, such as Preparing the Professoriate, that I benefited from greatly.

Thank you to the incredible staff who time and time again came through with the answers that made whatever I was trying to accomplish at a given moment possible: Carrie Baum-Lane, Jiewei Dan, Freha Legoas, Susan Marschalk, Dawn Newkirk, and William (Trevor) Quick.

Thank you to the Buckle lab members past and present, Jeffery Merrell and Drs. Paul Rudershausen and Brendan Runde, for being welcoming and allowing my participation in a number of remarkable experiences off and on shore.

Thank you to the undergraduate and high school students that I have taught and/or mentored in the classroom, laboratory, and/or through the NCSU BIT SURE, CEFS ASPIRE, and CAALS 3D programs for their time, interest, and dedication. I extend additional thanks to a number of students who exceeded expectations and are a true pleasure to know: Kenneth Erickson, Fatma Kahn, Fara Marin, Connor Neagle, Claire Pelletier, Sasha Pereira, and Kate Pottle. I am endlessly impressed with the ways you all have and continue to grow, succeed, deepen and discover passions, and develop your voice as scientists and individuals.

Thank you to my graduate school contemporaries for their efforts in making our shared experience as graduate students quite enjoyable. Specifically, thank you to Emilee Briggs, Dr. Eugene Cheung, Nathaniel Curtis, Erin Ducharme, Riley Gallagher, Dr. Jonathan Giacomini, Dr. Laura Hamon, Dr. Elsita Kiekebusch, Cara Kowalchyk, William Lee, Dr. Nicole Lindor, Dr. Jamie Mankiewicz, Dr. Andrew Mauer, Dr. Sarah Rajab, and Dr. Grayson Walker.

Lastly, thank you to my family and friends that have been by my side long before I was pursuing a graduate degree. Your love and support is invaluable and unmatched and I am continuously grateful for all of you and all that you do.

TABLE OF CONTENTS

LIST OF TABLES	xi
LIST OF FIGURES	xvii
CHAPTER 1: THE STATUS OF STRIPED BASS, <i>MORONE SAXATILIS</i>, AS A COMMERCIALLY READY SPECIES FOR U.S. MARINE AQUACULTURE	
Abstract	1
Introduction.....	2
History of the striped bass fishery and aquaculture and current market opportunity	5
Striped bass fishery status	5
Early striped bass aquaculture	7
Current market opportunity.....	10
Striped bass culture methods and tools	13
Domestication	14
Reproduction and larviculture.....	16
Rearing and growout.....	18
Genomic resources and tools	19
Future directions and challenges.....	20
Barriers and Opportunities.....	21
Establishing the <i>StriperHub</i>	23
Acknowledgements.....	24
References	26
Figures.....	48
CHAPTER 2: INSIGHTS INTO HETEROSIS OF HYBRID STRIPED BASS VIA MOTIF FINGERPRINTING	
Abstract.....	51
Introduction.....	52
Materials and Methods.....	54
Experimental Animals	54
Microsatellite Genotyping and Parentage Assignment.....	56
Muscle Histology	57
Statistical Comparisons.....	58
RNA-Sequencing of Hybrid Striped Bass White Muscle Tissue	59
Motif Fingerprint (MF) Discovery.....	60
Machine Learning Analysis	60
Mapping of Motif-Fingerprints to Genes.....	64
Allele-Level Resolution of Motif Fingerprints (Presence/Absence)	64
Pathway Analysis.....	65
Results.....	66
Morphometric Comparisons	66
Muscle Histology	68
Machine Learning Analysis	68
Motif-Fingerprint Mapping and Allele-Level Resolution	69
Pathway Analysis.....	71
Discussion.....	74

Conclusions.....	84
Acknowledgements.....	87
References.....	89
Tables.....	99
Figures.....	107

CHAPTER 3: TRANSCRIPTOMICS ANALYSIS OF STRIPED BASS SKELETAL MUSCLE REVEALS DISTINCT METABOLIC DIFFERENCES BETWEEN FISH EXHIBITING INFERIOR AND SUPERIOR GROWTH

Abstract.....	133
Introduction.....	134
Materials and Methods.....	136
Experimental Animals and Tissue Collection.....	136
Microsatellite Genotyping and Parentage Assignment.....	141
Muscle Histology.....	142
RNA-Sequencing and Quantitative Analysis of Gene Expression.....	143
Machine Learning Analysis.....	145
Pathway Analysis.....	148
Results.....	150
Experimental Animals Growth, Parentage, and Morphometric Comparisons.....	150
Muscle Histology.....	152
RNA-Sequencing and Quantitative Analysis of Gene Expression.....	153
Striped Bass Gene Expression and Growth.....	154
Striped Bass Gene Expression and Growth to Market Size.....	155
Striped Bass Gene Expression and Dam Parentage.....	156
Striped Bass Gene Expression and Sire Parentage.....	157
Striped Bass Gene Expression and Sire Size.....	158
Discussion.....	160
Acknowledgements.....	167
References.....	168
Tables.....	177
Figures.....	183

CHAPTER 4: MACHINE LEARNING WORKFLOWS FOR BIOLOGICAL DATA

Abstract.....	205
Introduction.....	206
Traditional Approaches to Data Analysis.....	208
An Overview of Machine Learning.....	213
How Learning Works and Cross-Validation Techniques.....	219
Evaluating Learning: Measures of Model Performance.....	221
Machine Learning Workflow.....	224
Specific Workflow Elements.....	224
Workflow Instructions.....	227
Examples of Workflow Applications.....	240
Machine Learning Considerations and Challenges.....	242
Conclusions.....	247

Acknowledgements.....	248
References.....	249
Tables.....	258
Figures.....	263

CHAPTER 5: IMPLICATIONS TO RESEARCH OF STRIPED BASS AND THEIR HYBRID

Research Directions: Metabolism and Muscle Growth in Fishes.....	268
References.....	276
Figure	280

APPENDICES

Appendix A.....	282
Appendix B.....	300
Appendix C.....	317
Appendix D.....	326
Appendix E.....	329
Appendix F.....	334
Appendix G.....	338
Appendix H.....	348

LIST OF TABLES

CHAPTER 2

Table 2.1	Descriptions of comparisons made between hybrid striped bass (HSB). Each comparison is generally referred to as the name listed in the left-most column. Strain refers to the geographic location of origin of the paternal fish (sire). The domestic sires were two years of age and bred at the North Carolina State University Pamlico Aquaculture Field Laboratory (PAFL) in Aurora, NC, USA. Other sires were three years of age and strain is based upon the waters of the states indicated. The sub-population size “(n=#)” of each group for a given comparison, referred to as “classes” in machine learning (ML) analyses, are also provided. The number prior to the forward slash is the number of the seventy-two (72) sacrificed fish, the number following the forward slash is the number of the forty (40) HSB of the sacrificed seventy-two for which sequencing data were generated. A brief description of each group is in the final column.....	99
Table 2.2.	Outcomes of negative control machine learning (ML) analyses of hybrid striped bass (HSB) expression data examined through two comparisons of growth performance at two months of age (Grade) and final harvest at fifteen months of age (Growth), and geographic origin of paternal parent, or sire strain (Strain). Each negative control was conducted by processing the optimal dataset* with randomized labels (group identifiers for a given comparison) associated with expression data ten separate times with each of the eight cross-validated algorithms (four algorithms x two cross-validated strategies). True learning by algorithms is said to occur if the mean percent correct classification of each negative control run is approximately what can be yielded through random chance. The number of attributes included in the optimal dataset and classes (groups) for each comparison are indicated followed by the percent correct classification that can be predicted under the assumption of random probability based on the number of classes (possible outcomes). The grand mean \pm standard deviation (SD) percent correct classification of all cross-validated ML algorithms on randomized optimal datasets for a given comparison is reported in the right-most column.	100
Table 2.3.	Mean \pm standard deviation weight (g) and total length (TL, mm) of a study population of aquacultured hybrid striped bass (HSB) and subpopulations thereof grouped for additional analyses and comparisons. The “Grade” comparison describes Top Grade (TG) and Runt HSB who at two months of age were sorted by size if expected to reach or exceed market size of ~680 g (1.5 lbs) by final harvest or not, respectively. HSB were grouped by weight at final harvest (fifteen months) as Large (LG) if above the mean TG weight, Small (SM) if below the mean Runt weight, and Intermediate (IN) if between the two means for the “Growth” comparison.	

Group sizes are provided in parentheses following the weight values for each comparison and subpopulation. Column “All (N=752)” represents the entire study population, “Sampled (n=72)” are HSB of this population sacrificed for white muscle tissue analyses, specifically, gene sequencing (“Sequenced (n=40)”) and histology (“Histology (n=18)”). Student’s *t*-Test (Grade) or one-way ANOVA and Tukey’s HSD post-hoc test (Growth, see: differentiating letters) were used to determine statistically significant differences between groups, the greatest p-value calculated for a comparison (i.e., between pairs) is provided in italic or as four asterisks (****) if $p \leq 0.0001$ 101

Table 2.4 Representation of striped bass (SB) sires among hybrid striped bass (HSB) offspring (N=72). The strain, or geographic location of sire origin is as follows: “DOM” are two year old, 5th generation domestic SB from the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA; “FL”, “SC”, “TX”, and “VA” refer to the three year old SB produced in hatcheries from SB caught in the waters of Florida, South Carolina, Texas, and Virginia, respectively; “N/A” represents the unknown sire of a HSB offspring that could not be matched via microsatellite genotyping. The number of sires represented among sacrificed HSB is listed in the “No.” column followed by the average \pm standard deviation sire weight (kg); DOM and VA sires were the only groups to significantly differ in weight (one-way ANOVA, Tukey’s HSD, $p=0.0061$) indicated by a double asterisk (**). The number of sacrificed HSB offspring sires of each strain produced is provided in the Progeny column and expressed as a percentage of the subpopulation in parentheses. The number of these sacrificed HSB that were included in sequencing analysis (n=40 of 72) are included in italic. The offspring that fell into Top Grade (TG) and Runt (R) groups for the Grade* comparison as well as Large (LG), Intermediate (IN), and Small (SM) groups for the Growth* comparison are provided in the rightmost columns. 102

Table 2.5 Mean \pm standard deviation weight (g) and total length (TL, mm) of hybrid striped bass (HSB, N=72) produced from different strains (i.e., geographic location of origin) of SB sires. Specifically, “DOM” are two-year-old, 5th generation domestic SB from the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA; “FL”, “SC”, “TX”, and “VA” refer to the three year old SB produced in hatcheries from SB caught in the waters of Florida, South Carolina, Texas, and Virginia, respectively. At two months of age HSB were sorted by size into Top Grade (TG) or Runt groups based upon expectation to reach or exceed market size of ~680 g (1.5 lbs) by final harvest or not, respectively. The number of HSB

(of 72 total) produced from sires of each strain is provided as “N=#” followed by the number of offspring belonging to the TG and Runt groups presented as “(# TG/#Runt)”. Grand means of weight and TL for all offspring of a given strain group are in boldface and means for TG and Runt HSB of a specific strain group are listed below grand means. Differentiating letters between grand means indicate significant differences between strain groups (one-way ANOVA and Tukey’s HSD post hoc test for weight: DOM vs SC p=0.0070, DOM vs TX p=0.0063, DOM vs VA p=0.0266 and TL: DOM vs SC p=0.0240). The p-values for Student’s *t*-Tests comparing weight and TL between TG and Runt fish of each strain are provided below the respective values in italic font. One HSB of the sample population described here is not represented as parentage could not be assigned, this fish weighed 441.00 g, was 310.00 mm in TL and had been in the Runt group..... 103

Table 2.6 Histological analysis of hybrid striped bass (HSB, N=18) white muscle tissue fibers between groups of projected growth performance (Grade). HSB were graded at two months of age into two groups, Top Grade and Runts, based on anticipated growth to meet or exceed market size (~680 g, or 1.5 lbs) or not, respectively, by final harvest at fifteen months of age. ImageJ software (Fiji, v.1.52a, National Institutes of Health, NIH, Bethesda, USA) was used to count and determine the area (μm^2) of all fibers entirely within the field of view (i.e., not marginated) of each image collected in triplicate for all individuals to evaluate hyperplastic (fiber amount) and hypertrophic (fiber area) muscle growth. The diameter of each scored fiber was calculated as a geometric derivative of its area to compute the average fiber diameter and the frequency of scored fibers of a certain diameter range for each slide image. Triplicate images were scored duplicate and the average \pm standard deviation values here are grand means of the average number and diameter calculated for each individual (i.e., slide) belonging to a given group by both scorers. A Student’s *t*-Test was used to compare values between groups and an asterisk is beside p-values indicating a statistically significant difference ($\alpha=0.05$). 104

Table 2.7 Official symbols and names of human (*Homo sapiens*) orthologs to the thirty-three (33) genes identified as those encoding the striped bass (*Morone saxatilis*, SB) and white bass (*M. chrysops*, WB) proteins mapped to from unique, twelve amino acid-long motif fingerprints (MFs) identified as informative to the differentiation of fish by growth performance (Grade and Growth comparisons) and/or paternal geographic location of origin (Strain comparison) via application of a machine learning (ML) workflow. The ML workflow reduced dimensionality of a 15,000 MFs dataset by identifying the MFs that yielded optimum classification performance (i.e., most individuals, or instances, correctly assigned comparison groups, or classes) by four distinct, cross-validated ML algorithms. There were 821 unique MFs among the 500 most-informative for each comparison. The

number of MFs mapping to each gene is provided in the MFs column and the percentage of all 821 MFs is provided in parentheses. MFs were concatenated into longer amino acid sequences and mapped to complete proteins in both the translated SB and WB genome assemblies (“Both” column), only the SB assembly (“SB” column), only the WB assembly (“WB” column), or were not able to be mapped completely (i.e., a 12 of 12 amino acid alignment) to either and are therefore considered undetermined (“Undetm.” column). MFs counted as “Both” represent those that either completely aligned to both the SB and WB assemblies..... 105

CHAPTER 3

Table 3.1 Descriptions of comparisons made between striped bass (SB, n=72) offspring. Each comparison is generally referred to as the name listed in the left-most column. The sub-population size “(n=#)” of each group for a given comparison, referred to as “classes” in machine learning (ML) analyses, are also provided along with a brief description of each group..... 177

Table 3.2 Outcomes of negative control machine learning (ML) analyses of striped bass (SB) gene expression data examined through five comparisons of growth performance (Growth, Market Size) and parentage (Dam, Sire, Sire Size). Each negative control was conducted by processing the optimal dataset* with randomized labels (group identifiers for a given comparison) associated with expression data ten separate times with each of the eight cross-validated algorithms (four algorithms x two cross-validated strategies). True learning by algorithms is said to occur if the mean percent correct classification of each negative control run is approximately what can be yielded through random chance. The number of attributes included in the optimal dataset and classes (groups) for each comparison are indicated followed by the percent correct classification that can be predicted under the assumption of random probability based on the number of classes (possible outcomes). The grand mean \pm standard deviation (SD) percent correct classification of all cross-validated ML algorithms on randomized optimal datasets for a given comparison is reported in the right-most column. 178

Table 3.3 Summary of striped bass (SB) growth and feeding data collected over the course of one year starting at five months of age. SB were raised in triplicate indoor recirculating aquaculture system (RAS) tanks (908 L/tank from four to eight months of age and 2006 L/tank thereafter). The rightmost column indicates the age of SB at each sampling point followed by the total population size (N) and mean \pm standard deviation (SD) weight (g) and total length (TL, mm). The amount of feed (g) offered (sum offered to all tanks over a given duration) for a given period between sampling and the mean \pm SD feed conversion ratio (FCR) are also provided. These values specific to each replicate tank are provided in

italics. Differentiating letters assigned to weight and/or TL values for a given sampling point indicate statistical differences between replicate tanks were identified via one-way ANOVA and Tukey’s HSD (alpha=0.05). The absence of letters indicates no significant differences were identified for a given measurement and timepoint. 179

Table 3.4 The mean \pm standard deviation (SD) weight (kg) and total length (TL, mm) of the sub-population of striped bass (SB) sacrificed for analysis (n=72 of 173). An equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (1.36 kg, or 3.0 lbs; n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) in subsequent analyses and the average weight and TL of those sub-groups are provided here. A Student’s *t*-Test was used to compare weight and length between Superior and Inferior groups for the “Growth” comparison described in the text and the one-way ANOVA and Tukey’s HSD tests were used to compare the four groups in the “Market Size” comparison described in the text. The p-value for all comparisons of weight and TL was p<0.0001 as indicated in the table below. Differentiating letters are used to indicate differences of weight and/or total length identified between groups in the Growth comparison and groups in the Market Size comparison. 180

Table 3.5 Weight (kg) and total length (TL, mm) of the striped bass (SB) offspring sacrificed for analysis (n=72) and grouped according to parentage: SB Dam, Sire Size, and Dam x Sire Size. Briefly, two female SB (dams) were crossed with twelve male SB (sires) that were either “Large” or “Small” in size (“Sire Size” comparison), whereby six males total (three from either size group) were crossed with a single female to produce twelve half-sibling families. The number of SB produced from either Dam, Large or Small sires, or specific Dam x Sire Size is listed in the table below (“Offspring”) followed by the mean \pm standard deviation (SD) weight and TL of SB offspring belonging to a given size group. Weight and TL metrics were compared between groups via Student’s *t*-Test (Dam, Sire Size) or one-way ANOVA and Tukey’s HSD post hoc test (Dam x Sire Size), p-values are listed below for each comparison and metric. Differentiating letters are used to indicate differences between weight of offspring in the Dam x Sire Size comparison..... 181

Table 3.6 The number of gene transcripts of the 32,018 measured that were identified as important to the comparisons of striped bass (SB) listed below based upon information gain ("Entropy") determined through the machine

learning (ML) approach or significant difference determined through traditional statistical approaches (false discovery rate p-value, “ $p \leq 0.05$ ”). The gene transcripts measured map back to 22,746 unique genes; the number of unique genes each set of transcripts map back to is provided in parentheses immediately following each value. The numbers provided in the “Optimal” row for each comparison are those gene transcripts and unique genes that were considered the optimal dataset* for each comparison. The number of transcripts and unique genes that were identified as important based upon information gain and significantly differed are listed in the “Shared” column, the values in this column associated with the “Optimal” row is this number of transcripts and genes that are also within the optimal dataset*. The final column, “Total” provides the total number of unique transcripts and genes identified as important to a given comparison based upon information gain and statistical tests..... 182

CHAPTER 4

Table 4.1	Definitions of common machine learning (ML) terminology. Comparable terms used to describe similar concepts in traditional statistics and/or biological research are provided in italics, if applicable. Defined terminology used in another provided definition are <u>underlined</u>	258
Table 4.2	Examples of machine learning (ML) algorithms within the four broad ML categories: (1) Traditional (Classical) Learning, (2) Reinforcement Learning, (3) Neural Nets and Deep Learning, (4) Ensemble Learning, and subtype(s) thereof as applicable.	259
Table 4.3	Kappa statistic (left) and Area Under the Receiver Operating Characteristic Curve (AUROC) values (right) and corresponding prediction strength of machine learning (ML) algorithms. The Kappa statistic is a measure of precision, specifically “interrater reliability”, or agreement between raters. The AUROC measures algorithm performance by characterizing the relationship between correct outcomes (true positives and true negatives) and incorrect outcomes (false positives and false negatives). Prediction strength varies from “worse than random” to “Optimal”, where “worse than random” indicates accuracy of predictions made was worse than what could be expected based on random chance and “optimal” is a perfectly trained algorithm.	261
Table 4.4	The percent correct classification of machine learning (ML) algorithms that can be predicted with the assumption of random probability based on the number of designated classes (groups) in a given dataset.....	262

LIST OF FIGURES

CHAPTER 1

- Figure 1.1 Commercial and recreational landings of Atlantic striped bass since the 1950s (a). Bars indicate the average landings per 2 years from 1950 to 2018. Spawning stock biomass (SSB) of Atlantic striped bass from the 1980s to 2016 (b). Hybrid striped bass production in the United States as reported since industry inception beginning in 1986 (c). Years that aquaculture production volumes were reported in the USDA Agriculture Census are indicated as black bars. Asterisks (*) on all panels mark the year the U.S. hybrid striped bass industry began production (1986). Data for these figures were provided by the Atlantic States Marine Fisheries Council (ASMFC) and National Oceanic and Atmospheric Administration (NOAA) National Marine Fisheries Service (NMFS) (a and b), and from Dr. James Carlberg (Kent SeaTech), Dr. Marc Turano (NC SeaGrant), Dr. Anita Kelly (University of Arkansas at Pine Bluff and Auburn University), and the USDA Agriculture Census (c). 48
- Figure 1.2 Domestic striped bass broodstock performance data collected for different age classes and generations (collected between March and June of each year, 2005–2020): Year 1 (45–60 weeks of age), Year 2 (80–104 weeks of age), Year 3 (136–154 weeks of age), and Year 4 (197–209 weeks of age). The x-axis is the age class grand mean and the y-axis is weight (g). The filial generation of captive breeding is indicated for the periods of 2004–2007 (F3), 2008–2011 (F4), 2012–2015 (F5), and 2016–2019 (F6). Gray shading indicates the target striped bass market size at between 1.36 and 2.27 kg (3.0 and 5.0 lb). Each datapoint for F3, F4, F5, and F6 represents a grand mean value of 3 or 4 different age class cohorts and hundreds of fish were measured for each age class per annum with the exceptions of F3 age class 1 (a single cohort), F6 age class 2 (2 cohorts), F6 age class 3 (2 cohorts), and F6 age class 4 (a single cohort). 49
- Figure 1.3 Domestic striped bass age class performance data (collected between March and June of each year, 2005–2020): (a) Year 1 (45–60 weeks of age) and (b) Year 2 (80–104 weeks of age). The filial generation of captive breeding is indicated for the periods of 2004–2007 (F3), 2008–2011 (F4), 2012–2015 (F5), and 2016–2019 (F6). Bars indicate grand mean values and error bars indicate standard deviation where there were three or four age class observations (annual performance data of hundreds of fish) for each filial generation; error bars are omitted where there were only one or two annual observations indicated as N=1 or N=2. The dashed line indicates target market size for striped bass at 1.36 kg (3.0 lb). 50

CHAPTER 2

Figure 2.1 Frequency distribution of hybrid striped bass (HSB) weight (g) measured from **(A)** the complete study population (N=752 HSB) and **(B)** the individuals sacrificed for additional analyses (n=72). HSB were produced and raised at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA and graded at two months of age into two groups, Top Grade and Runts, based on anticipated growth to reach or exceed market size for these fish (~680 g), or not, respectively. The weights shown were recorded at final harvest (fifteen months of age). The mean \pm standard deviation weight of the entire study population was 525.84 ± 145.36 g (sample population was 530.90 ± 140.19 g). Top Grade HSB from the study population (n=377) weighed 633.45 ± 107.35 g on average which significantly differed from the Runt HSB (n=375, 417.95 ± 86.65 g) and this difference was similarly observed between those in the sample population (Top Grade, n=36, 633.22 ± 106.12 g; Runt, n=36, 428.58 ± 84.10 g) (Student's *t*-Test, $p < 0.0001$). 107

Figure 2.2 Frequency distribution of hybrid striped bass (HSB) total length (TL, mm) measured from **(A)** the complete study population (N=752 HSB) and **(B)** the individuals sacrificed for additional analyses (n=72). HSB were produced and raised at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA and graded at two months of age into two groups, Top Grade and Runts, based on anticipated growth to reach or exceed market size for these fish (~680 g), or not, respectively. The TLs shown were recorded at final harvest (fifteen months of age). The mean \pm standard deviation TL of the entire study population was 324.20 ± 28.93 mm (sample population was 323.57 ± 28.00 mm). Top Grade HSB from the study population (n=377) were 346.73 ± 17.11 mm on average which significantly differed from the Runt HSB (n=375, 301.62 ± 19.072 mm) and this difference was similarly observed between those in the sample population (Top Grade, n=36, 345.33 ± 16.20 mm; Runt, n=36, 301.81 ± 18.80 mm) (Student's *t*-Test, $p < 0.0001$). 108

Figure 2.3 The average weight of hybrid striped bass (HSB) offspring produced using male striped bass (SB) of different geographic origin (strain). The "DOM" strain represents domestic SB males raised in captivity at the North Carolina State University's Pamlico Aquaculture Field Lab (NCSU PAFL) in Aurora, NC, USA. The remaining strains were produced in hatcheries from wild-caught SB from the waters of Florida (FL), South Carolina (SC), Texas (TX), and Virginia (VA). A one-way ANOVA test was used to determine if the weight of HSB produced from sires of different strains significantly differed ($p < 0.0001$). The value within each bar towards the x-axis represents the number of HSB in each group. Specific differences between the weights of offspring grouped by sire strain were identified via

Tukey's HSD post-hoc test, the results of which are indicated in the figure by differentiating letters and written here: The mean \pm standard deviation (SD) weight for HSB produced from DOM sires was 618.05 ± 145.99 g and did not significantly differ from the FL offspring (weight was 656.25 ± 71.01 g). DOM offspring were significantly greater in weight than HSB produced by SC sires (offspring weight was 456.64 ± 128.48 g; $p=0.0070$), TX sires (offspring weight was 470.20 ± 115.24 g; $p=0.0063$), and VA sires (offspring weight was 499.26 ± 107.70 g; $p=0.0266$)..... 109

Figure 2.4 A comparison of the average weight of hybrid striped bass (HSB) offspring produced using male striped bass (SB) of different geographic origin (sires) and belonging to one of two grade groups: Top Grade (TG) or Runts (R). HSB were graded at two months of age as projected to grow to or exceed market size (~680 g, or 1.5 lbs) by final harvest (TG fish) or not (R fish). Grade groups were reared separately through final harvest at fifteen months of age. The "DOM" strain represents domestic SB males raised in captivity at the North Carolina State University's Pamlico Aquaculture Field Lab (NCSU PAFL) in Aurora, NC, USA. The remaining strains were produced in hatcheries from wild-caught SB from the waters of Florida (FL), South Carolina (SC), Texas (TX), and Virginia (VA). Student's *t*-Tests were completed to compare the average weight of the hybrids in the TG and R HSB within each strain group. The value within each bar towards the x-axis represents the number of HSB in a given group. Significance is denoted in the figure as follows: (*) indicates $p<0.05$; (***) indicates $p<0.001$; (****) indicates $p<0.0001$. The mean \pm standard deviation weight (g) of the DOM TG offspring ($n=16$, 684.19 ± 108.65 g) were significantly larger than the DOM Runt HSB ($n=6$, 441.67 ± 52.68 g) ($p<0.0001$). The same trend was observed between SC TG offspring ($n=3$, 589.33 ± 136.66 g) and SC Runt HSB ($n=8$, 406.88 ± 88.71 g) offspring ($p=0.0261$), and VA TG ($n=9$, 580.67 ± 74.28 g) and VA Runt ($n=10$, 426.00 ± 75.51 g) offspring ($p=0.0003$). No significant difference was observed between TX TG offspring ($n=4$, 557.50 ± 87.90 g) and TX Runt HSB ($n=11$, 438.45 ± 110.087 g) offspring ($p=0.0751$). A comparison could not be completed for FL offspring ($n=4$), as all HSB were in the TG group (656.25 ± 71.014 g)..... 110

Figure 2.5 A flow-chart of the machine learning (ML) analysis used to reduce data dimensionality and identify the unique, 12 amino acid-long motif-fingerprint (MF) sequences that are differentially expressed between hybrid striped bass (HSB) in groups corresponding to three comparisons: (1) Grade: HSB were graded at two months of age based upon projected growth by the time of harvest where Top Grade (TG) fish were anticipated to reach or exceed market size and Runts not; (2) Strain: HSB produced from sires of five different geographic origins, Domestic (DOM) sires were produced and reared at the North Carolina State University Pamlico Aquaculture Field Lab (NCSU PAFL, Aurora, NC), and sires from Texas

(TX), South Carolina (SC), Virginia (VA), and Florida (FL) were of hatchery origin and caught in the surrounding respective waters; and (3) Growth: HSB were categorized at harvest based upon realized growth performance as Large (LG), HSB that exceeded the average weight of fish in the TG group (>633.45 g), Small (SM), HSB that did not reach the average weight of the fish in the Runts group (<417.95 g), and Intermediate (IN), HSB whose weight at the time of harvest fell between these two means. (A) The reduction of 15,000 MFs (attributes) identified among translated HSB sequencing data via Shannon's Entropy whereby MFs were assigned a weight corresponding to the amount of information it provides for decision making. The numbers under each comparison and groups (classes) thereof are the number of 15,000 MFs that were assigned an information gain value above 0.00 and therefore provide information to a given comparison. The MFs assigned information gain values above 0.00 were further processed by four ML algorithms each twice cross-validated (66.0 % split holdout method and 20-fold cross). Fewer and fewer top-ranking (i.e., highest information gain value) MFs were included in the input to identify points of model improvement and degradation with the inclusion or exclusion of MFs. (B) A Venn diagram of the shared and unique MFs among those ranked for each of the three comparisons. (C) Plots of ML model performance measured as percent correct classification of HSB into respective groups based upon included MF expression data averaged between cross-validation strategies for each of the four algorithms. Plots were used to identify points at which models were built on too many attributes (overfitting) or too few attributes (underfitting) based upon changes to performance and the occurrence of agreement between models. The shaded area indicates thresholds of optimum model performance based upon underfitting and overfitting. (D) A Venn diagram of the shared and unique MFs among those determined to be important for avoiding model overfitting (Top 500 ranked MFs for all three comparisons). (E) A Venn diagram of the shared and unique MFs among those determined to be important for avoiding model underfitting (Top 200 for Grade, 150 for Strain, and 100 for Growth). 111

Figure 2.6 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) hybrid striped bass (HSB) into classes (groups) for the Grade Comparison (see: Table 2.1) based upon expression patterns of unique, 12 amino acid-long motif fingerprints (MFs; attributes) quantitated from translated RNA-Seq data generated from these fish and identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered "top-ranked" among all 15,000 MFs measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were

sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K-fold cross-validation (20-folds) were the cross-validation strategies applied with each run of ML algorithm. Percent correct classification is plotted as an average calculated between the outcomes of each cross-validated algorithm run on subsets of the ranked attributes for this comparison. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance, “under”) and overfitting (i.e., too many highly-ranked attributes included, “over”) for this comparison, which were determined to be 200 and 500 attributes, respectively. This figure corresponds to the plot shown in Figure 2.5(B)..... 113

Figure 2.7 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% ,y-axis) hybrid striped bass (HSB) into classes (groups) for the Strain Comparison (see: Table 2.1) based upon expression patterns of unique, 12 amino acid-long motif fingerprints (MFs; attributes) quantitated from translated RNA-Seq data generated from these fish and identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 15,000 MFs measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K-fold cross-validation (20 folds) were the cross-validation strategies applied with each run of ML algorithm. Percent correct classification is plotted as an average calculated between the outcomes of each cross-validated algorithm run on subsets of the ranked attributes for this comparison. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance, “under”) and overfitting (i.e., too many highly-ranked attributes included, “over”) for this comparison, which were determined to be 150 and 500 attributes, respectively. This figure corresponds to the plot shown in Figure 2.5(B)..... 114

Figure 2.8 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% ,y-axis) hybrid striped bass (HSB) into classes (groups) for the Growth Comparison (see: Table 2.1) based upon expression patterns of unique, 12 amino acid-long motif fingerprints (MFs;

attributes) quantitated from translated RNA-Seq data generated from these fish and identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 15,000 MFs measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K-fold cross-validation (20-folds) were the cross-validation strategies applied with each run of ML algorithm. Percent correct classification is plotted as an average calculated between the outcomes of each cross-validated algorithm run on subsets of the ranked attributes for this comparison. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance, “under”) and overfitting (i.e., too many highly-ranked attributes included, “over”) for this comparison, which were determined to be 100 and 500 attributes, respectively. This figure corresponds to the plot shown in Figure 2.5(B)..... 115

Figure 2.9 The number of unique, twelve-amino acid long motif fingerprints (MFs) among those identified as highly informative in the differentiation of fish by growth performance (Grade and Growth comparisons) and/or paternal geographic location of origin (Strain comparison) (821 MFs total) displayed in a similar fashion to absolute abundance. The MFs mapped to thirty-three genes, the human (*Homo sapiens*) orthologue gene symbols are provided along the x-axis. These MFs were identified through application of a machine learning (ML) workflow to reduce dimensionality from 15,000 MFs to only those that yielded optimum classification performance (i.e., most individuals, or instances, correctly assigned comparison groups, or classes) by four distinct, cross-validated ML algorithms. The 500 most-informative MFs for each of the three comparisons were considered together and reduced to a total of 821 unique MFs. The MFs were mapped to the translated genome assemblies of the parental fish, striped bass (SB, paternal) and white bass (WB, maternal) and categorized as either “Both”, whereby complete homology (i.e., twelve-out-of-twelve amino acid match) was identified among the translated SB and WB sequences; “Undetm” or undetermined, whereby complete homology was not identified among either SB or WB sequences; “WB” if complete homology was exclusively identified among the WB sequence, and “SB” if complete homology was exclusively identified among the SB sequence. 116

Figure 2.10 The amount of unique, twelve-amino acid long motif fingerprints (MFs)

annotating back to each gene (x-axis) that were identified as having complete homology (i.e., twelve-out-of-twelve amino acid match) to “Both” the translated striped bass (SB, paternal) and white bass (WB, maternal) genome sequence assemblies; neither the SB or WB sequences and are therefore considered undetermined or “Undetm”, exclusively the WB sequence, or exclusively the SB sequence expressed as a relative percent (%) of all MFs annotating to a given gene. The MFs and subsequent genes examined here are those identified through the application of a machine learning (ML) workflow whereby 15,000 MFs were reduced in dimensionality to only those yielding the optimum classification performance of hybrid striped bass (HSB) offspring into comparison groups based on growth performance (Grade and Growth comparisons) and/or paternal geographic locations of origin (Strain comparison) by four cross-validated ML algorithms. The 500 most-informative MFs for each of the three comparisons were considered together and reduced to a total of 821 unique MFs, which were subsequently concatenated and aligned to protein sequences of the translated SB and/or WB genome assemblies. Human (*Homo sapiens*) orthologous were identified for subsequent pathway analyses and these gene symbols are provided here..... 118

Figure 2.11 Network of upstream regulatory molecules, causal networks, and downstream effects predicted to be associated to the input genes identified among hybrid striped bass (HSB). A subsample of the HSB study population (n=40) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups based on growth performance and sire strain (i.e., geographic location of origin). These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. There were 223 MFs that were among the most important for each comparison and these MFs concatenated to eight genes, which are included in the network here: apolipoprotein A-I (*APOA1*), carboxypeptidase B1 (*CPB1*), chymotrypsin-like elastase 1 (*CELA1*), chymotrypsin-like elastase 2A (*CELA2A*), hemoglobin subunit alpha 1 (*HBA1*), hemoglobin subunit beta (*HBB*), mitochondrially encoded ATP synthase membrane subunit 8 (*MT-ATP8*), and transferrin (*TF*). Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by

shape and line style, respectively, as follows: complexes or groups as double circles (e.g., Pdi), chemicals or drugs as a horizontal ovals (e.g., nicotinic acid), cytokines as squares (e.g., AGT), transmembrane receptors as vertical ovals (e.g., CUBN), peptidases as horizontal diamonds (e.g., CELA2A), transporters as irregular polygons (APOA1), mature microRNAs as semi-circles (e.g., miR-153-3p), and elements classified as “other” are displayed as circles (e.g., Marcks). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Arrows curving from a network element back to itself indicate a molecule interacts with itself (e.g., autophosphorylation). Predicted activity is indicated by orange for activation, blue for inhibition, pink or red for increased measurement, yellow for findings that are inconsistent with the up- or down-regulation of a downstream molecule, and grey indicates that an effect is not predicted. Color intensity indicates that it is more extreme in the dataset. 119

Figure 2.12 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in “Top Grade” (TG) hybrid striped bass (HSB), whereby TG is used to describe HSB that at two months of age were graded as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age. A subsample of the HSB study population (n=40, 20 HSB in TG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p=0.05$ indicated by “Threshold” line parallel to x-axis). Blue bars represent inhibited pathways, from left: Granzyme A Signaling, Dilated Cardiomyopathy Signaling Pathway, and Sirtuin Signaling Pathway. Orange bars represent activated pathways, from left: Oxidative Phosphorylation, LXR/RXR Activation, SNARE Signaling Pathway, ILK Signaling, Actin Cytoskeleton Signaling, Oxytocin Signaling Pathway, and Phagosome Formation. Grey bars indicate that

activity prediction cannot be made based upon the specific pathway construction and associated molecules, from left: Mitochondrial Dysfunction, FXR/RXR Activation, and Iron homeostasis signaling pathway. White bars have a z-score of zero indicating that the evidence for activation and inhibition are equal preventing a prediction from being made: Calcium Signaling and Estrogen Receptor Signaling. 121

Figure 2.13 Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among “Top Grade” (TG) hybrid striped bass (HSB), whereby TG is used to describe HSB that at two months of age were graded as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age. A subsample of the HSB study population (n=40, 20 HSB in TG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., TGFB1), transcription regulators are displayed as horizontal ovals (e.g., STAT3), peptidases are displayed as horizontal diamond (e.g., LONP1), enzymes are displayed as vertical diamonds (e.g., SIRT3), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, yellow for findings that are inconsistent with the up- or down-regulation of a downstream molecule, and grey indicates that an effect is not predicted..... 122

Figure 2.14 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in “Large” (LG) hybrid striped bass (HSB). HSB were graded at

two months of age as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age (referred to as Top Grade, TG), or not (referred to as Runts). LG is used to describe HSB that reached or exceeded the mean weight of HSB in the TG group. A subsample of the HSB study population (n=40, 13 HSB in LG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into growth performance groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher's Exact Test (p=0.05 indicated by "Threshold" line parallel to x-axis). Blue bars represent inhibited pathways, from left: Dilated Cardiomyopathy Signaling Pathway, Granzyme A Signaling, and Sirtuin Signaling Pathway. Orange bars represent activated pathways, from left: Oxidative Phosphorylation, LXR/RXR Activation, SNARE Signaling Pathway, ILK Signaling, Neutrophil Extracellular Trap Signaling Pathway, Actin Cytoskeleton Signaling, Oxytocin Signaling Pathway, and Phagosome Formation. Grey bars indicate that activity prediction cannot be made based upon the specific pathway construction and associated molecules, from left: Mitochondrial Dysfunction, FXR/RXR Activation, Iron homeostasis signaling pathway, and Cellular Effects of Sildenafil (Viagra). White bars have a z-score of zero indicating that the evidence for activation and inhibition are equal preventing a prediction from being made: Calcium Signaling. 124

Figure 2.15 Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among "Large" (LG) hybrid striped bass (HSB). HSB were graded at two months of age as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age (referred to as Top Grade, TG), or not (referred to as Runts). LG is used to describe HSB that reached or exceeded the mean weight of HSB in the TG group. A subsample of the HSB study population (n=40, 13 HSB in LG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible,

mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., TNF), transcription regulators are displayed as horizontal ovals (e.g., STAT3), peptidases are displayed as horizontal diamond (e.g., LONP1), enzymes are displayed as vertical diamonds (e.g., SIRT3), functions are displayed as octagons (e.g., Activation of neutrophils), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted..... 125

Figure 2.16 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in HSB produced from domestic (DOM) striped bass (SB) sires that were produced and reared at the North Carolina State University Pamlico Aquaculture Field Lab (NCSU PAFL, Aurora, NC). A subsample of the HSB study population (n=40, 12 HSB in DOM group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into growth performance groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal SB (*M. saxatilis*) fish, and examined further via pathway analysis. Pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test (p=0.05 indicated by “Threshold” line parallel to x-axis). Blue bars represent inhibited pathways, from left: Dilated Cardiomyopathy Signaling Pathway, Granzyme A Signaling, and Sirtuin Signaling Pathway. Orange bars represent activated pathways, from left: Oxidative Phosphorylation,

LXR/RXR Activation, SNARE Signaling Pathway, ILK Signaling, Neutrophil Extracellular Trap Signaling Pathway, Actin Cytoskeleton Signaling, Oxytocin Signaling Pathway, and Phagosome Formation. Grey bars indicate that activity prediction cannot be made based upon the specific pathway construction and associated molecules, from left: Mitochondrial Dysfunction, FXR/RXR Activation, Iron homeostasis signaling pathway, and Cellular Effects of Sildenafil (Viagra). White bars have a z-score of zero indicating that the evidence for activation and inhibition are equal preventing a prediction from being made: Calcium Signaling. 127

Figure 2.17 Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among HSB produced from domestic (DOM) striped bass (SB) sires that were produced and reared at the North Carolina State University Pamlico Aquaculture Field Lab (NCSU PAFL, Aurora, NC). A subsample of the HSB study population (n=40, 12 HSB in DOM group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal SB (*M. saxatilis*) fish, and examined further via pathway analysis. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., TGFBI), transcription regulators are displayed as horizontal ovals (e.g., STAT3), peptidases are displayed as horizontal diamond (e.g., LONP1), enzymes are displayed as vertical diamonds (e.g., SIRT3), mature microRNAs are displayed as half circles (e.g., miR-3150a-3p), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, yellow for findings that are inconsistent with the up- or

down-regulation of a downstream molecule, and grey indicates that an effect is not predicted..... 128

Figure 2.18 The LXR/RXR Activation pathway (liver X receptor/retinoid X receptor) enriched from genes identified among hybrid striped bass (HSB) exhibiting superior growth and/or of domestic sire strain (i.e., male striped bass were not of wild-origin). Specifically, apolipoprotein a-I (*APOA1*) and apolipoprotein E (*APOE*) are highlighted in red and are up-regulated in the dataset. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of other interacting elements indicated as being activated if in orange or inhibited in blue..... 130

Figure 2.19 The FXR/RXR Activation pathway (farnesoid X receptor/retinoid X receptor) enriched from genes identified among hybrid striped bass (HSB) exhibiting superior growth and/or of domestic sire strain (i.e., male striped bass were not from wild-origin). Specifically, apolipoprotein a-I (*APOA1*) apolipoprotein E (*APOE*), shown in red to indicate their up-regulation in the dataset. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of other interacting elements indicated as being activated if in orange or inhibited in blue..... 131

CHAPTER 3

Figure 3.1 Growth and feed consumption of a population of striped bass (SB; Initial N=194, Final N=173) raised in triplicate recirculating aquaculture system (RAS) tanks at Grinnell's Animal Health Laboratory (North Carolina State University, Raleigh, NC, USA) for a year from five to eighteen months of age. Fish were weighed every three months and feed was administered *ad libitum* for the entirety of the study to monitor consumption and enable the calculation of amount of feed consumed per number of fish over time and while accounting for mortality. The average weight in grams (g) of the population (left y-axis) is represented by the black solid line and bars represent the standard deviation. The average amount of feed consumed (g) per fish is represented by the dotted line and open circles. 183

Figure 3.2 Frequency distribution of the weights (g) of striped bass (SB) sacrificed at eighteen months of age (n=72 of 173) for muscle histology and gene expression analysis. The x-axis indicates the center value of each bin (e.g., the four fish in the 700 g bin weighed between 675 and 725 g). Fish were reared in triplicate recirculating aquaculture system (RAS) tanks at Grinnell's Animal Health Laboratory (North Carolina State University, Raleigh, NC, USA) for a year from five to eighteen months of age. 184

Figure 3.3 Mean weight (kg) of striped bass (SB) offspring (y-axis) produced from one of twelve SB sires (x-axis). Sires numbered 1–6 (“#-L/S”) were

crossed with the same SB female (“Dam A”) and are therefore of half sibling families. Similarly, sires numbered 7–12 were crossed with a different SB female (“Dam B”). Sires belonged to two different groups, Large, indicated by “-L”, or Small, indicated by “-S”, that significantly differed in weight and total length, TL (Large: 2.76 ± 0.15 kg weight, 570.67 ± 7.84 mm TL; Small: 1.75 ± 0.15 kg weight, 490.00 ± 21.14 mm TL; $p < 0.0001$). Vertical numbers indicate the number of offspring produced from each sire of the total sacrificed for this study ($n=72$). Standard deviation bars are present if more than one offspring was produced from a given sire. Differentiating letters indicate significant differences between mean weight of offspring born to individual sires as determined by one-way ANOVA and Tukey’s HSD post-hoc test ($p \leq 0.0201$). 185

Figure 3.4 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) striped bass (SB) into classes (groups) for the Growth Comparison (*see*: Table 3.1) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified *K*-fold cross-validation (*K* = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (965 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison, which were determined to be 10 and 300 attributes, respectively. 186

Figure 3.5 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) striped bass (SB) into classes (groups) for the Market Size Comparison (*see*: Table 3.1) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in

learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (K = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (284 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison, which were determined to be 75 and all 284 attributes, respectively..... 187

Figure 3.6 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) striped bass (SB) into classes (groups) for the Dam Comparison (*see*: Table 3.1) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (K = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (257 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison, which were determined to be 25 and 150 attributes, respectively. 188

Figure 3.7 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) striped bass (SB) into

classes (groups) for the Sire Comparison (*see*: Table 3.1) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified *K*-fold cross-validation (*K* = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (35) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison was determined to be including all 35 ranked attributes. 189

Figure 3.8 The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% ,y-axis) striped bass (SB) into classes (groups) for the Sire Size Comparison (*see*: Table 3.1) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified *K*-fold cross-validation (*K* = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (378 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too

many highly-ranked attributes included) for this comparison, which were determined to be 15 and 75 attributes, respectively. 190

Figure 3.9 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) designated as Superior (A) or Inferior (B) based upon growth performance. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those in the Superior group were significantly greater in weight (1.32 ± 0.18 kg/fish) and total length (TL, 449.00 ± 16.54 mm/fish) than those in the Inferior group (0.85 ± 0.12 kg/fish and 399.58 ± 15.96 mm/fish; Student's *t*-Test, $p < 0.0001$). Genes included in analysis (294 unique, analysis-ready molecules from a list of 300 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher's Exact Test ($p = 0.05$ indicated by "Threshold" line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made. 191

Figure 3.10 Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among striped bass (SB) designated as Inferior based upon growth performance. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36) and these groups significantly differed in weight and total length (Student's *t*-Test, $p < 0.0001$). Genes included in analysis (294 unique, analysis-ready molecules from a list of 300 gene transcripts total, 272 of which were up-regulated in Inferior SB) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance through the application of a machine learning workflow. Qiagen Ingenuity Pathway

Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., HGF), transcription regulators are displayed as horizontal ovals (e.g., MYC), mature microRNAs are displayed as half circles (e.g., miR-16-5p), functions are displayed as octagons (e.g., Cell Survival), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted. 193

Figure 3.11 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) designated as Under Size (A), Inferior Other (B), Market Size (C), or Superior Other (D) based upon growth to market size (1.36 kg, or 3.0 lbs) of these fish. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) and the remaining SB in each group were referred to as Superior Other and Inferior Other, respectively. SB in all groups significantly differed in weight and total length (one-way ANOVA and Tukey’s HSD post hoc test, $p < 0.0001$). Genes included in analysis (277 unique, analysis-ready molecules from a list of 284 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance to market size through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p = 0.05$ indicated by “Threshold” line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made

based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made..... 194

Figure 3.12 Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among striped bass (SB) designated as Under Size. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (1.36 kg, or 3.0 lbs; n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) and the remaining SB in each group were referred to as Superior Other and Inferior Other, respectively. SB in all groups significantly differed in weight and total length (one-way ANOVA and Tukey’s HSD post hoc test, $p < 0.0001$). Genes included in analysis (277 unique, analysis-ready molecules from a list of 284 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance to market size through the application of a machine learning workflow. A total of 194 of the 277 unique genes were identified as up-regulated in Under Size SB. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., HGF), transcription regulators are displayed as horizontal ovals (e.g., MYC), mature microRNAs are displayed as half circles (e.g., miR-16-5p), functions are displayed as octagons (e.g., Cell Survival), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted. 196

Figure 3.13 Network of upstream regulatory molecules and pathways predicted to

underlie observed patterns in gene expression among striped bass (SB) designated as Market Size. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (1.36 kg, or 3.0 lbs; n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) and the remaining SB in each group were referred to as Superior Other and Inferior Other, respectively. SB in all groups significantly differed in weight and total length (one-way ANOVA and Tukey’s HSD post hoc test, $p < 0.0001$). Genes included in analysis (277 unique, analysis-ready molecules from a list of 284 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance to market size through the application of a machine learning workflow. A total of 49 of the 277 unique genes were identified as up-regulated in Under Size SB. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., HGF), transcription regulators are displayed as horizontal ovals (e.g., MYC), mature microRNAs are displayed as half circles (e.g., miR-16-5p), functions are displayed as octagons (e.g., Cell Survival), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted. 198

Figure 3.14 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) produced from one of two female SB, Dam A (A) or Dam B (B). Briefly, SB had been produced by crossing two female SB with six different male SB (sires) each, whereby three of the sires each dam was crossed with were designated as “Large” or “Small” as they significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Offspring were reared in triplicate tanks of an indoor

recirculating aquaculture system (RAS) and sampled at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36) This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36), and these fish significantly differed in weight and total length (Student's *t*-Test, $p < 0.0001$). Genes included in analysis (139 unique, analysis-ready molecules from a list of 150 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher's Exact Test ($p = 0.05$ indicated by "Threshold" line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made..... 199

Figure 3.15 Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression in striped bass (SB) offspring produced by a single female, Dam B, and crossed with six males, three belonging to a "Large" group and three belonging to a "Small" group, which significantly differed in weight and total length (Student's *t*-Test, $p < 0.0001$). Genes included in analysis (139 unique, analysis-ready molecules from a list of 150 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. A total of 94 of these 139 genes were up-regulated in offspring produced from Dam B. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., AGT), mature microRNAs are displayed as half circles (e.g., miR-6825-5p), canonical pathways are displayed as hourglass hexagons (e.g., Signaling by Rho Family GTPases), and functions are displayed as octagons (e.g., Proliferation of neuroglia). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and

dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, yellow for findings that are inconsistent with the up- or down-regulation of a downstream molecule, and grey indicates that an effect is not predicted. 201

Figure 3.16 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) produced from one of twelve male SB Sires 1–12. Briefly, SB had been produced by crossing two female SB (Dam A and Dam B) with six different male SB (sires) each, whereby three of the sires each dam was crossed with were designated as “Large” or “Small” as they significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Offspring were reared in triplicate tanks of an indoor recirculating aquaculture system (RAS) and sampled at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth ($n=36$) and those that exhibited relatively poorer, or Inferior growth ($n=36$), and these fish significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Genes included in analysis (33 unique, analysis-ready molecules from a list of 35 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p=0.05$ indicated by “Threshold” line parallel to *x*-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated *z*-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made..... 202

Figure 3.17 Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) belonging to the “Large” (A) or “Small” (B) size group. Briefly, SB had been produced by crossing two female SB (Dam A and Dam B) with six different male SB (sires) each, whereby three of the sires each dam was crossed with were designated as “Large” or “Small” as they significantly differed in weight and total length (Student’s *t*-Test,

$p < 0.0001$). Offspring were reared in triplicate tanks of an indoor recirculating aquaculture system (RAS) and sampled at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth ($n=36$) and those that exhibited relatively poorer, or Inferior growth ($n=36$), and these fish significantly differed in weight and total length (Student's t -Test, $p < 0.0001$). Genes included in analysis (71 unique, analysis-ready molecules from a list of 74 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher's Exact Test ($p=0.05$ indicated by "Threshold" line parallel to x -axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z -score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made..... 203

CHAPTER 4

- Figure 4.1 Top-down and bottom-up approaches to processing. The top-down approach is more commonly associated with principles of hypothesis testing and reductionism to draw specific conclusions from general knowledge. The bottom-up approach is more commonly associated with principles of discovery-based, holistic, and integrative methodologies that are data-driven whereby conclusions are based solely on what is represented by the data and little to no assumptions are made prior to. Top-down and bottom-up approaches to processing are not mutually exclusive; they are connected through the handling of data (represented by the dotted line) and are implemented in many procedures and processes beyond data analysis and research. 263
- Figure 4.2 Four categories of machine learning (ML) and subtypes thereof. The dotted ovals encompassing subtypes of Traditional Learning, also referred to as Classical Learning, indicate which of these subtypes are often further described as being "supervised" or "unsupervised" depending on whether labels designating groupings (classes) of instances (e.g., samples) are or are not associated with input data, respectively. 264
- Figure 4.3 (A) Workflow scheme of a supervised machine learning (ML) analysis. The outcome of applying the ML workflow is a dataset reduced in dimensionality such that only the attributes most determinant (i.e., those

providing the most information) in the classification of instances (observations) into user-defined classes (groups) remain and a negative control for learning has been performed. Attributes are first reduced based upon Shannon’s Entropy and then again by determining the “optimal” number of attributes to yield superior classification by performing iterative runs of four orthogonal, twice cross-validated analytical algorithms on fewer and fewer attributes. Specifically, each attribute is assigned an information gain value based upon the amount of information provided to the classification of instances into classes and with higher values indicating greater information gain. Only those attributes with information gain above zero (or another user-selected cutoff) are included in the determination of the optimal dataset, which are those attributes of highest information gain value that account for (by avoiding) poor predictive ability based upon the inclusion of either too few or too many attributes (underfitting and overfitting, respectively). This optimal dataset is analysis ready (e.g., for pathway analysis). (B) A description of applying this workflow to an example dataset of 10,000 genes identified among twenty instances (biological replicates) that are split equally into two classes, “Good” and “Bad” based upon some observable quality. The example graph depicts algorithm performance measured as percent correctly classified instances (y-axis) into classes with the exclusion of fewer attributes in each iterative run of the cross-validated algorithms from left to right (x-axis). The grey box represents the threshold of underfitting and overfitting as including the top 50 and top 500 attributes, respectively..... 265

Figure 4.4 Example layout of a data file formatted for Weka machine learning (ML) software (University of Waikato, New Zealand; Eibe et al., 2016, www.cs.waikato.ac.nz/ml/weka/)..... 267

CHAPTER 5

Figure 5.1 Hierarchical clustering heat-map of differences in expression of motif-fingerprints (MFs, protein fragments associated to specific genes) between hybrid striped bass (HSB) of different grade and ultimate growth performance (Large and Small) and from three different families (A through C). Up-regulation is represented in red and down-regulation is represented on the blue to yellow color scale. The eight MFs outlined in black at the bottom of the heat-map are specific to Lipocalin-2 (*lcn2*) a gene involved in energy metabolism, innate immunity, and skeletal muscle regeneration. 280

CHAPTER 1. THE STATUS OF STRIPED BASS, *MORONE SAXATILIS*, AS A COMMERCIALLY READY SPECIES FOR U.S. MARINE AQUACULTURE

Linnea K. Andersen, Jason Abernathy, David L. Berlinsky, Greg Bolton, Matthew M. Booker, Russell J. Borski, Travis Brown, David Cerino, Michael Ciaramella, Robert W. Clark, Michael O. Frinsko, S. Adam Fuller, Steven Gabel, Batholomew W. Green, Eric Herbst, Ronald G. Hodson, Michael S. Hopper, Linas W. Kenter, Frank Lopez, Andrew S. McGinty, Barry Nash, Matthew Parker, Stacey Pigg, Steven Rawles, Kenneth Riley, Marc J. Turano, Carl D. Webster, Charles R. Weirich, Eugene Won, L. Curry Woods III, and Benjamin J. Reading

Manuscript published in *Journal of the World Aquaculture Society*:

Andersen, L.K., Abernathy, J., Berlinsky, D.L., Bolton, G., Booker, M.M., Borski, R.J., Brown, T., Cerino, D., Ciaramella, M., Clark, R.W., Frinsko, M.O., Fuller, S.A., Gabel, S., Green, B.W., Herbst, E., Hodson, R.G., Hopper, M.S., Kenter, L.W., Lopez, F., McGinty, A.S., Nash, B., Parker, M., Pigg, S., Rawles, S., Riley, K., Turano, M.J., Webster, C.D., Weirich, C.R., Won, E., Woods III, L.C., and Reading, B.J. 2021. The status of striped bass, *Morone saxatilis*, as a commercially ready species for US marine aquaculture. *Journal of the World Aquaculture Society*, 52(3), pp.710-730. DOI: 10.1111/jwas.12812.

Abstract

Striped bass, *Morone saxatilis*, is an anadromous fish native to the North American Atlantic Coast and is well recognized as one of the most important and highly regarded recreational fisheries in the United States. Decades of research have been conducted on striped bass and its hybrid (striped bass × white bass *Morone chrysops*) and culture methods have been established, particularly for the hybrid striped bass, the fourth largest finfish aquaculture industry in the United States (US \$50 million). Domesticated striped bass have been developed since the 1990s and broodstock are available from the government for commercial fry production using novel hormone-free methods along with traditional hormone-induced tank and strip spawning. No commercial-scale intensive larval rearing technologies have been developed at present and current fingerling production is conducted in fertilized freshwater ponds. Larval diets have not

been successfully used as first feeds; however, they have been used for weaning from live feeds prior to metamorphosis. Striped bass can be grown out in marine (32 ppt) or freshwater (<5 ppt); however, they require high hardness (200+ ppm) and some salinity (8–10 ppt) to offset handling stress. Juveniles must be 1–10 g/fish prior to stocking into marine water. Commercially available fingerling, growout, and broodstock feeds are available from several vendors. Striped bass may reach 1.36 kg/fish in recirculating aquaculture by 18 months and as much as 2.27 kg/fish by 24 months. Farm gate value of striped bass has not been determined, although seasonally available wild-harvested striped bass are valued at about US \$6.50 to US \$10.14 per kg and cultured hybrid striped bass are valued at about US \$8.45 to US \$9.25 per kg whole; the farm gate value for cultured striped bass may be as much as US \$10.00 or more per kg depending on demand and market. The ideal market size is between 1.36 and 2.72 kg/fish, which is considerably larger than the traditional 0.68 to 0.90 kg/fish for the hybrid striped bass market.

Introduction

Striped bass, *Morone saxatilis*, are a well-recognized fish native to the North American Atlantic Coast and regarded as the most popular recreational fishery in the United States. Striped bass are euryhaline, anadromous fish and juveniles typically remain in estuaries for 2–4 years prior to migrating to and from the north and south Atlantic Ocean seasonally as adults, ascending to rivers each spring to spawn (Callihan, Harris, & Hightower, 2015). In some cases, such as the Hudson and Cape Fear Rivers and Santee-Cooper reservoir, not all striped bass migrate into the ocean, as some may remain resident in freshwater (Haeseker, Carmichael, & Hightower, 1996; LeBlanc et al., 2020; Waldman, Dunning, Ross, & Mattson, 1990; Wirgin, Maceda, Tozer, Stabile, & Waldman, 2020). The life history and culture of this fish has been researched for

decades (Harrell, 1997; Harrell, Kerby, & Minton, 1990; McCraren, 1984), in part due to its use as the progenitor for creation of the original hybrid striped bass cross (striped bass × white bass, *Morone chrysops*; Palmetto bass) for stocking into natural and man-made impoundments for recreational fisheries, and the reciprocal hybrid striped bass (white bass × striped bass; Sunshine bass), which is raised in aquaculture. Hybrid striped bass is the fourth largest finfish aquaculture industry in the nation at a farm gate value of US \$50 million when accounting for sales of foodfish as well as fry and fingerling used for commercial growout (Reading et al., 2018; USDA, 2019). A domestic stock of striped bass has been bred for six generations in captivity and distributed across North America as broodfish for hybrid striped bass foodfish production and recreational fishery stock enhancement. However, the establishment of striped bass as a commercial aquaculture industry independent of hybrid striped bass is predominantly stagnant because of several challenges, including inconsistent market demand and lack of supportive regulations and demonstrated sustained market viability.

A recent increased focus on agricultural and coastal development and economic growth in the seafood sector has created an opportunity for establishing the striped bass aquaculture industry. Specifically, the seafood trade deficit in the United States is nearing US \$17 billion (NMFS, 2019) and 9 out of 10 seafood products consumed in the United States are of foreign import, half of which are aquaculture products (NMFS, 2020). Although the U.S. aquaculture industry (US \$1.52 billion in 2018, USDA, 2019) has grown in recent years, particularly in the production of bivalves such as clams and oysters, it has remained a minor aquaculture producer on a global scale (ranked 17th; NMFS, 2020). Expansion of finfish aquaculture, particularly of striped bass and other species, represents one of the greatest unrealized aquaculture industry growth potentials in the world (FAO, 2018; Lem, Bjorndal, & Lappo, 2014).

Only one-third of global aquaculture products are raised in marine waters, which presents an opportunity for industry expansion as these marine resources and species are currently underutilized in the United States and other countries (Froehlich, Gentry, & Halpern, 2018). The expansion of marine finfish production is hindered by the limited number of appropriate species choices. Atlantic salmon, *Salmo salar* and red drum, *Sciaenops ocellatus* are currently the only finfish species endemic to the United States that are cultured in significant quantities in coastal environments, and presently there is no appreciable aquaculture production of any premium white-fleshed marine finfish species, such as the striped bass, in the country. Candidate aquaculture species identified by the National Oceanic and Atmospheric Administration (NOAA) must command a premium price, have high consumer demand, and successfully adapt to rearing in localized environments for profitable production. The striped bass meets all these criteria and therefore has great potential for commercialization in the United States (Reading, 2017; Reading, Hinshaw, & Watanabe, 2014).

Our purpose is to review the current status of striped bass aquaculture and its potential as a U.S. aquaculture industry, primarily from an Atlantic state perspective. The industry originated and the seminal research was conducted in this region, which is well suited for the culture of striped bass. Section 1 provides an overview of the history and management of the striped bass fishery, the establishment of striped bass culture over time, and the current market opportunity. Section 2 describes the standard methods and tools available for striped bass culture, and Section 3 presents the future directions of the striped bass aquaculture industry, including challenges toward establishment of a culture industry for this species, to be addressed in part through the launch of the new *StriperHub* consortium.

History of the striped bass fishery and aquaculture and the current market opportunity

Striped bass fishery status

The endemic range of striped bass stretches from the St. Lawrence River in Canada to St. John's River in Florida and the importance of this fish to commercial and recreational fisheries dates to pre-colonial times (Hill, Evans, & Van Den Avyle, 1989). Overfishing and habitat degradation contributed to the collapse of the striped bass fishery in the 1980s and prompted the development and approval of the Interstate Fisheries Management Plan (ISFMP) for Atlantic Striped Bass by the Atlantic States Marine Fisheries Commission (ASMFC) in 1981 (ASMFC, 1981). Continued decline of striped bass populations led to the passage by Congress in 1984 of the Atlantic Striped Bass Conservation Act (98 Stat. 3187, 16 U.S.C. 5151-5158) that granted the Secretary of Commerce the authority to impose a total moratorium in any state that did not comply with ASMFC management guidelines. To date, the initial striped bass ISFMP put forth by the ASMFC has been amended six times. Notably, striped bass stocks were declared fully recovered in 1995 upon adoption of Amendment 5, which was preceded by the stipulation that focus should be on rebuilding the fishery rather than maximizing harvest yield per Amendment 4 (1989). After the declaration of recovery, per annum commercial striped bass harvest grew from about 1.5 million kg (3.4 million lb) in 1995 to about 2.7 million kg (6.0 million lb) in 2002 (Figure 1.1.a). Atlantic striped bass are currently managed by the ASMFC under Amendment 6 (2003) that established additional biological reference points for management, a commercial quota system, and bag and size limits for recreational fishing.

Although it does not currently have a direct role in managing the striped bass fishery, the NOAA also provides key research and scientific findings to the ASMFC and state agencies that continue to monitor and manage striped bass stocks. Moreover, a series of six addenda to

Amendment 6 were implemented beginning in 2007 to address items such as bycatch monitoring, how recruitment failure is considered, commercial harvest tagging, and the modeling of Atlantic striped bass as a single stock. A spawning stock biomass (SSB) assessment for striped bass conducted in 2013 estimated that current fishing mortality rates would reduce SSB below the 90.7 million kg (200.0 million lb) threshold over the next few years (Figure 1.1.b). In response to this 2013 stock assessment, the ASMFC approved Addendum IV in 2014 to reduce harvest levels by 25.0% in coastal states and 20.5% in the Chesapeake Bay. Addendum IV successfully reduced fishing mortality based on stock assessments conducted in 2016; however, striped bass SSB continued to decline. Commercial landings averaged about 3.2 million kg (7.0 million lb) annually from 2003 to 2014, and in 2017 the commercial quota was reduced to approximately 2.3 million kg (5.0 million lb) through Addendum IV (Figure 1.1.a). Recreational landings along the Atlantic coast reported by the National Marine Fisheries Service (NMFS) Office of Science and Technology (NOAA) were about 11.3 million kg (25.0 million lb) per year from 2007 to 2014 and harvests from 2015 to 2017 were reduced due to implementation of Addendum IV.

As the Albemarle Sound and Roanoke River striped bass stocks contribute minimally to the Atlantic striped bass population in comparison with the Chesapeake Bay and Delaware and Hudson Rivers, Addendum IV deferred management of this stock to the State of North Carolina. In 2017, commercial harvest in the Albemarle Sound and Roanoke River of North Carolina was estimated at 34,375 kg (75,783 lb) and recreational harvest estimated at 45,872 kg (101,131 lb). The North Carolina Division of Marine Fisheries (NC DMF) closed the striped bass season for commercial and recreational fishermen in all internal waters from just south of Oregon Inlet to

the South Carolina line in 2019. This is the Central Southern Management Area (CSMA) and includes the Tar, Neuse, Pamlico, and Cape Fear River systems.

A benchmark assessment in 2019 indicated that striped bass SSB was approximately 22.7 million kg (50.0 million lb) below the threshold of 91.6 million kg (202.0 million lb) and determined that the stock had been overfished since 2013. Addendum VI was initiated in 2019 as an adaptive management strategy to end overfishing and bring fishing mortality levels to the target level in 2020. Addendum VI specifically aims to reduce removals along the Atlantic coast by at least 18.0% and mandates the use of circle hooks and a 1-fish bag limit and 28 to 35-in. slot limit for recreational ocean fisheries and an 18-in. minimum size limit for the Chesapeake recreational fishery. States are still permitted to implement alternative regulations through conservation equivalency under Addendum VI. A proposed motion to initiate an amendment that will serve to address stock rebuilding and other management strategies is currently up for review in 2021 by the ASFMC.

Early striped bass aquaculture

Practices to culture striped bass were initially developed to improve production of fish for enhancing commercial and recreational fisheries of native Atlantic coastal stocks. This was later expanded to include non-native introduction to the Pacific Ocean in 1879 (Parks, 1978; Stone, 1882) and inland freshwater reservoirs beginning in 1957 (Stevens, 1975, 1984). The first published report of a successful hatch of striped bass eggs under artificial conditions was made in 1874 by Spencer Baird, the first commissioner of the U.S. Commission of Fish and Fisheries, later to become the U.S. Fish and Wildlife Service (USFWS) (Baird, 1874). In 1879, the USFWS hatched striped bass fry at a site located along the Abermarle Sound in North Carolina that had

been used as an American shad (*Alosa sapidissima*) hatchery (USFWS, 1882). These fry were sent to Washington D.C. and Baltimore, Maryland (USFWS, 1882). Seven years later in 1884, Stephen G. Worth reported construction of the first dedicated striped bass hatchery on the Roanoke River in Weldon, North Carolina (Worth, 1884). In a subsequent report, Worth (1904) described production of striped bass in that hatchery that included collecting over 2 million eggs and stocking “almost 300,000 fry” into the Roanoke River. The Edenton National Fish Hatchery was then established in North Carolina in 1898 by the USFWS with a similar purpose to Weldon (Woodroffe, 2012).

Beginning in the early 20th century, the USFWS began publishing manuals describing effective techniques for spawning, hatching, and releasing fry of cultured fish, including striped bass (Piper, 1982). By 1910, the basic technology of striped bass aquaculture was in place, but plans to augment marine fishery stocks were inexplicably dropped by the commission (Worth, 1910). Renewed interest in stock enhancement developed in the 1950s after the 1954–1955 discovery that a resident striped bass population had become established in the freshwater Santee-Cooper Reservoir of South Carolina (Scruggs Jr., 1957). The purpose of this new hatchery augmentation program was to establish new populations of striped bass in freshwater rivers and reservoirs throughout the southeastern United States, in states such as Kentucky, Alabama, Georgia, and South Carolina (Geiger & Parker, 1985; Kinman, 1988; Stevens, 1975). Striped bass were also being stocked as part of a fisheries management strategy to help control gizzard shad (*Dorosoma cepedianum*) populations, while providing anglers with a new recreational fishery (Anderson, 1966; Bonn, Bailey, & Bayless, 1976). By the 1980s, striped bass had been introduced into hundreds of reservoirs in at least 36 states (Stevens, 1984).

The first attempts to induce striped bass to spawn using hormones were made by Robert E. Stevens in the 1960s (Stevens, 1966, 1967). Within only a few years, procedures were developed that allowed the two principal hatcheries, the old hatchery established during the earliest years of striped bass culture in Weldon, North Carolina, and the newer hatchery built in Moncks Corner, South Carolina in 1961, to produce millions of striped bass fry annually (Mischke, 2012; Stevens, 1967). The first successful *Morone* hybridization cross was conducted vis-à-vis to the development of procedures to artificially spawn striped bass in captivity in the 1960s. This original cross hybrid, also referred to as the palmetto cross, was made using striped bass eggs and white bass sperm (milt) with the intention of creating a fish that had the hardiness and environmental tolerance of a white bass and would grow to the size of a striped bass, thus appealing to anglers.

Commercial aquaculture of hybrid striped bass began in the 1970s, but it did not gain market footing. It was not until moratoriums were imposed (Maryland 1985–90; Virginia 1989–90) following the collapse of the striped bass fishery in the 1980s (Figure 1.1.a) that the path for commercial hybrid striped bass aquaculture as a means of supplying a replacement product opened (Hodson, 1990; Hodson & Hayes, 1990). The initial pond and small tank aquaculture efforts were pioneered by Theodore I. J. Smith (South Carolina Department of Natural Resources), J. Howard Kerby and Melvin T. Huish (North Carolina State University, North Carolina Cooperative Research Unit), and Ronald G. Hodson (North Carolina State University, NOAA North Carolina Sea Grant), among others. By 1987, the National Coastal Resources Research and Development Institute in North Carolina developed a national research initiative to establish the feasibility of commercial production and profitability of hybrid striped bass reared in ponds. Reginal Harrell (University of Maryland, USDA Northeastern Regional Aquaculture

Center) coordinated the summation of these early efforts and methodologies to produce comprehensive reference manuals for culture and propagation of striped bass and its hybrids (Harrell, 1997; Harrell et al., 1990). The first commercial harvest of hybrid striped bass was in 1987 and the industry has since grown to produce 5.4 million kg (12.0 million lb) of hybrid striped bass (white bass eggs and striped bass milt; Reciprocal or Sunshine) annually with a farm gate value of US \$50 million (Reading, McGinty, et al., 2018; USDA, 2019) (Figure 1.1.c). Hybrid striped bass foodfish are raised in 19 states of the United States and approximately 50% of the production occurs in Texas, California, and Mississippi with the remaining production largely occurring throughout the Mid-Atlantic and southeast U.S. Commercial fingerling production occurs largely in Arkansas and North Carolina.

Current market opportunity

Barring any challenges to the expansion of the U.S. aquaculture industry, the market opportunity for striped bass exists, is strong, and is largely untapped. The seafood trade deficit and growth potential of aquaculture in the United States warrant the development of commercial marine aquaculture and recent evidence from seafood markets along the mid-Atlantic region indicate high demand for larger, white-fleshed marine fish with desired size of 1.36–2.27 kg (3.0–5.0 lb) per fish (Locals Seafood, Raleigh, NC, personal communication and unpublished data from current retail seafood markets). This demand cannot be met by currently available commercial aquaculture species including the hybrid striped bass, whose growth and feed efficiency rapidly decline after the fish reach 0.68 kg (1.5 lb) in size (Turano and Reading, unpublished data). However, some producers in Texas and Mississippi have reported rearing hybrid striped bass to 1.4 kg (3.0 lb) in 18–24 months (Treece and Associates, 2017).

Currently, tilapia (genus *Oreochromis*), pollock (*Gadus chalcogrammus*), cod (*Gadus morhua*), and catfish (genus *Ictalurus*) are ranked fourth, fifth, seventh, and eighth, respectively, among the 10 most popular seafoods in the United States (NFI, 2018). From a culinary perspective, the fillets from these finfishes possess sensory characteristics that are highly valued by professional chefs and discerning home cooks. With their slightly sweet flavor and relatively firm texture once cooked, wild and farmed striped bass can easily be prepared with recipes that have already been crafted for a number of white-meat finfish (NOAA, 2020a, 2020b; SeafoodSource, 2014a, 2014b). Given the mild flavor, the meat of striped bass can easily absorb an assortment of herbs and spices, allowing chefs and home cooks to create a variety of highly flavorful meal preparations. Like cod, pollock, and tilapia, striped bass also is a good source of nutritious, low-fat protein (NOAA, 2020c).

Striped bass, unlike hybrid striped bass, can be grown in “open” marine systems (e.g., coastal areas) or produced in freshwater land-based systems prior to marine transfer. Relative to other marine finfishes, the striped bass is a well-suited candidate to meet the seafood market demand, as the target market size of 1.36 kg (3.0 lb) for striped bass can be attained within approximately 24 months or less of growout. Furthermore, the reproduction, genetics, culture, and feed requirements of striped bass have been studied extensively largely through the development of the hybrid striped bass industry and the potential for striped bass aquaculture is already being established in preliminary small-scale studies in fresh, brackish, and marine environments. The feasibility of commercializing striped bass at a fairly rapid pace is already established as well, to the extent that a single commercial farm in northern Baja California, Mexico (Pacífico Aquaculture), produces enough fish to consistently supply product to various market outlets, including chain-grocery stores.

The stock assessment data indicating that the Atlantic striped bass fishery is in decline and the policies that are developing as a response further exacerbate the need to establish commercial aquaculture production of striped bass in the United States. Environmentally conscious aquaculture has a number of potential benefits for the striped bass fishery as it can provide economic development and readily available seafood supply to supplement the current, albeit declining, commercial striped bass fishery. The present-day economic and environmental scenario is very similar to the striped bass fishery decline and moratorium of the 1980s that jump-started hybrid striped bass aquaculture (Figure 1.1.c).

Fisheries data indicate potential market value of aquacultured striped bass. The decline in U.S. wild striped bass fishery landings from approximately 3.4 million kg (7.5 million lb) in 2008 to just slightly over 2.3 million kg (5.0 million lb) in 2017 was coincident with an increase in value from about US \$16 million in 2008 to about US \$23 million in 2017, indicating a classical supply and demand relationship. Thus, an underutilized current annual market of 1.1 million kg (2.5 million lb) of striped bass appears available along the Atlantic Coast of the United States alone, which cannot be filled by the presently declining commercial fishery. At present, the average dockside or “off-the-boat” price for whole striped bass in the commercial fishing industry is variable; however, the national average is about US \$10.14 per kg (US \$4.60 per lb) (US \$23 million/2.3 million kg). This suggests a farm gate value of at least US \$8.82–US \$11.02 per kg (US \$4.00–\$5.00 per lb) for aquaculture striped bass, which closely aligns or may be higher than the current farm gate value of cultured hybrid striped bass (US \$8.47 per kg for 0.57–0.91 kg sized fish, or US \$3.84 per lb for 1.25–2.0 lb fish; US \$9.26 per kg for 1.13 kg or larger fish, or US \$4.20 per lb for 2.5 lb or larger fish). Recent retail market prices for striped bass in urbanized areas in North Carolina, New England, and New York ranged from US \$26.45

to US \$41.89 per kg (US \$12.00–US \$19.00 per lb) for boneless, skin-on fillets of wild caught striped bass. Market surveys conducted with Locals Seafood in North Carolina found that marketing value-added, boneless, skin-on fillets of aquacultured striped bass in the mid-Atlantic region is feasible even with a final product price of US \$39.68 per kg (US \$18.00 per lb). Based on these survey data, we estimate the U.S. farm gate value for striped bass can be as low as US \$10.14 per kg (US \$4.60 per lb) and as high as US \$13.23 per kg (US \$6.00 per lb) based on a 50.0% to 70.0% mark-up margin. Furthermore, assessments have shown consumer willingness to pay premium prices for striped bass (Quagraine, 2019). These data show a clear economic and market potential for aquaculture production of striped bass, which already has a wide consumer acceptance and appeal.

Striped bass culture methods and tools

Considerable research on striped bass and its hybrids has been conducted and entire books (Harrell, 1997) and culture method guidelines have been published (Bonn et al., 1976; Harrell et al., 1990). In addition, many studies focusing on striped bass nutrition (Gatlin III, 1997; Small & Soares Jr, 1998; Small, Soares Jr., & Woods III, 2000; Webster & Lovell, 1990; Woods III & Soares Jr., 1996), health (Harms, Sullivan, Hodson, & Stotskopf, 1996; Noga, Kerby, King, Aucoin, & Giesbrecht, 1994; Noga, Wang, Grindem, & Avtalion, 1999; Plumb, 1997; Salger, Reading, & Noga, 2017), pond and recirculating aquaculture system (RAS) culture methodologies (Geiger & Turner, 1990; Geiger, Turner, Fitzmayer, & Nichols, 1985; Harrell, 1997; McGinty & Hodson, 2008; Turano, Borski, & Daniels, 2008), pond fingerling production (Ludwig, 1999, 2004; Ludwig, Perschbacher, & Edziyie, 2010; Ludwig & Tackett, 1991), fingerling production in biofloc production systems (Green, Rawles, Webster, & McEntire,

2018), stress mitigation (Harrell, 1992; Harrell & Moline, 1992; Kenter, Gunn, & Berlinsky, 2019), and high salinity tolerance (Kiilerich, Tipsmark, Borski, & Madsen, 2011; Tipsmark, Luckenbach, Madsen, & Borski, 2007; Tipsmark, Madsen, & Borski, 2004) have been published and are well known. Therefore, the ability to culture these fish in numerous environments on an experimental scale is not in question and, importantly, the wealth of studies conducted on striped bass has allowed for the development of a captively bred, domesticated broodstock (Garber & Sullivan, 2006; Hallerman, 1994).

Domestication

Most cultured fishes in the United States, and the world, originate from wild caught fish or fish that are not domesticated or selectively bred for genetic improvement (Gjedrem & Baranski, 2010; Knibb, 2000; Teletchea & Fontaine, 2014). Domestication is a process by which an organism is taken from its natural environment and then reared in a controlled setting, such as in agriculture. The breeding of these organisms that have been acclimated to and tolerate these culture conditions then produces offspring that are likely to thrive similarly or even better. A domesticated line of striped bass originally obtained from six distinct geographic stocks has been bred in captivity for six generations as part of the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* (Garber & Sullivan, 2006; Hodson et al., 1999; Reading, McGinty, et al., 2018; Woods III, 2001). Except for salmonids, this is the only marine aquaculture finfish species in the United States with an established domestic strain of fish that are available to producers and currently being used for commercial production.

Genetic improvement of finfish broodstock is a critical advancement for aquaculture industry success. Breeding programs provide fish that are selectively bred for optimal culture and

performance traits, such as disease resistance, growth rate and efficiency, acceptance of prepared diets, and tolerance to crowding and stress conditions among many others. Performance gains of domesticated fish can be dramatic in comparison to the wild-origin counterparts. For example, gains in body weight at harvest are estimated to be approximately 14% per generation of selectively bred Atlantic salmon (Gjedrem, 2010). Similarly, domesticated strains of striped bass have been shown to have superior performance for some culture traits (Reading, McGinty, et al., 2018).

There are marked improvements in domestic striped bass growth performance between filial generations captively bred over the last 17 years as evaluated by weight at age. For example, sixth-generation captive-bred domestic striped bass (F6) are about twice the size of third-generation fish (F3) by Age 2 and fifth-generation (F5) female striped bass are about 46% larger than F3 female striped bass at age of 4 years (Figure 1.2). When considering the average improvement in domestic striped bass growth performance for each captive bred generation, we see 33.8% growth gain between F3 and F4, 26.9% growth gain between F4 and F5, and 24.0% growth gain between F5 and F6. These are fish reared in outdoor tanks and pools at semi-commercial density. The timeframe required for domestic striped bass to grow to about 1,000 g (2.20 lb) in these conditions, which is the desired market size for the hybrid striped bass, has been dramatically reduced by 69% through breeding between the F3 and F6 generations (Figure 1.3). Furthermore, the time to grow to the desired market size of 1.36 kg (3 lb), which was identified as a target for white-fleshed marine fish such as striped bass, is about 32 months for F3 generation, 29 months for F4 generation, 28 months for F5 generation, and 24 months for F6 generation fish. Thus, selective breeding has taken the F3 generation fish, which were not economically feasible to grow to this market size over a 32-month timeframe, to within the

economically feasible timeframe of 24 months or less by the F6 generation of breeding. Overall, this is a 75% reduction in the growout time to market obtained through just three generations of selective breeding. The F7 generation of domestic striped bass, first created in 2020, will likely have further improved growth performance over the next 4 or so years (Figure 1.2).

Recent studies have demonstrated that female reproductive potential in the domesticated striped bass is superior to that of equal-sized females captured from the wild using a manual strip spawning method (Locke, Sugg, Sullivan, & Turano, 2013). Additionally, domestic striped bass have an improved dress-out weight (0.5–4.0% increase compared with wild-origin fish (Reading, McGinty, et al., 2018), and, importantly, a 13–25% significantly better feed conversion efficiency ($p < .05$), with feed conversion ratio (FCR) values < 1.1 (Kenter, Kovach, Woods III, Reading, & Berlinsky, 2018). The presumed FCR for striped bass raised at commercial density is approximately 1.5 or slightly higher. Collectively, this domestic striped bass broodstock program has produced a fish suitable for commercial growout economics. However, the use of wild-origin striped bass stocks may be critical for offshore culture in some regions due to escapement concerns (e.g., Northeast Atlantic and Gulf of Mexico), and as such, it is important to extend reproduction and larviculture technology of these fish to those regions as appropriate for the U.S. striped bass aquaculture industry to thrive.

Reproduction and larviculture

A major constraint to the culture of any marine fish species is the complexity of larval rearing and ability to produce a reliable source of juveniles for culture (Planas & Cunha, 1999). However, this bottleneck has already been addressed in the culture of striped bass, which have comparatively simple requirements for larviculture and are similar to that of salmonids, one of

the only successful marine finfish aquaculture industries in the United States. Early research was conducted to understand the female striped bass reproductive cycle (Berlinsky & Specker, 1991; Swanson & Sullivan, 1991; Tao, Hara, Hodson, Woods III, & Sullivan, 1993; Woods III & Sullivan, 1993) and endocrine events that occur during ovary maturation (King et al. 1994,b; Zohar, 1989). These studies were followed by others that employed environmental (temperature, photoperiod) manipulation to phase-shift the reproductive cycle in order to induce out-of-season spawning (Blythe et al. 1994,b; Clark, Henderson-Arzapalo, & Sullivan, 2005) and to better understand egg quality and reproductive performance (Chapman, Reading, & Sullivan, 2014; Reading et al., 2014, 2018; Reading, Hiramatsu, & Sullivan, 2011; Reading, Williams, Chapman, Islam Williams, & Sullivan, 2013; Schilling et al., 2014; Sullivan, Chapman, Reading, & Anderson, 2015; Williams et al. 2014,b). Both hormone-induced and nonhormone-induced tank spawning of striped bass has been achieved (Andersen et al., 2021; Smith & Whitehurst, 1990; Woods III, Woiwode, McCarthy, Theisen, & Bennett, 1990), although considerably more attention has been focused on inducing ovulation and manual strip spawning for in vitro fertilization (Hodson & Sullivan, 1993; Mylonas et al., 1993; Woods III & Sullivan, 1993). We recently described commercially scalable methods to batch spawn domestic striped bass en masse in tanks without any hormone applications, and these procedures can be used to produce many millions of larvae necessary for commercial production (Andersen et al., 2021; Reading et al., 2016, 2018b, 2018c, 2018d, 2019). Sperm cryopreservation and storage has also been characterized (Frankel, Theisen, Guthrie, Welch, & Woods III, 2013; He & Woods III, 2004; Jenkins-Keeran & Woods III, 2002; Woods III et al., 2018). Additionally, the osmoregulatory apparatus that enables striped bass to be euryhaline is highly geared for life in seawater even as a resident in freshwater environments (Kiilerich et al., 2011; Tipsmark et al., 2004). As such,

striped bass larvae are tolerant to half-strength seawater as early as 1 day post-hatch (dph), and growth and survival of 5 dph larvae raised in 20, 40, and 60% seawater was found to be as great as those raised in freshwater (Lal, Lasker, & Kuljis, 1977).

Larval striped bass can be raised to fingerlings at a commercial scale in earthen ponds using natural productivity through fertilization (Harrell, 1997; Ludwig, 1999). This infrastructure is currently in place at many aquaculture operations utilizing pond systems, in particular at commercial hybrid striped bass fingerling operations. Pond sizes for larviculture are typically smaller than for growout and therefore not available at all commercial hybrid striped bass rearing facilities. Intensive larval rearing for fingerling production in tank systems is generally constrained to the use of live feeds, and challenges are not as well described as compared to other life stages. Further research on commercially scalable methods of intensive larval rearing is needed and currently being conducted. Collectively, the larval and juvenile seed-stock supply for striped bass is presently achievable at commercial scale in the United States.

Rearing and growout

Striped bass have been shown to adapt well to and exhibit high survival in both RAS technologies and cages. Laboratory-scale RAS studies show that striped bass exhibit equivalent growth performance in freshwater, brackish, and saltwater environments (Kenter et al., 2018). Experimental-scale studies of striped bass in cage culture show that fish grow better than hybrid striped bass in brackish water with little impact on survivorship (Woods, Kerby, & Huish, 1983). Significant progress has been made on growth biology in striped bass including seasonally based feeding protocols; characterization of growout temperature (Harrell, 1992); demonstration that a

range of salinities are equally effective in regulating growth (Harrell, 1992; Kenter et al., 2018); nutrient requirements, endocrine and growth physiology (Picha et al., 2009, 2014; Picha, Turano, Beckman, & Borski, 2008; Picha, Turano, Tipsmark, & Borski, 2008; Won & Borski, 2013); and experimental scale studies suggesting a potential for culture of 1.36–2.27 kg (3.00–5.00 lb) fish. However, none of this research has provided insight into commercial scaling or use of stocking densities typical of intensive culture requirements or economic analyses for the full production cycle from egg to plate of domestic or wild striped bass. Currently, data suggest that striped bass can be grown in cages and under RAS at different salinities. Despite this work, one major constraint has been a lack of demonstration that striped bass can be economically cultured at commercial scale.

Genomic resources and tools

The striped bass is a priority species for the United States Department of Agriculture (USDA) National Animal Genome Research Support Program (NRSP-8) and as such considerable progress in establishing genomic resources for striped bass has been accomplished. The striped bass genome assembly was recently updated (2019) through a combinatorial approach of short-read sequencing (Illumina, San Diego, CA), long-read sequencing (Pacific Biosciences, Menlo Park, CA), and Chicago® and Dovetail™ Hi-C + HiRise™ scaffolding (Dovetail Genomics, Scotts Valley, CA). This genome assembly (NCSU_SB_2.0) is publicly available under GenBank accession GCA_004916995.1 and is currently in the annotation pipeline. This genome assembly has a total sequence length of 598,109,5477 base pairs and consists of 629 scaffolds (Abdelrahman et al., 2017; Andersen, Baltzegar, Fuller, Abernathy, & Reading, 2019; Reading, McGinty, et al., 2018).

Other genomic resources available for striped bass include a medium-density genetic linkage map of 289 polymorphic microsatellite DNA markers (Liu et al., 2012), 23,000 unigene sequences from a multi-tissue transcriptome (Li, Beck, Fuller, & Peatman, 2014; GenBank accession GBAA000000000), and a well-annotated transcriptome of 11,200 unigene sequences derived from ovary representative of all stages of oocyte growth and maturation (Reading et al., 2012; GenBank accession SRX007394). A number of studies have reported the development of microsatellite DNA markers for striped bass (Brown, Baltazar, & Hamilton, 2005; Couch et al., 2006; Han, Li, Leclerc, Hays, & Ely, 2000; Rexroad et al., 2006; Skalski, Couch, Garber, Weir, & Sullivan, 2006). Epigenetic studies on striped bass are limited to sperm methylation profiles and their correlation to fertility (Woods III et al., 2018). Additional resources are also available for closely related Moronids including a reference genome sequence assembly for white bass (Abernathy et al., 2019), multi-tissue transcriptome of 22,000 unigene sequences for white bass (Li et al., 2014; GenBank accession GAZY000000000), and 1,730 unigene sequences for white perch (*Morone americana*) (Schilling et al., 2014; GenBank accession GAQS000000000).

These resources collectively provide excellent tools for selective breeding, marker-assisted selection, and domestication, as well as for functional studies on the biology and aquaculture of striped bass. For example, the genome and transcriptome data empower proteomic analyses (Andersen et al., 2019; Reading et al., 2012, 2013; Schilling et al., 2014, 2015; Williams et al., 2014).

Future directions and challenges

Striped bass is an aquaculture species that is well positioned for commercial production. The current extent of consumer visibility, established market size and product price-point,

knowledge of the biology and culture, and infrastructure for commercial seed production and rearing of striped bass all support the likelihood of its success as an aquaculture industry. Furthermore, the fish is euryhaline, which means it can be reared in fresh, brackish, or marine water in both coastal and inland systems throughout the United States. Culture methods of striped bass are well established and therefore no major hurdles remain regarding the technology to produce the fish. Recent efforts have established a reliable hatchery larval production system, which in the past has been considered a bottleneck to commercial-scale production (McCraren, 1984). One of the only limitations to developing a striped bass industry is the lack of current commercial U.S. producers and data to support the economic viability of commercial production.

Barriers and opportunities

Significant barriers remain primarily the full commercial-scale demonstration and detailed economics of production and marketing to show that striped bass aquaculture is solvent. Development and expansion of the striped bass aquaculture industry in the United States has great potential if the following conditions are addressed:

1. Identifying domestic producers for commercial production and providing adequate fish to consistently supply seafood markets;
2. Demonstrating profitability through production, marketing, processing, and economics;
3. Clarification and general reduction of regulatory permitting and licensing procedures; and
4. Promoting comprehensive extension education, technical training, marketing, and product visibility to consumers and stakeholders.

The establishment of a conglomerate group of stakeholders and partners would enable a centralized demonstration of the technologies and outreach necessary to commercialize striped bass production. This would include items such as demonstrating the culture of adequate volumes of fish for commercialization and marketing using diverse aquaculture systems (pond, cage, and RAS or combinations thereof), development of business models for demonstrating profitability, and establishing extension activities to disseminate this information. Current finfish aquaculture infrastructure exists that can provide support for producing and marketing striped bass at commercial scale. Collaboration with social scientists and seafood distributors to better understand seafood marketing, consumer preferences, market depth, supply and demand, retail pricing, and the provision of additional outreach about striped bass aquaculture are also crucial (Pigg & Reading, 2018; Ryan et al., 2018).

Venture capital investment will be required for the next phase of industry development and upscaling once commercial striped bass production and marketing has been demonstrated. Additionally, engagement in extensive outreach, including extension programming and technology transfer, is required to provide the necessary aquaculture and marketing training tools to support the growth of this industry. This would include working with state NOAA Sea Grant programs along with the USDA and state cooperative extension agents at Land Grant Universities in the region to address social, behavioral, economic, and policy priorities associated with striped bass aquaculture.

Clarity on policies relevant to commercial aquaculture (e.g., production, product transport) is also imperative to the development of a striped bass aquaculture industry. Presidential Executive Order 13921, 2020 (“Promoting American Seafood Competitiveness and Economic Growth”) was issued in 2020 with the intent to:

“...improve the competitiveness of American industry; ensure food security; provide environmentally safe and sustainable seafood; support American workers; ensure coordinated, predictable, and transparent Federal actions; and remove unnecessary regulatory burdens.”

Among the specific actions outlined in the order to achieve these goals is the renewal of a:

“...focus on long-term strategic planning to facilitate aquaculture projects, we can protect our aquatic environments; revitalize our Nation's seafood industry; get more Americans back to work; and put healthy, safe food on our families' tables.”

Several of the legislative hurdles hindering development of the U.S. marine aquaculture industry are addressed by the executive order, such as requiring environmental reviews of aquaculture projects to be completed within 2 years.

Establishing the StriperHub

The *StriperHub* is a Sea Grant-supported network that formed to facilitate striped bass aquaculture. The aim of the hub is to overcome barriers to industry development and expansion through demonstration and promotion of commercial-level culture, economics, and marketing of U.S. striped bass. North Carolina Sea Grant is leading the initiative and coordinating the *StriperHub* network, which is made up of several Sea Grant programs, USDA and other federal scientists, industry partners, and university researchers focused on consolidating and streamlining commercialization efforts in various culture environments. Detailed analyses of economics and marketing, baseline farm gate value and market depth, estimations of production economics, and demonstration of the potential for commercial culture scaling necessary for adoption and growth of the commercial striped bass aquaculture industry are some of the

priorities of the *StriperHub*. The *StriperHub* has been active since 2020 in organizing project meetings, developing a web presence, creating recipes, and conducting research. Activities to date have resulted in successful commercial aquaculture production and the first farmed domestic striped bass will be available in U.S. markets in 2021.

In addition to being identified as a candidate aquaculture species by the NOAA, establishing a commercial striped bass aquaculture industry relies on the continued efforts of stakeholders, scientists, legislators, policymakers and their institutions in conducting research, performing assessments, developing business models and marketing strategies, and adopting clear permitting and licensing procedures for producers and vendors. These goals will be best realized if all of these stakeholder groups are able to synergize in a coordinated effort to serve as a nexus for information that can be disseminated to commercial producers and the public through the additional avenues of communication, outreach, education, and extension created through the *StriperHub*.

Acknowledgements

The authors thank Dr. Craig V. Sullivan for his longstanding contributions to the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry*. This work was supported by funding provided from the following sources: The National Oceanic and Atmospheric Administration (NOAA) and National Sea Grant (E/2019-AQUA-02, a project to establish Striped Bass Aquaculture Hub, the *StriperHub*), North Carolina Sea Grant, the United States Department of Agriculture (USDA), National Institute of Food and Agriculture (NIFA), Foundation for Food and Agriculture Research New Innovator Award (FFAR), and North Carolina Sea Grant (R/MG-1411). Striped bass is a priority species for the

USDA National Research Support Project 8 (NRSP-8; National Animal Genome Research Project) and funding to support the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* was provided by the NRSP-8 and the USDA NIFA (Hatch Multistate Project), the USDA Agricultural Research Service (ARS, Harry K. Dupree Stuttgart National Aquaculture Research Center), the North Carolina State University College of Agriculture and Life Sciences and College of Sciences, the North Carolina Agricultural Foundation William White Endowment, and various industry stakeholders including Premex, ADM, Clay Chappell, Locals Seafood, and several others who wish to remain anonymous. This is publication number 122 from the North Carolina State University Pamlico Aquaculture Field Laboratory.

The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author and do not necessarily reflect the views of the United States Department of Commerce or the National Oceanic and Atmospheric Administration. The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the USDOC, NOAA, USDA, or USDA-ARS.

References

- Abdelrahman, H., El Hady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., Beck, B., Blackburn, H., Bosworth, B., Buchanan, J., Chappell, J., Daniels, W., Dong, S., Dunham, R., Durland, E., Elaswad, A., Gomez-Chiarri, M.*, Gosh, K., Guo, X.*, Hackett, P., Hanson, T., Hedgecock, D., Howard, T., Holland, L., Jackson, M., Jin, Y., Khalil, K., Kocher, T.*, Leeds, T., Li, N., Lindsey, L., Liu, S., Liu, Z.*, Martin, K., Novriadi, R., Odin, R., Olds, B., Palti, Y.*, Peatman, E., Proestou, D.*, Qin, G., Reading, B.J.*, Rexroad, C.*, Roberts, S.*, Rye, M., Salem, M.*, Severin, A., Shi, H., Shoemaker, C., Stiles, S., Tan, S., Tang, K.F.J., Thongda, W., Tiersch, T., Tomasso, J., Tri Prabowo, W., Vallejo, R., van der Steen, H., Vo, K., Waldbieser, G., Wang, H., Wang, X., Xiang, J., Yang, Y., Yant, R., Yuan, Z., Zeng, Q., and Zhou, T. on behalf of the The Aquaculture Genomics, Genetics and Breeding Workshop. 2017. Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. *BMC Genomics*, 18:191. DOI: 10.1186/s12864-017-3557-1.
- Abernathy, J., Andersen, L.K., Rawles, S., Schilling, J., Fuller, A., McEntire, M., Reading, B.J., Beck, B.H., and Peatman, E. 2019. Development and enhancement of white bass *Morone chrysops* resources for genetic improvement of hybrid striped bass. In Abstracts of Aquaculture Triennial 2019: The Big Easy Choice! (World Aquaculture Society). March 7-11. New Orleans, Louisiana, USA.
- Andersen, L.K., Baltzegar, D.A., Fuller, A., Abernathy, J., and Reading, B.J. 2019. The striped bass *Morone saxatilis* genome sequence assembly. In Abstracts of Aquaculture Triennial 2019: The Big Easy Choice! (World Aquaculture Society). March 7-11. New Orleans, Louisiana, USA.

- Andersen, L.K., Clark, R.W., McGinty, A.S., Hopper, M.S., Kenter, L.W., Salger, S.A., Schilling, J., Hodson, R.G., Kovach, A.I., Berlinsky, D.L. and Reading, B.J. 2021a. Volitional tank spawning of domestic striped bass (*Morone saxatilis*) using human chorionic gonadotropin (hCG) and gonadotropin releasing hormone analogue (GnRHa)-induced 'pace-setting' females. *Aquaculture*, 532, p.735967. DOI: 10.1016/j.aquaculture.2020.735967.
- Andersen, L.K., Clark, R.W., Hopper, M.S., Hodson, R.G., Schilling, J., Daniels, H.V., Woods III, L.C., Kovach, A.I., Berlinsky, D.L., Kenter, L.W. and McGinty, A.S. 2021b. Methods of domestic striped bass (*Morone saxatilis*) spawning that do not require the use of any hormone induction. *Aquaculture*, p.736025. DOI: 10.1016/j.aquaculture.2020.736025.
- Anderson, J.C. 1966. Production of striped bass fingerlings. *Progressive Fish-Culturist*, 28:162164.22. DOI: 10.1577/1548-8640(1966)28[162:POSBF]2.0.CO;2.
- ASMFC (Atlantic States Marine Fisheries Council). 1981. Interstate fisheries management plan for the striped bass of the Atlantic Coast from Maine to North Carolina. *Fisheries Management Report No. 1*.
- Baird, S.F. 1874. Report of the commissioner. Pages I-XCII, In: Report of the Commissioner for 1872 and 1873. U.S. Commission of Fish and Fisheries, Washington, D.C.
- Berlinsky, D.L. and Specker, J.L. 1991. Changes in gonadal hormones during oocyte development in the striped bass, *Morone saxatilis*. *Fish Physiology and Biochemistry*, 9:51-62. DOI: 10.1007/BF01987611.

- Blythe, W.G., Helfrich, L.A., and Libey, G. 1994a. Induced maturation of striped bass *Morone saxatilis* exposed to 6, 9, and 12 month photothermal regimes. *Journal of the World Aquaculture Society*, 25:183-192. DOI: 10.1111/j.1749-7345.1994.tb00180.x.
- Blythe, W.G., Helfrich, L.A., and Sullivan, C.V. 1994b. Sex steroid hormone and vitellogenin levels in striped bass (*Morone saxatilis*) maturing under 6-, 9-, and 12-month photothermal cycles. *General and Comparative Endocrinology*, 94:122-134. DOI: 10.1006/gcen.1994.1066.
- Bonn, E.W., Bailey, W.M., and Bayless, J.D. (Eds.). 1976. Guidelines for striped bass culture. Southern Division, American Fisheries Society. Bethesda, Maryland, USA.
- Brown, K.M., Baltazar, G.A., and Hamilton, M.B. 2005. Reconciling nuclear microsatellite and mitochondrial marker estimates of population structure: breeding population structure of Chesapeake Bay striped bass (*Morone saxatilis*). *Heredity*, 94(6):606-615. DOI: 10.1038/sj.hdy.6800668.
- Callihan, J.L., Harris, J.E., and Hightower, J.E. 2015. Coastal migration and homing of Roanoke River striped bass. *Marine and Coastal Fisheries*, 7(1):301-315. DOI: 10.1080/19425120.2015.1057309.
- Chapman, R.W., Reading, B.J., and Sullivan, C.V. 2014. Ovary transcriptome profiling via artificial intelligence reveals a transcriptomic fingerprint predicting egg quality in striped bass, *Morone saxatilis*. *PLoS ONE*, 9(5):e96818. DOI: 10.1371/journal.pone.0096818.
- Clark, R.W., Henderson-Arzapalo, A., and Sullivan, C.V. 2005. Disparate effects of constant and annually-cycling daylength and water temperature on reproductive maturation of striped bass (*Morone saxatilis*). *Aquaculture*, 249:497-513. DOI: 10.1016/j.aquaculture.2005.04.001.

Couch, C.R., Garber, A.F., Rexroad III, C.E., Abrams, J.M., Stannard, J.A., Westerman, M.E., and Sullivan, C.V. 2006. Isolation and characterization of 149 novel microsatellite DNA markers for striped bass, *Morone saxatilis*, and cross-species amplification in white bass, *Morone chrysops*, and their hybrid. *Molecular Ecology Notes*, 6(3):667-669. DOI: 10.1111/j.1471-8286.2006.01292.x.

Executive Order No. 13921, 85 Fed. Reg. 28471 (May 7, 2020).

FAO (Food and Agriculture Organization of the United Nations). 2018. The State of World Fisheries and Aquaculture (SOFIA) – Meeting the Sustainable Development Goals. Food and Agriculture Organization, Rome, Italy.

Frankel, T.E., Theisen, D.D., Guthrie, H.D., Welch, G.R., and Woods III, L.C. 2013. The effect of freezing rate on the quality of striped bass sperm. *Theriogenology*, 79(6):940-945. DOI: 10.1016/j.theriogenology.2013.01.009.

Froehlich, H.E., Gentry, R.R., and Halpern, B.S. 2018. Global change in marine aquaculture production potential under climate change. *Nature Ecology & Evolution*, 2(11):1745-1750. DOI: 10.1038/s41559-018-0669-1.

Garber, A.F. and Sullivan, C.V. 2006. Selective breeding for the hybrid striped bass (*Morone chrysops*, Rafinesque x *M. saxatilis*, Walbaum) industry: status and perspectives. *Aquaculture Research*, 37:319-338. DOI: 10.1111/j.1365-2109.2005.01439.x.

Gatlin III, D.M. 1997. Nutrition and feeding of striped bass and hybrid striped bass. Pages 235-252 in: R.M. Harrell, (Ed.). *Striped bass and other Morone culture*. Elsevier Science B.V., Amsterdam, The Netherlands.

- Geiger, J.G. and Parker, N.C. 1985. Survey of striped bass hatchery management in the southeastern United States. *The Progressive Fish-Culturist*, 47(1):1-13. DOI: 10.1577/1548-8640(1985)47<1:SOSBHM>2.0.CO;2.
- Geiger, J.G. and Turner, C.J. 1990. Pond fertilization and zooplankton management techniques for production of fingerling striped bass and hybrid striped bass. Pages 79-98 in: R.M. Harrell, J.H. Kerby, and R.V. Minton (Eds.). *Culture and propagation of striped bass and its hybrids*. Striped Bass Committee, Southern Division, American Fisheries Society, Bethesda, Maryland, USA.
- Geiger, J.G., Turner, C.J., Fitzmayer, K.M., and Nichols, W.C. 1985. Feeding habits of larval and fingerling striped bass and zooplankton dynamics in fertilized rearing ponds. *The Progressive Fish-Culturist*, 47:213-223.23. DOI: 10.1577/1548-8640(1985)47<213:FHOLAF>2.0.CO;2.
- Gjedrem, T. 2010. The first family-based breeding program in aquaculture. *Reviews in Aquaculture*, 2(1):2-15. DOI: 10.1111/j.1753-5131.2010.01011.x.
- Gjedrem, T. and Baranski, M. 2010. *Selective breeding in aquaculture: an introduction* (Vol. 10). Springer Science & Business Media, Berlin, Germany.
- Green, B.W., Rawles, S.D., Webster, C.D., and McEntire, M.E. 2018. Effect of stocking rate on growing juvenile Sunshine bass, *Morone chrysops* x *M. saxatilis*, in an outdoor biofloc production system. *Journal of the World Aquaculture Society*, 49:872-836. DOI: 10.1111/jwas.12491.
- Haeseker, S.L., Carmichael, J.T., and Hightower, J.E. 1996. Summer distribution and condition of striped bass within Albemarle Sound, North Carolina. *Transactions of the American*

- Fisheries Society, 125(5):690-704. DOI: 10.1577/1548-8659(1996)125<0690:SDACOS>2.3.CO;2.
- Hallerman, E.M. 1994. Toward coordination and funding of long-term genetic improvement programs for striped and hybrid bass *Morone* sp. Journal of the World Aquaculture Society, 25(3):360-365. DOI: 10.1111/j.1749-7345.1994.tb00219.x.
- Han, K., Li, L., Leclerc, G.M., Hays, A.M., and Ely, B. 2000. Isolation and characterization of microsatellite loci for striped bass (*Morone saxatilis*). Marine Biotechnology, 2(5):405-408. DOI: 10.1007/s101260000014.
- Harms, C.A., Sullivan, C.V., Hodson, R.G., and Stotskopf, M.K. 1996. Clinical pathology and histopathology of net-stressed striped bass with “red tail”. Journal of Aquatic Animal Health, 8:82-86. DOI: 10.1577/1548-8667(1996)008<0082:CPAHCO>2.3.CO;2.
- Harrell, R.M. 1992. Stress mitigation by use of salt and anesthetic for wild striped bass captured for brood stock. The Progressive Fish-Culturist, 54:228-233. DOI: 10.1577/1548-8640(1992)054<0228:SMBUOS>2.3.CO;2.
- Harrell, R.M. (Ed.). 1997. Striped bass and other *Morone* culture. Elsevier Science B.V., Amsterdam, The Netherlands.
- Harrell, R.M., Kerby, J.H., and Minton, R.V. 1990. Striped bass and hybrid striped bass culture: the next twenty-five years. Pages 253-261 in: R.M. Harrell, J.H. Kerby, and R.V. Minton (Eds.). Culture and propagation of striped bass and its hybrids. Striped Bass Committee, Southern Division, American Fisheries Society, Bethesda, Maryland, USA.
- Harrell, R.M. and Moline, M.A. 1992. Comparative stress dynamics of brood stock striped bass *Morone saxatilis* associated with two capture techniques. Aquaculture, 23:58-63. DOI: 10.1111/j.1749-7345.1992.tb00751.x.

- He, S. and Woods III, L.C. 2004. Changes in motility, ultrastructure, and fertilization capacity of striped bass *Morone saxatilis* spermatozoa following cryopreservation. *Aquaculture*, 236(1-4):677-686. DOI: 10.1016/j.aquaculture.2004.02.029.
- Hill, J., Evans, J.W., and Van Den Avyle, M.J. 1989. Species profiles: life histories and environmental requirements of coastal fishes and invertebrates (South Atlantic): striped bass. Georgia Cooperative Fishery and Wildlife Research Unit Athens. Fish and Wildlife Service, U.S. Department of the Interior, Washington, DC. Biological Report 82(11.118).
- Hodson, R.G. 1990. Hybrid striped bass biology and life history. Texas Agricultural Extension Service publication no. 2416.
- Hodson, R.G., Clark, R.W., Hopper, M.S., McGinty, A.S., Weber, G.M., and Sullivan, C.V. 1999. Reproduction of domesticated striped bass: commercial mass production of fingerlings. *UJNR Aquaculture. 28th Panel Proceedings*, 23-32.
- Hodson, R.G. and Hayes, M. 1990. Hybrid striped bass pond production of foodfish. Texas Agricultural Extension Service publication no. 2411.
- Hodson, R.G. and Sullivan, C.V. 1993. Induced maturation and spawning of domestic and wild striped bass (*Morone saxatilis*) broodstock with implanted GnRH analogue and injected hCG. *Journal of Aquaculture and Fisheries Management*, 24:271-280. DOI: 10.1111/j.1365-2109.1993.tb00562.x.
- Jenkins-Keeran, K. and Woods III, L.C. 2002. The cryopreservation of striped bass *Morone saxatilis* semen. *Journal of the World Aquaculture Society*, 33(1):70-77. DOI: 10.1111/j.1749-7345.2002.tb00480.x.

- Kenter, L.W., Gunn, M.A., and Berlinsky, D.L. 2019. Transport stress mitigation and the effects of preanesthesia on striped bass. *North American Journal of Aquaculture*, 81(1):67-73. DOI: 10.1002/naaq.10072.
- Kenter, L.W., Kovach, A.I., Woods III, L.C., Reading, B.J., and Berlinsky, D.L. 2018. Strain evaluation of striped bass (*Morone saxatilis*) cultured at different salinities. *Aquaculture*, 492:215-225. DOI: 10.1016/j.aquaculture.2018.04.017.
- Kiilerich, P., Tipsmark, C.K., Borski, R.J., and Madsen, S.S. 2011. Differential effects of cortisol and 11-deoxycorticosterone on ion transport protein mRNA levels in gills of two euryhaline teleosts, Mozambique tilapia (*Oreochromis mossambicus*) and striped bass (*Morone saxatilis*). *Journal of Endocrinology*, 209(1):115. DOI: 10.1530/JOE-10-0326.
- King, W.V., Thomas, P., Harrell, R.M., Hodson, R.G., and Sullivan, C.V. 1994a. Plasma levels of gonadal steroids during final oocyte maturation of striped bass, *Morone saxatilis*. *General and Comparative Endocrinology*, 95:178-191. DOI: 10.1006/gcen.1994.1115.
- King, W.V., Thomas, P., and Sullivan, C.V. 1994b. Hormonal regulation of final maturation of striped bass oocytes in vitro. *General and Comparative Endocrinology*, 96:223-233. DOI: 10.1006/gcen.1994.1177.
- Kinman, B.T. 1988. Evaluation of striped bass introductions in Lake Cumberland. Kentucky Department of Fish and Wildlife Resources bulletin no. 83.
- Knibb, W. 2000. Genetic improvement of marine fish- which method for industry? *Aquaculture Research*, 31:11-23. DOI: 10.1046/j.1365-2109.2000.00393.x.
- Lal, K., Lasker, R., and Kuljis, A. 1977. Acclimation and rearing of striped bass larvae in sea water. *California Fish and Game*, 63:210-218.

- LeBlanc, N.M., Gahagan, B.I., Andrews, S.N., Avery, T.S., Puncher, G.N., Reading, B.J., Buhariwalla, C.F., Curry, R.A., Whiteley, A.R., and Pavey, S.A. 2020. Genomic population structure of striped bass (*Morone saxatilis*) from the Gulf of St. Lawrence to Cape Fear River. *Evolutionary Applications*, In Press. DOI: 10.1111/eva.12990.
- Lem, A., Bjorndal, T., and Lappo, A. 2014. Economic analysis of supply and demand for food up to 2030– Special focus on fish and fishery products. Food and Agriculture Organization (FAO) Fisheries and Aquaculture Circular. No. 1089. Rome, Italy.
- Li, C., Beck, B.H., Fuller, S.A., and Peatman, E. 2014. Transcriptome annotation and marker discovery in white bass (*Morone chrysops*) and striped bass (*Morone saxatilis*). *Animal Genetics*, 45(6):885-887. DOI: 10.1111/age.12211.
- Liu, S., Rexroad, C.E., Couch, C.R., Cordes, J.F., Reece, K.S., and Sullivan, C.V. 2012. A microsatellite linkage map of striped bass (*Morone saxatilis*) reveals conserved synteny with the three-spined stickleback (*Gasterosteus aculeatus*). *Marine Biotechnology*, 14(2):237-244. DOI: 10.1007/s10126-011-9407-2.
- Locke, S.H., Sugg, N., Sullivan, C.V., and Turano, M.J. 2013. Domesticated broodstock for hybrid striped bass farming: pioneering industry implementation. Final Report North Carolina Sea Grant Project no. 11-AM-07.24.
- Ludwig, G.M. 1999. Zooplankton succession and larval fish culture in freshwater ponds. Southern Regional Aquaculture Center (SRAC) publication no. 700.
- Ludwig, G.M. 2004. Hybrid striped bass: fingerling production in ponds. Southern Regional Aquaculture Center (SRAC) publication no. 302.
- Ludwig, G.M., Perschbacher, P., and Edziyie, R. 2010. The effect of the dye Aquashade® on water quality, phytoplankton, zooplankton, and sunshine bass, *Morone chrysops* × *M.*

- saxatilis*, fingerling production in fertilized culture ponds. Journal of the World Aquaculture Society, 41:40-48. DOI: 10.1111/j.1749-7345.2009.00331.x.
- Ludwig, G.M. and Tackett, D.L. 1991. Effects of using rice bran and cottonseed meal as organic fertilizers on water quality, plankton, and growth and yield of striped bass, *Morone saxatilis*, fingerlings in ponds. Journal of Applied Aquaculture, 1(1):79-94. DOI: 10.1300/J028v01n01_07.
- McCraren, J.P. (Ed.). 1984. The aquaculture of striped bass: a proceedings. University of Maryland Sea Grant Publication, College Park, Maryland, USA.
- McGinty, A.S. and Hodson, R.G., 2008. Hybrid striped bass: hatchery phase. Southern Regional Aquaculture Center (SRAC) publication no. 301.
- Mischke, C.C. (Ed.). 2012. Aquaculture pond fertilization: impacts of nutrient input on production. John Wiley & Sons, Hoboken, New Jersey, USA.
- Mylonas, C.C., Swenson, P., Woods III, L.C., Jonsson, E., Jonsson, J., Stefansson, S., and Zohar, Y. 1993. GnRHa-induced ovulation and sperm production of striped bass, Atlantic and Pacific salmon using controlled release devices. World Aquaculture Society/European Aquaculture Society Special Publication, 19:418.
- NFI (National Fisheries Institute). 2018. Top 10 List for Seafood Consumption. National Fisheries Institute U.S. Per-Capita Consumption By Species in Pounds. Available at: <https://www.aboutseafood.com/about/top-ten-list-for-seafood-consumption/>.
- NMFS (National Marine Fisheries Service). 2019. Imports and Exports of Fishery Products Annual Summary, 2018. U.S. Department of Commerce, NOAA Current Fishery Statistics No. 2018-2. Available at: <https://www.st.nmfs.noaa.gov/Assets/commercial/trade/Trade2018.pdf>.

- NMFS (National Marine Fisheries Service). 2020 Fisheries of the United States, 2018. U.S. Department of Commerce, NOAA Current Fishery Statistics No. 2018 Available at: <https://www.fisheries.noaa.gov/national/commercial-fishing/fisheries-united-states-2018>.
- NOAA (National Oceanic and Atmospheric Administration). 2020a. Alaska Pollock. U.S. Department of Commerce, NOAA FishWatch U.S. Seafood Facts. Available at: <https://www.fishwatch.gov/profiles/alaska-pollock>.
- NOAA (National Oceanic and Atmospheric Administration). 2020b. Atlantic Cod. U.S. Department of Commerce, NOAA FishWatch U.S. Seafood Facts. Available at: <https://www.fishwatch.gov/profiles/atlantic-cod>.
- NOAA (National Oceanic and Atmospheric Administration). 2020c. Atlantic Striped Bass. U.S. Department of Commerce, NOAA FishWatch U.S. Seafood Facts. Available at: <https://www.fishwatch.gov/profiles/atlantic-striped-bass>.
- Noga, E.J., Kerby, J.H., King, W., Aucoin, D.P., and Giesbrecht, F., 1994. Quantitative comparison of the stress response of striped bass (*Morone saxatilis*) and hybrid striped bass (*Morone saxatilis* x *Morone chrysops* and *Morone saxatilis* x *Morone americana*). American journal of veterinary research, 55(3):405-409.
- Noga, E.J., Wang, C., Grindem, C.B., and Avtalion, R., 1999. Comparative clinicopathological responses of striped bass and palmetto bass to acute stress. Transactions of the American Fisheries Society, 128(4):680-686. DOI: 10.1577/1548-8659(1999)128<0680:CCROSB>2.0.CO;2.
- Parks, N.B. 1978. The Pacific Northwest Commercial Fishery for Striped Bass, 1922-74. Marine Fisheries Review, 40(1):18-20.

- Picha, M.E., Biga, P.R., Galt, N., McGinty, A.S., Gross, K., Hedgpeth, V.S., Siopes, T.D., and Borski, R.J. 2014. Overcompensation of circulating and local insulin-like growth factor-1 during catch-up growth in hybrid striped bass (*Morone chrysops* × *Morone saxatilis*) following temperature and feeding manipulations. *Aquaculture*, 428:174-183. DOI: 10.1016/j.aquaculture.2014.02.028.
- Picha, M.E., Strom, C.N., Riley, L.G., Walker, A.A., Won, E.T., Johnstone, W.M., and Borski, R.J. 2009. Plasma ghrelin and growth hormone regulation in response to metabolic state in hybrid striped bass: effects of feeding, ghrelin and insulin-like growth factor-I on in vivo and in vitro GH secretion. *General and Comparative Endocrinology*, 161(3):365-372. DOI: 10.1016/j.ygcen.2009.01.026.
- Picha, M.E., Turano, M.J., Beckman, B.R., and Borski, R.J. 2008a. Endocrine biomarkers of growth and applications to aquaculture: a minireview of growth hormone, insulin-like growth factor (IGF)-I, and IGF-binding proteins as potential growth indicators in fish. *North American Journal of Aquaculture*, 70(2):196-211. DOI: 10.1577/A07-038.1.
- Picha, M.E., Turano, M.J., Tipsmark, C.K., and Borski, R.J. 2008b. Regulation of endocrine and paracrine sources of insulin-like growth factors and growth hormone receptor during compensatory growth in hybrid striped bass (*Morone chrysops* × *Morone saxatilis*). *Journal of Endocrinology*, 199:81-94. DOI: 10.1677/JOE-07-0649.
- Pigg, S. and Reading, B.J. 2018. Knowing bass: Accounting for information environments in designing online public outreach. *Open Library of Humanities*, 4(2). DOI: 10.16995/olh.377.
- Piper, R.G. 1982. Fish hatchery management (No. 2175). US Department of the Interior, Fish and Wildlife Service.

- Planas, M. and Cunha, I. 1999. Larviculture of marine fish: problems and perspectives. *Aquaculture*, 177(1-4):171-190. DOI: 10.1016/S0044-8486(99)00079-4.
- Plumb, J.A. 1997. Infectious diseases of striped bass. Pages 271-314 in: R.M. Harrell, (Ed.). *Striped bass and other Morone culture*. Elsevier Science B.V., Amsterdam, The Netherlands.
- Quagraine, K.K. 2019. Consumer Willingness to Pay for a Saline Fish Species Grown in the US Midwest: The Case of Striped Bass, *Morone saxatilis*. *Journal of the World Aquaculture Society*, 50(1), pp.163-171. DOI: 10.1111/jwas.12464.
- Reading, B.J. 2017. Marine Fish Aquaculture: Selective Breeding and Reproduction. Invited Seminar, Marine Fish Aquaculture Scoping Workshop and development of the Marine Aquaculture Survey. March 23-24. Harbor Branch Oceanographic Institute, Florida Atlantic University, Fort Pierce, Florida, USA.
- Reading, B.J., Andersen, L.K., Ryu, Y.W., Mushirobira, Y., Todo, T., and Hiramatsu, N. 2018e. Oogenesis and Egg Quality in Finfish: Yolk Formation and Other Factors Influencing Female Fertility. *Fishes*, 3(4):45. DOI: 10.3390/fishes3040045.
- Reading, B.J., Berlinsky, D.L., Woods III, L.C., Hodson, R.G., McGinty, A.S., Abernathy, J., and Fuller, S.A. 2019. *The status of striped bass, Morone saxatilis, as a commercially ready species for marine aquaculture*. Invited Symposium, Status of Marine Finfish Species for US Aquaculture. In *Abstracts of Aquaculture Triennial 2019: The Big Easy Choice!* (World Aquaculture Society). March 7-11. New Orleans, Louisiana, USA.
- Reading, B.J., Chapman, R.W., Schaff, J.E., Scholl, E.H., Opperman, C.H., and C.V. Sullivan. 2012. *An ovary transcriptome for all maturational stages of the striped bass (Morone*

saxatilis), a highly advanced perciform fish. BMC Research Notes, 5:111. DOI: 10.1186/1756-0500-5-111.

Reading, B.J., Clark, R.W., Hopper, M.S., Berlinsky, D.L., Kenter, L., Hodson, R.G., and McGinty, A.S. 2016. *Organized group-spawning of domestic striped bass Morone saxatilis*. In: Abstracts of Aquaculture Triennial 2016: All In For Aquaculture (World Aquaculture Society). February 22-26. Las Vegas, Nevada, USA.

Reading, B.J., Clark, R.W., McGinty, A.S., Hopper, M.S., Andersen, L.K., Ducharme, E.E., Rajab, S., Kenter, L.W., and Berlinsky, D.L. 2018b. *NC State University Updates on The National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry*. Invited Symposium, Striped Bass Growers Association Industry Forum. Aquaculture America 2018. February 19-22. Las Vegas, Nevada, USA.

Reading, B.J., Clark, R.W., McGinty, A.S., Hopper, M.S., Salger, S.A., Andersen, L.K., Kenter, L.W., and Berlinsky, D.L. 2018c. *Methods of domestic striped bass Morone saxatilis spawning that do not require the use of any hormonal induction procedures*. Invited Symposium, Striped Bass Growers Association Industry Forum. Aquaculture America 2018. February 19-22. Las Vegas, Nevada, USA.

Reading, B.J., Clark, R.W., McGinty, A.S., Hopper, M.S., Salger, S.A., Andersen, L.K., Kenter, L.W., and Berlinsky, D.L. 2018d. *Methods of domestic striped bass Morone saxatilis spawning that do not require the use of any hormonal induction procedures*. In Abstracts of Aquaculture America 2018 (World Aquaculture Society). February 19-22. Las Vegas, Nevada, USA.

- Reading, B.J., Hinshaw, J., and Watanabe, W.O. 2014a. Research Panel. The NC Marine Fish Culture Workshop. November 20-21. The NOAA Center for Coastal Fisheries and Habitat Research, Beaufort, North Carolina, USA.
- Reading, B.J., Hiramatsu, N., Schilling, J., Molloy, K.T., Glassbrook, N., Mizuta, H., Luo, W., Baltzegar, D.A., Williams, V.N., Hara, A., and Sullivan, C.V. 2014b. Lrp13 is a novel vertebrate lipoprotein receptor that binds vitellogenins in teleost fishes. *Journal of Lipid Research*, 55(11):2287-2295. DOI: 10.1194/jlr.M050286.
- Reading B.J., Hiramatsu N., and Sullivan, C.V. 2011. Disparate binding of three types of vitellogenin to multiple forms of vitellogenin receptor in white perch. *Biology of Reproduction*, 84:392-9. DOI: 10.1095/biolreprod.110.087981.
- Reading, B.J., McGinty, A.S., Clark, R.W., Hopper, M.S., Woods III, L.C., and Baltzegar, D.A. 2018a. Genomic enablement of temperate bass aquaculture (Family *Moronidae*). In: H. Wang and X. Liang (Eds.) *Breeding and Culture of Perch and Bass*. Science China Press (Chinese Academy of Sciences), Beijing, China.
- Reading, B.J., Williams, V.N., Chapman, R.W., Islam Williams, T., and Sullivan, C.V. 2013. Dynamics of the striped bass (*Morone saxatilis*) ovary proteome reveal a complex network of the translasome. *Journal of Proteome Research*, 12:1691-1699. DOI: 10.1021/pr3010293.
- Rexroad, C., Vallejo, R., Coulibaly, I., Couch, C., Garber, A., Westerman, M., and Sullivan, C. 2006. Identification and characterization of microsatellites for striped bass from repeat-enriched libraries. *Conservation Genetics*, 7(6):971-982. DOI: 10.1007/s10592-006-9122-0.

- Ryan, S.F., Adamson, N.L., Aktipis, A., Andersen, L.K., Austin, R., Barnes, L., Beasley, M.R., Bedell, K.D., Briggs, S., Chapman, B. and Cooper, C.B., Corn, J.O., Creamer, N.G., Depoundorne, J.A., Domenico, P., Driscoll, E., Godwin, J., Hjarding, J., Hupoundert, J., Isard, S., Just, M.G., Kar Gupta, K., Lopez-Uribe, M.M., O'Sullivan, J., Landis, E.A., Madden, A.A., McKenney, E.A., Nichols, L.M., Reading, B.J., Russell, S., Sengupta, N., Shapiro, L.R., Shell, L.K., Sheard, J.K., Shoemaker, D.D., Sorger, D.M., Starling C., Thakur, S., Vatsavai, R.R., Weinstein, M., Winfrey, P., and Dunn, R.R. 2018. The role of citizen science in addressing grand challenges in food and agriculture research. *Proceedings of the Royal Society B*, 285(1891):p.20181977. DOI: 10.1098/rspb.2018.1977.
- Salger, S.A., Reading, B.J., and Noga, E.J. 2017. Tissue localization of piscidin host-defense peptides during striped bass (*Morone saxatilis*) development. *Fish & shellfish immunology*, 61:173-180. DOI: 10.1016/j.fsi.2016.12.034.
- Schilling, J., Loziuk, P.L., Muddiman, D.C., Daniels, H.V., and Reading, B.J. 2015a. Mechanisms of egg yolk formation and implications on early life history of white perch (*Morone americana*). *PloS one*, 10(11). DOI: 10.1371/journal.pone.0143225.
- Schilling, J., Nepomuceno, A., Muddiman, D.C., Schaff, J.E., Daniels, H.V., and Reading, B.J. 2014. Compartment proteomics analysis of white perch (*Morone americana*) ovary using support vector machines. *Journal of Proteome Research*, 13(3):1515-1526. DOI: 10.1021/pr401067g.
- Schilling, J., Nepomuceno, A.I., Planchart, A., Yoder, J.A., Kelly, R.M., Muddiman, D.C., Daniels, H.V., Hiramatsu, N., and Reading, B.J. 2015b. Machine learning reveals sex-

- specific 17 β -estradiol-responsive expression patterns in white perch (*Morone americana*) plasma proteins. *Proteomics*, 15(15):2678-2690. DOI: 10.1002/pmic.201400606.
- Scruggs Jr., G.D. 1957. Reproduction of resident striped bass in Santee-Cooper Reservoir, South Carolina. *Transactions of the American Fisheries Society*, 85:144–159. DOI: 10.1577/1548-8659(1955)85[144:RORSBI]2.0.CO;2.
- SeafoodSource. 2014a. Catfish Product Profile. SeafoodSource Seafood Handbook. Available at: <https://www.seafoodsource.com/seafood-handbook/finfish/catfish>.
- SeafoodSource. 2014b. Tilapia Product Profile. SeafoodSource Seafood Handbook. Available at: <https://www.seafoodsource.com/seafood-handbook/finfish/tilapia>.
- Skalski, G.T., Couch, C.R., Garber, A.F., Weir, B.S., and Sullivan, C.V. 2006. Evaluation of DNA pooling for the estimation of microsatellite allele frequencies: a case study using striped bass (*Morone saxatilis*). *Genetics*, 173(2):863-875. DOI: 10.1534/genetics.105.053702.
- Small, B.C. and Soares, Jr, J.H. 1998. Estimating the quantitative essential amino acid requirements of the striped bass, *Morone saxatilis*, using filet A/E ratios. *Aquaculture Nutrition*, 4:225-232. DOI: 10.1046/j.1365-2095.1998.00075.x.
- Small, B.C., Soares Jr., J.H., and Woods III, L.C. 2000. Optimization of feed formulation for mature female striped bass. *North American Journal of Aquaculture*, 62:290-293. DOI: 10.1577/1548-8454(2000)062<0290:OOFFFM>2.0.CO;2.
- Smith, J.M. and Whitehurst, D.K. 1990. Tank spawning methodology for the production of striped bass. Pages 73-77 in R.M. Harrell, J.H. Kerby and R.V. Minton (Eds.) *Culture and propagation of striped bass and its hybrids*. Striped Bass Committee, Southern Division, American Fisheries Society, Bethesda, Maryland, USA.

- Stevens, B. 1984. Striped bass culture in the United States. In Chester, A.J. (Ed.) Sampling statistics in the Atlantic menhaden fishery (9). National Oceanic and Atmospheric Administration (NOAA), National Marine Fisheries Service (NMFS). The University of California, California, USA.
- Stevens, R.E. 1966. Hormone-induced spawning of striped bass for reservoir stocking. *Progressive Fish-Culturist*, 28:19-28. DOI: 10.1577/1548-8640(1966)28[19:HSOSBF]2.0.CO;2.
- Stevens, R.E. 1967. A final report on the use of hormones to ovulate striped bass, *Morone saxatilis* (Walbaum). Proceedings of the Southeastern Association of Game and Fish Commissioners, 18:523-538.
- Stevens, R.E. 1975. Current and future considerations concerning striped bass culture and management. The Proceedings of the Southeastern Association of Fish and Wildlife Agencies, 28(69).
- Stone, L. 1882. Report on overland trip to California with living fishes. Pages 637-644 in United States Commission of Fish and Fisheries. Report of the Commissioner for 1879. Washington: Government Printing Office.
- Sullivan, C.V., Chapman, R.W., Reading, B.J., and Anderson, P.E. 2015. Transcriptomics of mRNA and egg quality in farmed fish: Some recent developments and future directions. *General and Comparative Endocrinology*, 221:23-30. DOI: 10.1016/j.ygcen.2015.02.012.
- Swanson, P. and C.V. Sullivan. 1991. Isolation of striped bass, *Morone saxatilis*, pituitary hormones. *American Zoologist*, 31(5):3A.

- Tao, Y., Hara, A. Hodson, R.G., Woods, III, L.C., and C.V. Sullivan. 1993. Purification, characterization and immunoassay of striped bass (*Morone saxatilis*) vitellogenin. *Fish Physiology and Biochemistry*, 12:31-46. DOI: 10.1007/BF00004320.
- Teletchea, F. and Fontaine, P. 2014. Levels of domestication in fish: implications for the sustainable future of aquaculture. *Fish and Fisheries*, 15(2):181-195. DOI: 10.1111/faf.12006.
- Tipsmark, C.K., Luckenbach, J.A., Madsen, S.S. and Borski, R.J. 2007. IGF-I and branchial IGF receptor expression and localization during salinity acclimation in striped bass. *American Journal of Physiology (Regul Integr Comp Physiol)*, 292:R535–R543. DOI: 10.1152/ajpregu.00915.2005.
- Tipsmark, C.K., Madsen, S.S., and Borski, R.J. 2004. Effect of salinity on expression of branchial ion transporters in striped bass (*Morone saxatilis*). *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 301(12):979-991. DOI: 10.1002/jez.a.119.
- Treece and Associates. (2017). *The Texas aquaculture industry – 2017*. Lampasas, TX: Granvil D. Treece.
- Turano, M.J., Borski, R.J., and Daniels, H.V. 2008. Effects of cyclic feeding on compensatory growth of hybrid striped bass (*Morone chrysops* × *M. saxatilis*) foodfish and water quality in production ponds. *Aquaculture Research*, 39(14):1514-1523. DOI: 10.1111/j.1365-2109.2008.02023.x.
- USDA (United States Department of Agriculture). 2019. *2017 Census of Agriculture, 2018 Census of Aquaculture*, 3(2). Publication AC-17-SS-2.

- USFWS (United States Fish and Wildlife Service). 1882. United States Commission of Fish and Fisheries Report of The Commissioner Part VII for 1879. Government Printing Office, Washington, D.C., USA.
- Waldman, J.R., Dunning, D.J., Ross, Q.E., and Mattson, M.T. 1990. Range dynamics of Hudson River striped bass along the Atlantic coast. Transactions of the American Fisheries Society, 119(5):910-919. DOI: 10.1577/1548-8659(1990)119<0910:RDOHRS>2.3.CO;2.
- Webster, C. D. and Lovell, R. T. 1990. Response of striped bass larvae fed brine shrimp from different sources containing different fatty acid composition. Aquaculture 90:49-61. DOI: 10.1016/0044-8486(90)90282-R.
- Williams, V.N., Reading, B.J., Amano, H., Hiramatsu, N., Schilling, J., Islam Williams, T., Gross, K., and Sullivan, C.V. 2014a. Proportional accumulation of yolk proteins derived from multiple vitellogenins is precisely regulated during vitellogenesis in striped bass (*Morone saxatilis*). Journal of Experimental Zoology, 321(6):301-315. DOI: 10.1002/jez.1859.
- Williams, V.N., Reading, B.J., Hiramatsu, N., Amano, H., Glassbrook, N., Islam Williams, T., and Sullivan, C.V. 2014b. Multiple vitellogenins and product yolk proteins in striped bass, *Morone saxatilis*: molecular characterization and processing during oocyte growth and maturation. Fish Physiology and Biochemistry, 40(2):395-415. DOI: 10.1007/s10695-013-9852-0.
- Wirgin, I., Maceda, L., Tozer, M., Stabile, J., and Waldman, J. 2020. Atlantic coastwide population structure of striped bass *Morone saxatilis* using microsatellite DNA analysis. Fisheries Research, 226:105506. DOI: 10.1016/j.fishres.2020.105506.

- Won, E.T. and Borski, R.J. 2013. Endocrine regulation of compensatory growth in fish. *Frontiers in Endocrinology*, 4:74. DOI: 10.3389/fendo.2013.00074.
- Woodroffe, J.R. 2012. Historical Ecology of Striped Bass Stocking in the Southeastern United States. Master's thesis. East Carolina University, Greenville, North Carolina, USA.
- Woods III, L. C. 2001. Domestication and strain evaluation of striped bass (*Morone saxatilis*). *Aquaculture*, 202(3-4), 343-350. DOI: 10.1016/S0044-8486(01)00783-9.
- Woods III, C.L., Kerby, J.H., and Huish, M.T. 1983. Estuarine cage culture of hybrid striped bass. *Journal of the World Mariculture Society*, 595-612. DOI: 10.1111/j.1749-7345.1983.tb00113.x.
- Woods III L.C., Li Y., Ding Y., Liu J., Reading B.J., Fuller S.A., and Song J. 2018 DNA methylation profiles correlated to striped bass sperm fertility. *BMC Genomics* 19:244. DOI: 10.1186/s12864-018-4548-6.
- Woods III, L.C. and Soares Jr., J.H. 1996. Nutritional requirements of domestic striped bass broodstock. *Proceedings of the Second World Fisheries Congress*, 1:107.
- Woods III, L.C. and Sullivan, C.V. 1993. Reproduction of striped bass (*Morone saxatilis*) brood stock: monitoring maturation and hormonal induction of spawning. *Journal of Aquaculture and Fisheries Management*, 24:213-224. DOI: 10.1111/j.1365-2109.1993.tb00543.x.
- Woods, III, L.C., Woiwode, J.G., McCarthy, M.A., Theisen, D.D., and Bennett, R.O. 1990. Technical notes: noninduced spawning of captive striped bass in tanks. *The Progressive Fish-Culturist*, 52:201-202. DOI: 10.1577/1548-8640(1990)052<0201:TNNSOC>2.3.CO;2.

- Worth, S.G. 1884. Report upon the propagation of striped bass at Weldon, N.C., in the spring of 1884. *Bulletin of the United States Fish Commission*, 4:225-230.
- Worth, S.G. 1904. The recent hatching of striped bass and possibilities with other commercial species. *Transactions of the American Fisheries Society*, 33:223-230. DOI: 10.1577/1548-8659(1904)34[223:TRHOSB]2.0.CO;2.
- Worth, S.G. 1910. Progress in hatching striped bass. *Transactions of the American Fisheries Society*, 39:155-159. DOI: 10.1577/1548-8659(1909)39[155:PIHSB]2.0.CO;2.
- Zohar, Y. 1989. Endocrinology and fish farming: aspects in reproduction, growth, and smoltification. *Fish Physiology and Biochemistry*, 7(1-6):395-405. DOI: 10.1007/BF00004734.

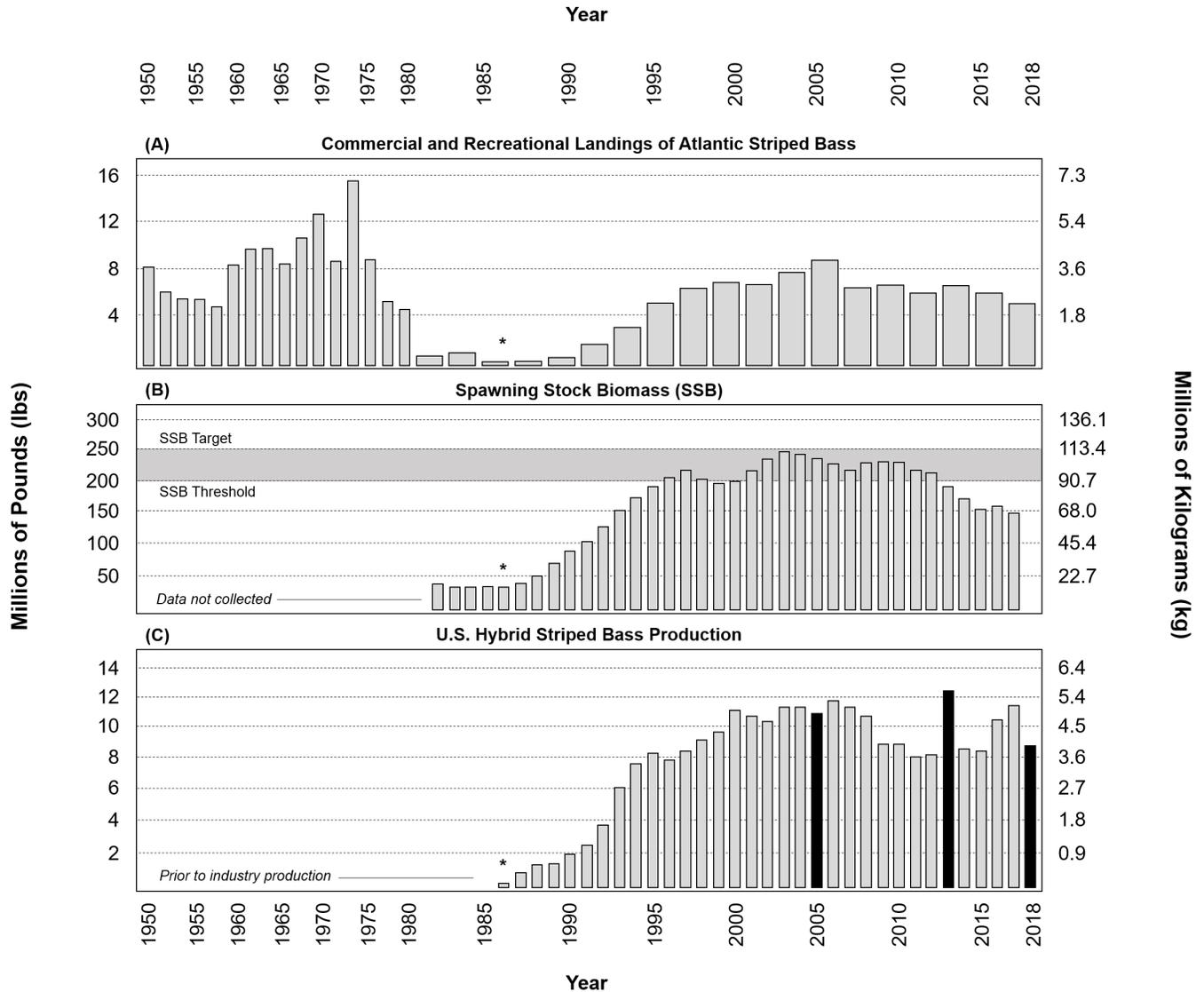


Figure 1.1. Commercial and recreational landings of Atlantic striped bass since the 1950s (a). Bars indicate the average landings per 2 years from 1950 to 2018. Spawning stock biomass (SSB) of Atlantic striped bass from the 1980s to 2016 (b). Hybrid striped bass production in the United States as reported since industry inception beginning in 1986 (c). Years that aquaculture production volumes were reported in the USDA Agriculture Census are indicated as black bars. Asterisks (*) on all panels mark the year the U.S. hybrid striped bass industry began production (1986). Data for these figures were provided by the Atlantic States Marine Fisheries Council (ASMFC) and National Oceanic and Atmospheric Administration (NOAA) National Marine Fisheries Service (NMFS) (a and b), and from Dr. James Carlberg (Kent SeaTech), Dr. Marc Turano (NC SeaGrant), Dr. Anita Kelly (University of Arkansas at Pine Bluff and Auburn University), and the USDA Agriculture Census (c).

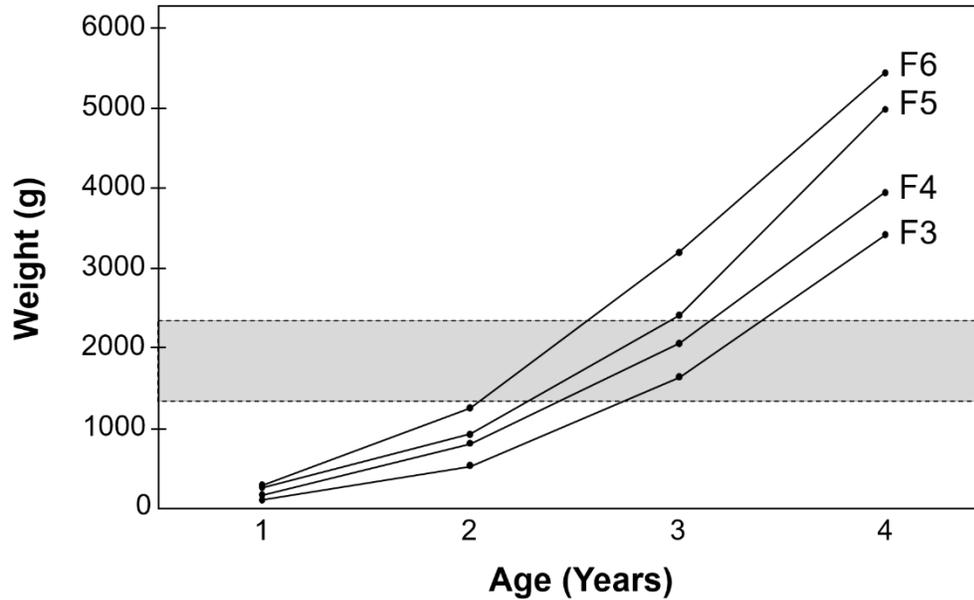
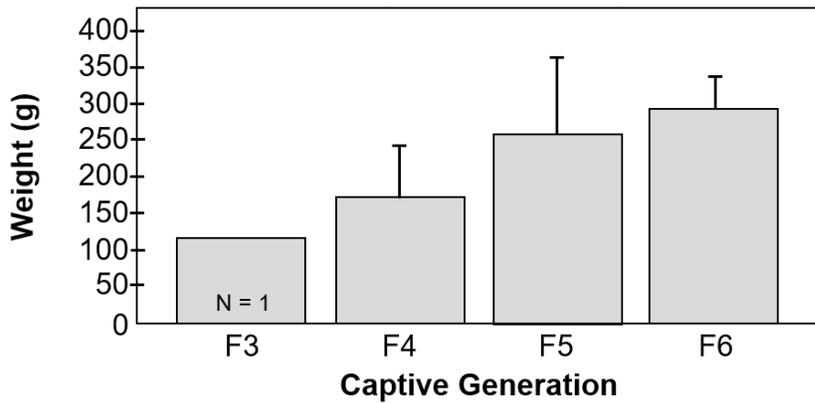


Figure 1.2. Domestic striped bass broodstock performance data collected for different age classes and generations (collected between March and June of each year, 2005–2020): Year 1 (45–60 weeks of age), Year 2 (80–104 weeks of age), Year 3 (136–154 weeks of age), and Year 4 (197–209 weeks of age). The x-axis is the age class grand mean and the y-axis is weight (g). The filial generation of captive breeding is indicated for the periods of 2004–2007 (F3), 2008–2011 (F4), 2012–2015 (F5), and 2016–2019 (F6). Gray shading indicates the target striped bass market size at between 1.36 and 2.27 kg (3.0 and 5.0 lb). Each datapoint for F3, F4, F5, and F6 represents a grand mean value of 3 or 4 different age class cohorts and hundreds of fish were measured for each age class per annum with the exceptions of F3 age class 1 (a single cohort), F6 age class 2 (2 cohorts), F6 age class 3 (2 cohorts), and F6 age class 4 (a single cohort).

A) Size of Domestic Striped Bass at Age 1 Year



B) Size of Domestic Striped Bass at Age 2 Years

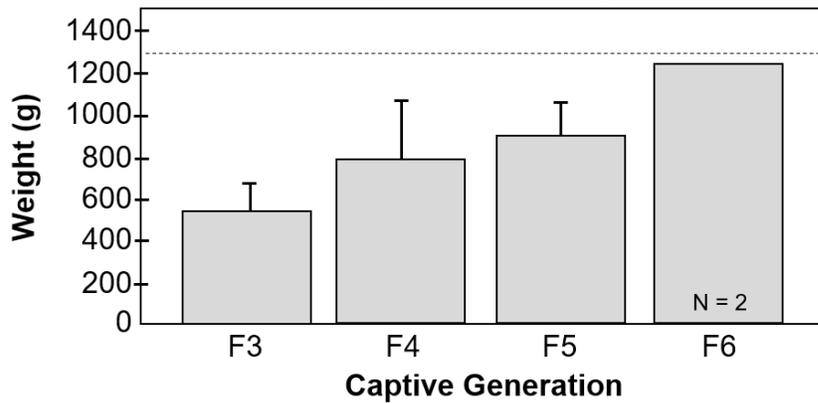


Figure 1.3. Domestic striped bass age class performance data (collected between March and June of each year, 2005–2020): (a) Year 1 (45–60 weeks of age) and (b) Year 2 (80–104 weeks of age). The filial generation of captive breeding is indicated for the periods of 2004–2007 (F3), 2008–2011 (F4), 2012–2015 (F5), and 2016–2019 (F6). Bars indicate grand mean values and error bars indicate standard deviation where there were three or four age class observations (annual performance data of hundreds of fish) for each filial generation; error bars are omitted where there were only one or two annual observations indicated as N=1 or N=2. The dashed line indicates target market size for striped bass at 1.36 kg (3.0 lb).

CHAPTER 2. INSIGHTS INTO HETEROSIS OF HYBRID STRIPED BASS VIA MOTIF-FINGERPRINTING

Abstract

A novel machine learning (ML) data analysis pipeline was developed and employed to identify the genes that are most determinant in the regulation of economically important traits in hybrid striped bass (HSB; female white bass, WB, *Morone chrysops* x *M. saxatilis*, male striped bass, SB). Specifically, trained ML models were applied to predict HSB growth performance at two critical time points of production, juvenile grading (2-3 months of age) and final market harvest (15 months of age), as well as sire strain. Five sire strains were represented, corresponding to the geographic location of origin of the SB sires, as follows: Domestic (DOM), Florida (FL), South Carolina (SC), Texas (TX), and Virginia (VA), whereby DOM SB were fifth generation domestic bred at the North Carolina State University Pamlico Aquaculture Field Laboratory as part of the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry*. Transcriptomes of HSB white skeletal muscle tissue (n=40 individuals) collected at final harvest were scanned against a genomic library of unique protein motif fingerprints (MFs) each twelve amino acids in length. A total of 15,000 MFs that significantly varied in expression, calculated as mean read count in each of the six reading frames, between any sample or technical replicate thereof were identified and analyzed through the ML workflow. These data were first reduced by information gain (Shannon's Entropy), whereby MFs determined to provide no information to ML algorithms towards the correct classification of samples into growth performance or strain groups were eliminated from subsequent analyses. The MFs were further reduced by performing recursive exclusion of less informative MFs to identify points of model overfitting and underfitting, or when too many or

too few attributes, respectively, are included in the dataset and may negatively impact model performance. The top 500 MFs for each comparison remained that, without duplicates across comparisons, were 821 MFs that mapped to thirty-three genes that can be regarded as potential targets for breeding, other genetic manipulation, and future research efforts. Moreover, the examination of individual MFs mapped to translated regions of the SB and WB reference genome assemblies have enabled the determination of instances in which the expression of one allele specific to the SB or WB parent is more influential for growth performance in the hybrid offspring. This is the only study that has been conducted to date that explores gene expression at the allele-level in HSB to understand heterosis, or hybrid vigor, displayed by these fish. The pathway analysis conducted to better understand the relationships between gene expression and growth in HSB revealed molecular interactions and cellular pathways supporting lipid metabolism and function of signaling pathways (JAK/STAT and AMPK) in fish of superior growth phenotype and that these and related functions are identified among the DOM HSB that also exhibited superior growth traits.

Introduction

The farming of aquatic organisms (aquaculture) has been reported as the fastest growing sector of food production in the world and is the source of half of the fish consumed worldwide. The United States (US) is ranked 17th in production and roughly 80.0 % of seafood is imported leading to a trade deficit in excess of \$17 billion USD. Thus, there is great socioeconomic potential of domestic aquaculture production and for some species this is beginning to be realized. Hybrid striped bass (HSB, a cross between white bass, WB, *Morone chrysops* x *M.*

saxatilis, striped bass, SB) is currently the fourth most valuable cultured finfish in the US and recently experienced a farmgate value peak of \$50 million USD.

The seminal research on SB and WB crosses was completed in effort to produce a fish that expressed hybrid vigor (i.e. heterosis, offspring showing superior qualities than either parent), as has been done for many other crops and animals, including other fish species, involved in agricultural production (Rahman et al., 1995, 2018; You et al., 2015). This research was conducted in part at or by collaborators of North Carolina State University (NCSU, Raleigh, NC) and the hybridization of both species was successful in terms of creating a production industry (Hodson, 1990). Understanding how the parental genomes from these two different species combine to produce the hybrid phenotypes is not clearly understood.

Here we describe a genome-wide motif fingerprinting analyses of SB and WB that allowed for insight into the genetic mechanisms underlying phenotypes in HSB. Further, the results show genes present in the HSB and derived from exclusively WB that are relevant to growth and perhaps other culture traits. This is the only study that has been conducted to date that explores gene expression at the allele-level in these fish and in the context of parental strain and important culture traits. As such, the findings here represent novel information regarding the combination of the parental genomes and potential inheritance patterns of genes from the WB (maternal) parent and how they relate to growth. Lastly, these findings empower genomic selection as well as predictive phenomics in the future and will be incorporated into the breeding program for these fish.

Materials and Methods

Experimental Animals

This study was carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health (National Research Council, 1996) and all procedures were approved by the Institutional Animal Care and Use Committee of North Carolina State University (protocols 10-042-A and 19-065-O). The described activities of spawning, rearing, and sampling fish for this project took place at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL, Aurora, NC, USA).

HSB were produced by crossing ninth generation domestic (F9) three-year-old white bass (WB, *Morone chrysops*) female ‘dams’ (mean \pm standard deviation, SD, weight was 0.77 ± 0.15 kg) and male striped bass (*M. saxatilis*) ‘sires’ of wild and domestic origin (mean \pm SD weight was 3.73 ± 0.77 kg). Wild-origin sires were three years of age and represented four geographically distinct strains; sires of each strain had been bred in captivity from parent fish caught in the waters of Florida (FL), South Carolina (SC), Texas (TX), and Virginia (VA). Domestic-origin sires (DOM) were two years of age and fifth generation domestic (F5). The domestic fish (WB dams and DOM SB sires) were bred and reared as part of the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry*, coordinated by Dr. Benjamin J Reading (NCSU) and Dr. S. Adam Fuller (U.S. Department of Agriculture, Agricultural Research Service) housed at the NCSU PAFL. Males were anesthetized (MS-222, Sigma-Aldrich, St. Louis, MO, USA), verified for spermiation by applying gentle pressure to their ventral surfaces, and treated with 165 IU human chorionic gonadotropin (hCG)/kg body weight (Chorulon®, Merck Animal Health, Kenilworth, NJ, USA) via injection

into the dorsal lymphatic sinus approximately 40 h prior to producing the HSB crosses. Females were treated with 300 IU hCG/kg in a similar manner approximately 24 h prior to stripping to induce ovulation. Steps to verify spermiation, evaluate oocyte stage and female eligibility, and egg incubation were completed according to the established protocols for artificial production of HSB (Rees and Harrell, 1990). Eggs from each dam were divided into five equal parts and fertilized *in vitro* with milt from one sire of each strain to control for paternal fertility effects (Table 2.1). Caudal fin samples of each broodfish were collected and stored in 95.0 % EtOH for microsatellite genotyping and parentage assignment.

Hatchery incubation and rearing of HSB offspring completed according to the industry-standard practice and the two-phase production cycle for HSB (D'Abramo and Frinsko, 2008). Briefly, larvae from each spawn were pooled and stocked into three replicate ponds (0.1 ha each). At two months of age HSB were harvested, combined equally in number from each pond into several replicate tanks, and graded (i.e., sorted by size) into two groups: Top Grade (TG), or those fish anticipated to reach or exceed market size by the conclusion of the production cycle (16–18 months), and Runts, or those deemed unlikely to reach market size by that time. HSB of each grade group were re-stocked into replicate ponds (n=2,630 fish/0.1 ha pond) until seven months of age when fish were sampled and stocked into replicate indoor flow-through tanks (75 fish/2,400 L tank). HSB were moved to replicate outdoor flow-through tanks at twelve months of age to reduce tank density (n=125 fish/5,800 L tank, three tanks per grade group).

At the final harvest of fifteen months of age, a sub-sample of HSB (n=72) was randomly selected from the replicate tanks (n=12 per tank) and sacrificed for the collection of tissue samples and measurement of morphometric data. Caudal fin clips were stored in 95.0 % EtOH for microsatellite genotyping and parentage assignment. White, fast-twitch skeletal muscle tissue

samples (approximately 3.5 mm in size) were excised from the left side of forty of the sacrificed individuals just below the anterior margin of the first dorsal fin and stored in 1.0 mL of RNA-Later (Life Technologies, CA, USA) for 36 h prior to removing the RNA-Later and freezing samples at -80.0 °C. Selected samples were those from individual fish that represented: the ten largest fish by weight, the ten smallest fish by weight, among the eighteen randomly selected individuals (three from each replicate tank), or two individuals chosen to represent sires that would otherwise not have been based on the selection criteria and that allowed for the representation of TG and Runt fish to be equal between groups (n=20 per grade group, TG and Runt, **Table 2.1**). Duplicate white muscle tissue samples were collected from eighteen (n=18 of 72) of the sacrificed fish (four from each group of twelve randomly selected from each replicate tank) and stored in Bouin's Solution (Sigma-Aldrich) and stored at 4.0 °C prior to processing for histology.

Additional groupings of HSB were designated at the time of final sampling representative of actual growth performance (i.e., as opposed to the grade assigned at two months of age). Specifically, the HSB that weighed above the mean weight for TG at the time of final harvest were designated as “Large” (LG), those that weighed below the mean weight for Runt HSB were designated as “Small” (SM), and “Intermediate” (IN) fish were those that fell in between the two means by weight (**Table 2.1**).

Microsatellite Genotyping and Parentage Assignment

Microsatellite genotyping and subsequent parentage assignment was completed as described in Kenter et al. (2018). Briefly, genomic DNA was extracted from the fin clips collected from the dams, sires, and sampled HSB offspring using eleven microsatellite markers

identified in Couch et al. (2006) and Rexroad et al. (2006). Fragment analysis of PCR-amplified DNA was completed at the Yale University DNA Analysis Facility (New Haven, CT, USA) using an automated DNA sequencer (3730xl 96-capillary genetic analyzer; Applied Biosystems, Foster City, CA). Peak Scanner™ software (v.2.0, Applied Biosystems, Waltham, MA, USA) was used to manually score peaks and sort scores into allelic bins. Parentage was assigned using CERVUS software (v.3.0.7, Field Genetics, London, UK), which calculates a log-likelihood of each candidate parent being the true parent and ultimately determines the level of confidence in the parentage assignment (Kalinowski et al., 2007). A 1.0 % genotyping error rate was assumed. Sibship assignment using the maximum likelihood method employed in the COLONY program (v.2.0.6.3, Institute of Zoology, London, UK) was conducted for any offspring with a level of confidence in parental assignment <80.0 % (Jones and Wang, 2010).

Muscle Histology

Muscle samples were prepared for histological analysis at the NCSU College of Veterinary Medicine (CVM, Raleigh, NC, USA). Briefly, each muscle tissue sample (n=18) was washed in ethanol, embedded tissue in paraffin, sectioned each of the tissue samples into 5 µm cross sections perpendicular to the muscle fiber, stained the tissue with hematoxylin and eosin stain (H&E), and mounted each cross section onto microscope slides for muscle morphometry analysis. Identifying slide labels, which corresponded to sample identification numbers, were concealed until after enumerations had been made to eliminate potential bias in image processing and analysis.

Three unique images of the muscle fibers on each slide were taken using an Olympus CH microscope (4X) with a Celestron digital microscope imager camera (10X). ImageJ software

(Fiji, v.1.52a, National Institutes of Health, NIH, Bethesda, USA) was used to count and determine the area (μm) of all fibers entirely within the field of view (i.e., not marginal fibers or those with a limited area due to the margin of view) of each image to evaluate hyperplastic (fiber number) and hypertrophic (fiber area) muscle growth. The diameter of each scored fiber was calculated as a geometric derivative of its area to compute the average fiber diameter and the frequency of scored fibers of a certain diameter range for each slide image. Specifically, the frequency (i.e., count) of fibers scored on each slide image that fell into 10–20 μm , 10–25 μm , 10–50 μm , 10–75 μm , 10–100 μm , >75 μm , >100 μm , and >150 μm diameter bins were recorded. Image processing and subsequent calculations were completed in duplicate (i.e., by two scorers) and the total number of fibers, average fiber diameters, and number of fibers per diameter bin were pooled prior to analysis to account for any scorer-error. The ROUT outlier test was used to identify and subsequently remove outliers (Motulsky and Brown, 2006).

Statistical Comparisons

Statistical comparisons between the weight and TL of HSB of differing grade, growth performance, and strain groups were conducted using a Student's *t*-Test, one-way ANOVA, and Tukey's Honestly Significant Difference (HSD) post-hoc test, as appropriate ($\alpha=0.05$; JMP® Pro v.14.0.0, SAS Institute, Cary, NC, USA). The total number of fibers, average fiber diameter, and number of fibers of the different diameter bins were compared between grade and growth performance groups of HSB offspring in a similar fashion. A simple linear regression was used to assess the relationship between the average total fiber number and average fiber diameter to the weight and TL of offspring.

RNA-Sequencing of Hybrid Striped Bass White Muscle Tissue

Library preparation (RNA) and next-generation sequencing for the forty HSB white muscle tissue samples selected as described above was performed by the NCSU Genomic Sciences Laboratory (GSL, Raleigh, NC, USA). Briefly, total RNA extraction was performed using the RNeasy Fibrous Tissue Mini Kit (Qiagen, Inc., Hilden, Germany). Sample quality and concentration was evaluated using an Agilent Bioanalyzer 2100 with an RNA 6000 Nano chip (Agilent, Santa Clara, CA, USA). The NEBNext Poly(A) mRNA Magnetic Isolation Module oligo-dT beads (New England Biolabs, NEB, Ipswich, MA, USA) were used for messenger (mRNA) purification. Complementary DNA (cDNA) libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit and NEBNext Multiplex Oligos for Illumina sequencing (Illumina Inc., San Diego, CA, USA). Per the manufacturer-specified protocol, mRNA was chemically fragmented and primed with random oligos for first strand cDNA synthesis, dUTPs were incorporated during second-strand cDNA synthesis to accomplish strand-specificity, and cDNA was purified, end repaired, and “a-tailed” for adaptor ligation. Samples were then selected for a final library size of 400–550 base pairs (bp) using sequential AMPure XP bead isolation (Beckman Coulter, Brea, CA, USA). Library enrichment was performed and specific indexes for each sample were added during the protocol-specified PCR amplification. Amplified library fragments were purified and an Agilent 2100 Bioanalyzer with a High Sensitivity DNA chip was used to verify quality and final concentration. The final libraries were pooled in equimolar amounts and sequenced on the Illumina HiSeq 2500 platform across eight lanes with 125 bp paired-end reads (200 million reads per lane), which yielded sixteen technical replicates of sequence data for each of the forty individuals for which samples were submitted.

Motif Fingerprint (MF) Discovery

Motif fingerprint (MF) discovery analyses were performed by Orion Integrated Biosciences, Inc. (Manhattan, Kansas, USA). Motifs are 12-amino acid-long (aa, protein), non-redundant (unique), position-independent sub-sequences identified from data for which taxonomic classification is known, therefore serving as unique ‘fingerprints’ for the given taxonomic group, species, or individual thereof from which it was identified. For example, a twelve amino-acid sequence unique to those identified amongst protein sequence data of fishes in the Family *Moronidae* could be considered a MF for that taxonomic group. Orion Integrated Biosciences performed MF discovery by scanning a library of 3.5 million MFs associated to taxonomically-relevant species (i.e., Family *Moronidae* or subtaxa thereof) against all six possible reading frames of the translated *M. chrysops* (DOM_MoChry_2.0; GCA_019097615.1) and *M. saxatilis* (NCSU_SB_2.0; GCA_004916995.1) reference genome assembly sequences and the whole-transcriptome data generated from the forty HSB for which RNA-Seq data was generated. Expression values for MFs were calculated as the mean read count across each of the six frames, whereby a “hit” is only counted if there is a complete, twelve-out-of-twelve amino acid match in any of the reading frames. Orion Integrated Biosciences identified 15,000 MFs of which the patterns of expression were the most varied between individual fish and/or replicates thereof, thus considered to be the most determinant of observed variation in the population.

Machine Learning Analysis

The 15,000 MFs Orion Integrated Biosciences identified as being of high variance between individuals and/or replicates were further reduced in dimensionality (i.e., number) through the application of a supervised machine learning (ML) workflow applied using Weka

software (v.3.8, University of Waikato, New Zealand). This workflow was applied separately for each of the three comparisons of: 1) Grade (Top Grade, TG vs. Runts); (2) Growth (Large, LG vs. Small, SM vs. Intermediate, IN); and (3) Strain (Domestic, DOM vs. Florida, FL vs. Texas, TX vs. South Carolina, SC vs. Virginia, VA sires) to identify which MFs (attributes), and subsequently genes, are the most informative to the classification of individuals (instances) into the user-defined groups (classes). Four ML algorithms were used for model building and subsequent determination of informative attributes: a support vector machine (SVM; sequential minimal optimization, SMO), an artificial neural network (ANN; multilayer perceptron, MLP), a decision tree (J48), and an ensemble (i.e., combination of models) decision tree (Random Forest). Two cross-validation strategies, the holdout method and the stratified K -fold cross-validation, were applied with each algorithm to designate the training and test subsets of data necessary for model building and evaluation. The 66.0 % split was used as the holdout method, whereby 66.0 % of data is randomly designated as the training set from which the algorithm learns and the other 34.0 % of data is designated as the test set used to validate the model and measure performance. In the K -fold cross-validation, the data was partitioned into twenty subsets of data (i.e., 20-fold cross). The K -fold cross-validation then runs several successive models where each run is the model training on the entire dataset minus a random fold and the random fold that is excluded changes with each run. Three metrics were used to evaluate model performance: (1) percent correct classification, or how well the model classified each instance to its true class based upon the expression values associated with the provided attributes; (2) Area Under the Receiver Operating Characteristic Curve (AUROC, ROC Curve), which is the power of the analysis as a function of the Type I Error, or the plot of the true positive rate (TPR) against

the false positive rate (FPR); and (3) Kappa Statistic (Cohen's Kappa Coefficient), a measure of randomness, or the possibility of the model correctness as a function of chance.

Each cross-validated algorithm was first applied to the complete labeled dataset to establish baseline model performance. Information gain ("Shannon's Entropy") values were then calculated for each attribute. Information gain is a numerical value representative of the amount of information gained from the inclusion of a given attribute in learning. Attributes with a lower calculated entropy value are more informative when considering its association in learning (i.e., training). Each attribute was then assigned a rank based on information gain values, whereby the top-ranked attribute (i.e., "number 1") was that with the lowest calculated entropy value. Any attributes with an information gain value of 0.0 (no information) were removed from subsequent analyses, thus representing the first step in reduction of data dimensionality.

Data dimensionality was further reduced using recursive elimination by identifying the minimum number of highest-ranking MFs required to maintain optimal classification determined for the four algorithms. Specifically, subsets of the ranked attributes (i.e., information gain weight above 0.0) dataset were created such that each subset included fewer and fewer top-ranked attributes to allow for the evaluation of changes in model performance (model fit). The iterative recursive exclusion of ranked attributes allows for the identification of model overfitting and underfitting, or when too many or too few attributes, respectively, are included in the dataset and may negatively impact model performance. For each comparison, the ranked attributes were first reduced to include only the top 3,000 attributes, then the top 2,000, subsequently partitioned by 250 attributes at a time (1,750, 1,500, 1,250, 1,000, 750, 500, 250), then by 50 attributes (200, 150, 100), by 25 attributes (75, 50, 25), and finally only the top ten (10) were included. Models were not run for performance comparison beyond the inclusion of 3,000 MFs of highest rank, as

at least two models (SMO and Random Forest) were able to correctly assign 100.00 % of the MFs to groups when the top-ranked 1,000–2,000 MFs were included in the dataset. The MLP model was not able to be completed beyond the inclusion of the top 1,250 ranked MFs for each of the three comparisons due to a limitation in computer processing power that is required to complete a memory-intensive model such as the MLP.

Model performance (percent correct classification) for each cross-validated algorithm was plotted against the number of attributes included in each iterative run. The number of MFs required to avoid model underfitting (i.e., too few MFs included in analysis leading to diminished model performance) was determined as the point at which the four models had the greatest agreement of percent correct classification performance with the inclusion of the fewest high-ranking MFs (**Table 2.2**). The number of MFs required to avoid model overfitting (i.e., too many MFs included in analysis leading to diminished model performance) was determined by identifying the point at which the performance of at least one model began to decline with the inclusion of a greater number of high-ranking MFs. The top-ranked MFs identified as those necessary to avoid underfitting for each comparison are referred to as “optimal”, however, those MFs identified as necessary to avoid overfitting for each comparison were included in mapping and subsequent analysis as described in sections below.

A negative control for learning was established by running the cross-validated algorithms on randomized versions of the optimal dataset such that the class labels (e.g., TG or Runt) are not known to be associated with data truly representative of a given class (Schilling et al., 2014, 2015). This randomization process was repeated ten times for each comparison and the average values of performance metrics were calculated. Mean values were approximately what can be predicted from random assignment based on the number of classes for the Grade, Growth, and

Strain comparisons, thus it was concluded that true learning occurred when using correctly labeled data to build models (**Table 2.2**).

Mapping Motif-Fingerprints to Genes

The MFs identified as those to avoid overfitting for each comparison (ranked as the top 500 for Grade, Growth, and Strain) were combined and duplicates were removed to yield a list of 821 unique MFs. These MFs were then concatenated to longer (>12 amino acids) sequences by aligning MFs to determine the longer sequence of amino acids that subsequently map (i.e., align) to the same protein. The National Center for Biotechnology Information (NCBI) BLASTp® (Basic Local Alignment Search Tool for amino acid protein sequences) was used to compare the portion of amino acid sequence that included the MFs against the annotated *M. saxatilis* reference genome (NCSU_SB_2.0, accession: GCA_004916995.1; percent identity ranged from 90.00–100.00 % and e-values ranged from 4.00E-09–0.064) (Altschul et al., 1990, 1997).

Allele-Level Resolution of Motif Fingerprints

The number of hits of each MF in the translated SB and/or WB genome assembly sequences were examined further to identify any allele-level distinctions between the MFs. Specifically, the extent to which MFs that mapped to the same encoding gene, referred to here as a “set”, were representative of, and therefore could be inherited from, both parent species was examined further as follows:

If complete homology of MFs in a set were only found (i.e., measured as a hit) among the translated genome sequence of one parent, SB (paternal) or WB (maternal), the allele was considered exclusively representative of that parent species (i.e., overdominance). If complete

homology of MFs in a set were found among both translated genome sequences, the allele was considered ambiguous and representative of both parent species (i.e., overdominance or codominance). If complete homology of some MFs in a set were exclusively identified among the sequence data for one parent and different MFs of the same set were exclusively identified among the sequence data of the other parent, the alleles were considered distinct but representative of both parent species (i.e., codominance). If complete homology of MFs in a set were not identified among sequence data for either the SB or WB parent, allele(s) were considered “undetermined” (i.e., overdominance or codominance) and the closest relative of family *Moronidae* was used as a reference (*Dicentrarchus labrax*). The undetermined MFs were aligned to the translated SB and WB genome sequence data to the best extent possible (i.e., with increased confidence the greater number of amino acid matches) in effort to elucidate any reason(s) complete homology was not found (e.g., single mismatch, genome assembly gap).

Pathway Analysis

Qiagen Ingenuity Pathway Analysis (IPA; Qiagen) software was used to identify enriched pathways (“Canonical Pathways”), make predictions of the molecules and cellular processes that may underlie observed patterns in gene expression (“Upstream Regulators” and “Causal Networks”), and identify which of the genes identified from the 821 unique MFs identified between the top 500 for each comparison are the most informative to pathway and network building (“Top Analysis-Ready Molecules”) for these HSB or a specific comparison group thereof (Krämer et al., 2014). The designation of a pathway or network as “enriched” is based upon the number of associated molecules from the dataset and calculated p-values measuring whether the presence of these molecules in the dataset and expression thereof is likely

due to random chance. A z-score calculated from the \log_2 fold change of gene expression between groups is used to predict the activity state of a molecule (activated or inhibited) and ultimately predict what biological mechanisms may underlie differences in gene expression and subsequently phenotype or other variation between groups. The Qiagen Ingenuity® Knowledge Base (i.e., molecule, network, and pathway information) is specific to human (*Homo sapiens*), mouse (*Mus musculus*), and rat (*Rattus norvegicus*) systems, and therefore the orthologous human gene symbols for the SB genes identified from concatenated MFs were obtained through NCBI Genbank® and input into IPA.

Results

Morphometric Comparisons

The mean \pm standard deviation (SD) weight of the entire study population of HSB (N=752) was 525.84 ± 145.36 g and mean TL was 324.20 ± 28.93 mm by final harvest at fifteen months of age (**Figures 2.1** and **2.2**). The subpopulation of HSB sacrificed for this study (n=72) were representative of the study population as the mean weight and TL were 530.90 ± 140.19 g and 323.57 ± 28.00 mm, respectively (**Figures 2.1** and **2.2**). Of the subpopulation of sacrificed HSB, those included in sequencing analysis (n=40) weighed an average of 525.40 ± 169.48 g and were 320.53 ± 33.34 mm in TL and those included in the histology analysis (n=18) weighed 516.06 ± 107.85 g and were 321.17 ± 21.53 mm in TL. The HSB in TG and Runt groups significantly differed by weight and TL in all populations/study groupings of HSB (**Table 2.3**, **Figures 2.1** and **2.2**). Similarly, HSB designated as Large (LG), Intermediate (IN), or Small (SM) at final harvest for the Growth comparison also significantly differed in weight and TL in all populations/study groupings (**Table 2.3**). Although the average weight of the population was

approximately 155 g (0.34 lb) shy of market size for these fish (~680 g, or 1.5 lb) at fifteen months of age, a total of 121 HSB of the entire population (i.e., of 752) reached or exceeded market size (16.09 %); thirteen of these fish were among those in the sacrificed subpopulation, eleven of which were included in the sequencing analysis and two in the histology analysis.

Parentage of SB sires was successfully assigned to all but one of the sacrificed HSB (multiplex PCR failure for one offspring) and all five strains were represented by at least one sire among these fish (**Table 2.4**). The DOM and VA sires that were represented among the sacrificed HSB offspring were the only sire strain groups that significantly differed in weight ($p=0.0061$; **Table 2.4**). The specific representation of SB strains among the sacrificed HSB offspring and those HSB included in sequencing analysis is provided in **Table 2.4**, as well as the number of HSB produced from specific sire strains and grouped into TG or Runt for the Grade comparison and LG, IN, or SM for the Growth comparison. The DOM strain had the greatest representation (i.e., majority) among the sacrificed offspring (30.56 %) (**Table 2.4**) and these HSB weighed significantly more than the offspring spawned from the SC, TX, and VA sires ($p\leq 0.0266$) (**Figure 2.3**). A greater number of DOM HSB in the sample population were in the TG group, and all sampled HSB produced by the FL sire ($n=4$) were in the TG group (**Tables 2.4 and 2.5, Figure 2.4**). As such, more of the DOM and FL HSB were found to be in the LG growth group compared to IN or SM. This was not the case for the HSB in the SC, TX, or VA strain groups, where more of the sampled HSB belonged to the Runt group and subsequently the IN and/or SM group(s) for the growth comparison (**Tables 2.4 and 2.5**). The TG HSB produced from DOM, SC, TX, and VA sires were larger in both weight and TL than the Runt HSB in the same strain groups, however, a statistically significant difference between the weight and TL of TG and Runt HSB within strain groups was only observed for DOM HSB ($p<0.0001$), SC HSB

($p \leq 0.0443$), and VA HSB ($p \leq 0.0003$) (**Table 2.5, Figure 2.4**). The HSB that parentage could not be assigned to weighed 441.00 g, was 310.00 mm in TL and had been in the Runt group.

Muscle Histology

The HSB sampled for histology significantly differed in weight between grade groups ($p=0.0013$), however, the grade groups did not significantly differ in TL ($p=0.0804$) (**Table 2.6**). The only statistical difference in muscle fiber comparisons between groups was that the HSB in the Runt group had a significantly greater number of fibers in the 10–25 μm diameter bin (56.78 ± 29.11 fibers) than the TG fish (30.11 ± 16.39 fibers, $p=0.0329$) (**Table 2.6**). The simple linear regressions conducted to predict the relationship between muscle tissue histology metrics and HSB morphometrics identified significant correlations between HSB weight and fibers of 10–25 μm in diameter, HSB TL and fibers of 10–20 μm and of 10–25 μm . The fitted regression model for HSB weight is: $\text{Fibers } 10\text{--}25 \mu\text{m} = 118.73 - 135.12 * \text{weight}$ ($p=0.0384$, $R^2=0.24$). The fitted regression models for HSB TL are: $\text{Fibers } 10\text{--}20 \mu\text{m} = 189.66 - 0.47 * \text{TL}$ ($p=0.0247$, $R^2=0.28$) and $\text{Fibers } 10\text{--}25 \mu\text{m} = 271.94 - 0.68 * \text{TL}$ ($p=0.0052$, $R^2=0.40$).

Machine Learning Analysis

The reduction of 9,600,000 expression values for the 15,000 MFs across the sixteen technical replicates of the forty HSB for which RNA-Seq and MF discovery was completed is depicted as a flowchart in **Figure 2.5**.

Of the 15,000 MFs identified as highly varying between individual HSB and replicates thereof, 12,473 were assigned an information gain value over 0.0 for the Grade classification (TG, Runts), 13,811 MFs for Strain (DOM, FL, SC, TX, VA), and 12,969 for Growth (LG, SM,

IN) (**Figure 2.5(A)**). Of these ranked MFs, 11,191 were shared between all three comparisons as shown in **Figure 2.5(B)**, a Venn diagram of the number of shared and unique MFs identified among those ranked for the three comparisons. The plotted performance of each model for each of the three comparisons are shown in **Figure 2.5(C)** and individually (i.e., for each comparison) in **Figures 2.6–2.8**. The threshold for overfitting, or the maximum number of attributes input for agreement of optimal classification performance between cross-validated algorithms, was identified as the top 500 ranked MFs for all three comparisons. The threshold for underfitting, or the minimum number of attributes input for agreement of optimal classification performance between cross-validated algorithms, was identified as the top 200 MFs for the Grade comparison, top 100 for Growth, and top 150 for Strain (**Figures 2.5(D-E)–2.8**).

Motif Fingerprint Mapping and Allele-Level Resolution

The amino acid sequences concatenated from the 821 unique MFs between the top 500 ranked for each comparison mapped to thirty-three (33) unique genes (**Table 2.7**). Eight of these genes, apolipoprotein A-I (*APOA1*), carboxypeptidase B1 (*CPB1*), chymotrypsin-like elastase 1 (*CELA1*), chymotrypsin-like elastase 2A (*CELA2A*), hemoglobin subunit alpha 1 (*HBA1*), hemoglobin subunit beta (*HBB*), mitochondrially encoded ATP synthase membrane subunit 8 (*MT-ATP8*), and transferrin (*TF*), were mapped to by the 223 MFs that were shared between the top 500 for each comparison (**Table 2.7** and **Figure 2.5(D)**). These genes were also mapped to by MFs shared between the top 500 of only two comparisons or unique to a single comparison. Fifteen of the thirty-three genes were mapped to by MFs that were not shared between the top 500 ranked attributes for any comparisons (i.e., were uniquely among only the top 500 ranked attributes for one of the three comparisons). Specifically, MFs that mapped to mitochondrial

genes encoding the NADH:ubiquinone oxidoreductase core subunit 1 through subunit 4L (*MT-ND1*, *MT-ND2*, *MT-ND3*, *MT-ND4*, and *MT-ND4L*) and retinol binding protein 4 (*RBP4*) were unique to the top 500 attributes identified for the Grade comparison and MFs mapping to apolipoprotein E (*APOE*) and lipocalin 2 (*LCN2*) were unique to the top 500 attributes identified for the Growth comparison (**Table 2.7**). Similarly, MFs mapping to aggrecan (*ACAN*), ATPase sarcoplasmic/endoplasmic reticulum Ca²⁺ transporting 1 (*ATP2A1*), CD74 molecule (*CD74*), down-regulator of transcription 1 (*DRI*), glycerol-3-phosphate dehydrogenase 1 (*GPD1*), myosin heavy chain 4 (*MYH4*), and troponin T2, cardiac type (*TNNT2*) genes were uniquely identified among the top 500 attributes for the Strain comparison (**Table 2.7**). The remaining genes were mapped to from MFs that were in some combination shared between the top 500 for two comparisons and/or unique to one comparison. For example, a portion of the MFs that mapped to mitochondrially encoded cytochrome c oxidase subunit II (*MT-COX2*) were shared between the top 500 identified for Grade and Growth, some were unique to the top 500 for Grade, and some unique to Growth.

The MFs that mapped to each gene categorized as having complete homology (i.e., a twelve-out-of-twelve amino acid match) to the translated genome assembly sequence of both SB (paternal) and WB (maternal), exclusively to SB, exclusively to WB, or neither and therefore considered undetermined are presented in terms of absolute (i.e., of 821 MFs total) and relative (i.e., as a percentage of 100.00 % total for each gene) abundance in **Figures 2.9** and **2.10**, respectively. Over half of the unique MFs (477 of 821, 58.10 %) identified among the top 500 for each comparison were found to have complete homology to the translated genome assembly sequences for both SB and WB and these MFs mapped to twenty of the thirty-three genes (**Table 2.7**). The MFs that mapped to five of these genes were completely homologous to both SB and

WB sequences. The other fifteen genes were mapped to by a combination of MFs that were completely homologous to both SB and WB sequences, exclusive to SB, exclusive to WB, and/or categorized as undetermined, indicative of distinct alleles (**Table 2.7**). The twenty MFs (20 of 821, 2.44 %) exclusive to the SB (paternal) sequence mapped to three genes, however, MFs homologous to both SB and WB sequences, exclusive to WB, and/or categorized as undetermined also mapped to these three genes, suggesting that these may be alleles, however, it is unclear. The 191 MFs (191 of 821, 23.26 %) exclusively homologous to the WB (maternal) sequence mapped to seventeen of the thirty-three genes (**Table 2.7**). Eight of these genes were only identified by MFs exclusive to the translated WB sequence data, indicating distinct WB alleles, and MFs homologous to both SB and WB, exclusive to SB, and/or categorized as undetermined mapped to the other nine, suggesting that these may be alleles (**Table 2.7**). The remaining 133 MFs of 821 (16.20 %) categorized as undetermined mapped to nineteen of the thirty-three genes (**Table 2.7**). Sixteen of these genes were also mapped to by MFs identified in SB and WB sequence data, exclusive to SB, and/or exclusive to WB as mentioned above, and three of these genes were mapped by only undetermined MFs (9 of 133, 6.77 %) (**Table 2.7**).

Pathway Analysis

An interaction network for the eight genes mapped by MFs that were shared (i.e., identified) among the top 500 ranked for all three comparisons (223 of 821) is shown in **Figure 2.11**. These eight genes were all up-regulated in TG HSB and LG HSB for the Grade and Growth comparisons, respectively. Two of these eight genes, *APOA1* and *TF*, were up-regulated among DOM HSB, three in VA HSB (*CELA1*, *CELA2A*, and *CPBI*), two in FL HSB (*HBB* and *MT-ATP8*), and one in SC HSB (*HBA1*). All eight genes were down-regulated among TX HSB.

The Grade pathway analysis identified twenty-nine of the thirty-three genes as being up-regulated in TG HSB, including all mitochondrially-encoded genes that were exclusive to the translated WB (maternal) genome sequence assembly. The top (i.e., greatest enrichment and/or prediction of activity) canonical pathways identified for TG fish are shown in **Figure 2.12**, and the top five for this group are, in order of descending significance: Oxidative Phosphorylation (activated), Mitochondrial Dysfunction (activity not predicted), Granzyme A Signaling (inhibited), LXR/RXR Activation (activated), and FXR/RXR Activation (activity not predicted). Of the thirty-three input molecules those considered the top analysis-ready molecules (i.e., providing the most information for pathway and network building) that were up-regulated in TG fish and in order of descending expression value (most important to least important) are: *CPBI*, *CELA1*, *CELA2A*, *TF*, hemopexin (*HPX*), *APOA1*, *ACAN*, *LCN2*, *RBP4*, and *APOE*. The top analysis-ready molecules that were down-regulated in TG fish were, in order of ascending expression value: *ATP2A1*, aldolase, fructose-bisphosphate A (*ALDOA*), *GPDI*, and myosin heavy chain 8 (*MYH8*). A graphical summary of upstream regulators and causal networks predicted to underlie the observed patterns of gene expression in the data and that are determined to be significant (Fisher's Exact Test, $p \leq 0.05$) for the TG fish is shown in **Figure 2.13**.

As only four molecules were identified as up-regulated in Runt fish (*ALDOA*, *ATP2A1*, *GPDI*, and *MYH8*) the canonical pathways and other elements of the Core Analysis in IPA did not have the same depth as that for the TG fish. The top five canonical pathways for Runt HSB in order of descending significance: Glycerol-3-phosphate Shuttle, Calcium Signaling, Glycerol Degradation I, Calcium Transport I, and Sucrose Degradation V (Mammalian), however, predictions of activation and inhibition were not made as the limited dataset resulted in evidence (i.e., from the Ingenuity® Knowledge Base) to support either being equal across these genes and

pathways. Similarly, there were not enough connectable entities among the up-regulated genes in Runt HSB to generate a cohesive network diagram of relationships between upstream regulators and casual networks, although those associated to the gene expression patterns presented in Runt HSB include: dystrophin (DMD, regulator), insulin-like growth factor 1 (IGF1, regulator and network), dihydrotestosterone (regulator), SIX homeobox 1 (SIX1, regulator), zinc finger protein 224 (ZNF224, regulator), TNF superfamily member 12 (TNFSF12, network), glycogen (network), sonic hedgehog signaling molecule (SHH, network), and ERBB receptor feedback inhibitor 1 (ERRFI1, network).

Twenty-two of the thirty-three genes were up-regulated in HSB of the LG Growth group, all of which were also up-regulated in TG Grade fish. Four of the seven genes up-regulated in TG fish that were not up-regulated in LG fish were found to up-regulated in IN fish (*CELA1*, *CELA2A*, *CPBI*, and myosin binding protein C2, *MYBPC2*), and the remaining three in SM fish (myosin heavy chain 1, *MYH1*, 2, *MYH2*, and 4 *MYH4*). The top canonical pathways identified for LG fish are shown in **Figure 2.14**, and in order of descending significance are: Oxidative Phosphorylation (activated) and Mitochondrial Dysfunction (activity not predicted), Calcium Signaling (activity unknown), Dilated Cardiomyopathy Signaling Pathway (inhibited), and Granzyme A Signaling (inhibited). The top up-regulated, analysis-ready molecules for the LG HSB in order of descending expression value are: *TF*, *HPX*, *APOA1*, *ACAN*, *LCN2*, *APOE*, *RBP4*, hemoglobin subunit epsilon 1 (*HBE1*), *HBA1*, and *HBB*. The top down-regulated analysis-ready molecules for the LG HSB in order of ascending expression value are: *ATP2A1*, *GPD1*, *MYBPC2*, and *MYH4*. A graphical summary of upstream regulators and causal networks predicted to underlie the observed patterns of gene expression in the data for the LG fish is shown in **Figure 2.15**.

Seven of the thirty-three genes were up-regulated in HSB bred from DOM sires, all of which were also up-regulated in TG fish and LG fish: *ACAN*, *APOA1*, *APOE*, *DRI*, *HPX*, *RBP4*, and *TF*. The top canonical pathways identified for the DOM fish were identical to those identified for the LG fish including by predictions of activation or inhibition, as indicated by shading in **Figures 2.14** and **2.16**. The top up-regulated and analysis-ready molecules for the HSB with DOM sires in order of descending expression value are: *APOA1*, *TF*, *HPX*, *ACAN*, *APOE*, *CPB1*, *CELA1*, *RBP4*, *CELA2A*, and *DRI*. The top down-regulated analysis-ready molecules for DOM HSB in order of ascending expression value are: *MYBPC2*, *MYH4*, and *MYH2*. A graphical summary of upstream regulators and causal networks predicted to underlie the observed patterns of gene expression in the data for the DOM fish is shown in **Figure 2.17**.

Discussion

The bimodal distribution of weight and TL observed among the HSB is representative of a true effect underlying the variability of growth performance in these fish. There were no instances of HSB graded as TG or Runts at two months of age later being designated as SM or LG, respectively, for the Growth comparison at fifteen months of age, indicating that this underlying effect is, to some extent, persistent throughout the lifecycle of these fish, predictable at the juvenile grading phase of production, and perhaps heritable.

The Runt HSB seemed to display hyperplastic muscle fiber recruitment as these fish had greater number of fibers ≤ 25 μm compared to those in the TG group on average, although the average diameter of fibers was similar between groups. This result differs from the anticipated findings based on a similar study of HSB sampled at twenty-two months of age whereby the top ten percent of fish by weight had significantly more fibers of smaller diameter than the bottom

ten percent of fish by weight (Rajab, 2020). The HSB sampled by Rajab (2020), however, were seven months older and those larger in size weighed approximately 950 g/fish compared to approximately 630 g/fish for TG fish included in this study. It is possible that a change to muscle fiber recruitment does not occur until HSB are closer to or even past the typical timeframe of commercial harvest (sixteen to eighteen months) after they have reached a certain size. Further, the HSB examined by Rajab (2020) had been reared common garden and were not graded based upon projected growth to market size as the HSB described herein were. As such, it may be worth considering that the TG fish included in this study had already switched to hypertrophic muscle growth and the Runt HSB, which had been lagging in growth relative to TG fish as early as two months of age, remained in hyperplasia through at least the sampling period at fifteen months of age. The simple linear regression models show that there are significant relationships ($p \leq 0.0384$) between HSB weight and HSB TL and muscle fibers 10–25 μm , however, the R^2 values are not indicative of these characteristics being strongly determinant of the observed variation in number of muscle fibers of this diameter bin ($R^2 = 0.24\text{--}0.40$).

A strain effect was observed whereby HSB bred from different sires varied in growth performance outcomes (**Table 2.5** and **Figures 2.3, 2.4**). Although FL HSB offspring had the highest average weight, this strain was only represented by a few individuals all of which were determined to be produced by a single sire, subsequently limiting analysis and suggestive of a lack of hardiness in terms of survivability at the spermatozoa/fertilization and/or larval stage. The DOM sires had the greatest representation among the offspring sampled and these HSB offspring were of the largest size compared to the SC, TX, and VA wild-strain counterparts (**Tables 2.4** and **2.5**). The high representation of DOM offspring and their superior growth performance relative to HSB produced from wild sires indicate greater robustness to culture

conditions, a goal of domestication. Kenter et al. (2018) compared production metrics (growth, feed conversion ratio, fillet yield) of pure SB produced from SB broodfish of varying strain, including those strains described in the present study and, in some cases, using the same sires described here to produce offspring. Kenter et al. (2018) found that SB produced from DOM broodfish grew faster and had a lower FCR within the first year and into the second year of the trials compared to SB produced from wild-strain broodfish, although this growth advantage of SB produced from DOM broodfish was not maintained throughout the entirety (i.e., full two years) of the experiment. The superior performance of DOM SB through the first approximately year and a half described by Kenter et al. (2018) are consistent with the findings of the present study in HSB and indicate that there are benefits (e.g., growth rate) of SB domestication efforts and that these benefits are observable in both pure SB and HSB, however, there is not ample evidence to conclusively establish that effects of strain and/or domestication are underlying the runt effect observed here (i.e., there are likely multiple factors).

The patterns of expression of MFs mapping to the same gene but differently expressed between groups seems to support a strain effect to some extent. For example, undetermined MFs that were unique to the top 500 for the Strain comparison mapped to two genes, specifically six MFs that map to *MYH4* and the two that map to *DRI*. Four of the six MFs that map to *MYH4* concatenate and all have the greatest mean expression intensity among FL HSB and that is notably greater (5–13 fold) than the mean expression among HSB of other strains. This trend is similarly observed for the other two MFs that mapped to *MYH4*, which concatenate and maximum expression was identified among VA HSB and minimum among DOM HSB. The two MFs that map to *DRI* do not concatenate, although the greatest mean expression of one MF is in FL HSB and lowest in DOM HSB, and the other is greatest in DOM HSB and lowest in TX

HSB. It is possible that the inability to reconcile the amino acid sequence with either parent genome sequence is because the allele (or mutation) is unique to an individual sire or the wild strain of SB and therefore may not be represented in the reference SB genome assembly. Further investigation is required to determine whether these MFs and subsequently genes are strain-specific markers and/or the result of another phenomenon. A single undetermined MF that was shared between the top 500 attributes identified for the Strain and Growth comparison (greatest mean expression among TG HSB, LG HSB, and DOM HSB) mapped to a third gene, *HBE1*.

The genes mapped by MFs that were shared (i.e., identified) among the top 500 ranked for all three comparisons (223 of 821, indicated by an asterisk * in **Table 2.7**) are considered as being of high importance generally in HSB as they may be representative of both the runt effect, strain effect, interaction between the two, or other physiological processes underlying the phenotypes in HSB. The importance with which these genes are considered is supported by their shared functionality and/or involvement in broad cellular processes such as molecular transport, cell death and survival (necrosis and/or apoptosis), and connective tissue development. Of particular note is the *APOAI* gene, the role of which in cholesterol binding, lipid homeostasis and transport, and similar actions is well characterized across vertebrates (Tall and Yvan-Charvet, 2015). The *APOAI* gene was mapped to by the greatest number of MFs (168 of 821) compared to other genes in the dataset (**Figure 2.9**), and nearly all MFs aligned with complete homology to the translated sequence of both SB and WB reference genomes (**Table 2.7**). The majority of *APOAI*-MFs were shared between the top 500 for all three comparisons (102 MFs) and the prevalence of MFs associated to *APOAI* is reflected by this gene being identified as a top-analysis molecule for the TG, LG, and DOM groups examined more in-depth below. Moreover, *APOAI* has a high connectedness and relative centrality to other elements shown in

the interaction network of shared genes and predicted upstream regulators and downstream effects (**Figure 2.11**) and is associated to two of the significant canonical pathways identified for each comparison, LXR/RXR Activation (activated) and FXR/RXR Activation (activity not predicted) (**Figures 2.12, 2.14, and 2.16**).

The LXR/RXR pathway is directly involved in regulating lipid metabolism, inflammation, and catabolism of cholesterol to bile acid as shown alongside other pathway outputs in **Figure 2.18** (Mangelsdorf et al., 1992; Spector et al., 1979). Identifying this pathway as activated among skeletal muscle samples of HSB of superior growth performance (TG, LG, and/or DOM) is consistent with skeletal muscle being a primary regulator of lipid metabolism and the extent to which these processes are critical for growth (Morales et al., 2017). The FXR/RXR pathway shown in **Figure 2.19** is involved in linking the regulation of bile acids to lipoprotein, lipid, and glucose metabolism (Fojo and Brewer, 1992; Heyman et al., 1992). The persistence of high concentrations of bile acids has been linked to metabolic syndromes marked by insulin resistance and glucose tolerance and skeletal muscle disorders, such as sarcopenia, that result in muscle weakness, mass loss, and decline in strength, suggesting that the FXR/RXR pathway is likely activated in HSB demonstrating superior growth traits (Abrigo et al., 2022). Moreover, the three network diagrams of upstream regulators predicted for each group TG, LG, and DOM, include activation of sterol regulatory element-binding transcription factor 1 (SREBF1), which encodes proteins involved in lipid biosynthesis and more specifically, the activation of low density lipoprotein (LDL) receptor via signal transduction of insulin and IGF1, hormones known to enhance somatic growth (Streicher et al., 1996) (**Figures 2.13, 2.15, and 2.17**). In these networks for HSB, SREBF1 is inferred to interact with peroxisome proliferative activated receptor, gamma, coactivator 1 beta (PPARGC1B), known to stimulate transcription of

molecules known to support mitochondrial function, including the expression of mitochondrially-encoded genes, specifically estrogen receptor alpha (ER α), nuclear respiratory factor 1 (NRF1), and glucocorticoid receptor (GR) when in the presence of glucocorticoids (Kressler et al., 2002; Li et al., 2018; Ling et al., 2004).

The two genes mapped from MFs identified uniquely among those most important to the Growth comparison, *APOE* and *LCN2*, are also involved in lipid transport and had greater expression among HSB in the TG and LG groups (*APOE* was up-regulated in DOM HSB and *LCN2* was up-regulated in FL HSB). Specifically, both *APOE* and *LCN2* encode proteins from the lipocalin family, which transport hydrophobic molecules (e.g., steroids, lipids, retinoids) and are generally associated with metabolism, homeostasis, and immune response (Flower, 1996). The LXR/RXR and FXR/RXR pathways also include *APOE* (**Figures 2.18 and 2.19**), although *LCN2* is not associated to these or other top pathways. Expression of *apoE* in finfish has been examined as an indicator of lipid metabolism at various life stages. For example, *apoE* expression has been found to positively correlate to lipid nutrition and synthesis of very low density lipoprotein (VLDL) in turbot (*Scophthalmus maximus*) at the endo-exotrophic stage of development and to high feeding frequency in Yangtze sturgeon (*Acipenser dabryanus*) (Chen et al., 2021; Poupard et al., 2000). The characterization of proteins identified in teleosts that are structurally similar to the mammalian lipocalin 2 is most commonly in the context of supporting immune response in particular to bacterial infections (Rolig et al. 2018; Zhou et al., 2020). However, the role of lipocalin 2 in critical growth-related functions such as mediating appetite suppression during metabolic states of wasting and post-injury skeletal muscle regeneration has recently been examined more in-depth for mammalian counterparts (Olson et al., 2021; Rebalka et al., 2018). Our identification of a lipocalin 2-like gene in HSB suggests that further

investigation into functionality of the protein may lead to fruitful insight into the role of this gene broadly and specific differences in muscle use, fiber recruitment, and retention in fishes.

Although not depicted outright in the network diagrams generated from this dataset (**Figures 2.9, 2.13, 2.15, and 2.17**), leptin (LEP) was identified as an activated upstream regulator associated to the *APOA1* gene, along with four other genes up-regulated in HSB of superior growth performance (TG, LG, DOM, and/or FL HSB), *HPX*, myosin heavy chain 7, *MYH7*, *MT-ND1* and *MT-ND4L*. Leptin is a major metabolic hormone in vertebrates that has been shown to stimulate energy expenditure (glycolysis) through a signal transducer and activator of transcription 3 (STAT3)-mediated mechanism in tilapia (*Oreochromis mossambicus*), as well as IGF1 in HSB (Douros et al., 2018; Won et al., 2016). Not surprisingly, all three networks feature STAT3 as activated, and in TG and DOM HSB, as having an indirect activation interaction with IGF1 (**Figures 2.13, 2.15, and 2.17**).

In the TG and DOM groups, transforming growth factor beta 1 (TGFB1) is shown to directly activate STAT3 and indirectly interact with IGF1, which in turn indirectly interacts with STAT3 (**Figures 2.13 and 2.17**). In LG HSB, STAT3 is shown to be directly and indirectly activated by leukemia inhibitory factor (LIF) and tumor necrosis factor (TNF), respectively, and subsequently indirectly promote several functions, representative of an immune response (activation of cells, leukocytes, myeloid cells, neutrophils, phagocytes, and inflammatory response) (**Figure 2.15**). Both LIF and TNF encode pro-inflammatory cytokines that signal via the Janus kinase-signal transducer and activator of transcription (JAK/STAT) pathway (Seif et al., 2017). Although the JAK/STAT pathway has major roles in processes such as cell membrane-to-nucleus signaling and immune system development, STAT3 signaling also has a role in the regulation of several skeletal muscle cell types, including muscle stem cells,

myofibers, and macrophages (Sala and Sacco, 2016; Seif et al., 2017). Specifically, the activation of the JAK/STAT pathway in skeletal muscle is known to promote muscle hypertrophy by increasing satellite cell proliferation as well as in muscle wasting (Moresi et al., 2019). Further, high intensity and/or constant muscle movement, such as that of a fish swimming, is known to elicit an immune response regardless of the extent to which muscle fibers are damaged whereby leukocytes and neutrophils are released into circulation (Van de Vyver et al., 2016). Thus, the activation of these molecules is expected to coincide with a phagocytic response activated after high intensity movement and/or continuous muscle contraction, suggesting that the functions associated with immune response in TG, LG, and DOM HSB are a result of muscle contraction and development, rather than a response to a pathogen (Ahman et al., 2020; Kim & Lee, 2017; Van de Vyver et al., 2016)

The enrichment and activation of the Oxidative Phosphorylation pathway across groups of HSB exhibiting superior growth phenotypes is also consistent with frequent skeletal muscle contraction, as this pathway represents ATP production via oxidation-reduction reactions occurring in the mitochondria (Korzeniewski, 2003). Seven of the eight mitochondrially encoded genes that were mapped to from MFs exclusively homologous to the translated WB genome sequence (*MT-ATP6*, *MT-CO2*, *MT-ND1*, *MT-ND2*, *MT-ND3*, *MT-ND4*, and *MT-ND4L*) (**Table 2.7**) were associated to this pathway and several other top canonical pathways identified for TG, LG, and DOM HSB. Specifically, Mitochondrial Dysfunction (activity not predicted), Granzyme A Signaling (inhibited), and Sirtuin Signaling Pathway (inhibited) (**Figures 2.12, 2.14, and 2.16**), as well as Estrogen Receptor Signaling (activity unknown) in TG HSB (**Figure 2.12**) and the Neutrophil Extracellular Trap Signaling Pathway (activated) in LG and DOM HSB (**Figures 2.14 and 2.16**). Interestingly, although also up-regulated in DOM HSB, the mitochondrially-

encoded genes had the greatest mean expression in FL HSB. The up-regulation of these genes in HSB of superior size and that the MFs mapping to these genes were only identified among the translated WB sequence suggests that the superior growth phenotype observed in these FL HSB offspring is the result of a WB maternal effect or additive effect of both parental fish, rather than a SB-specific sire strain effect.

The activity of the Mitochondrial Dysfunction pathway could not be predicted based upon the expression of associated genes, however, as oxidative phosphorylation, which has been linked to a more effective metabolism in the context of fish undulation and skeletal muscle growth, is activated in HSB exhibiting superior growth traits and mitochondrial defects are associated to muscle atrophy and downstream impairments such as insulin resistance, it is suspected that the HSB exhibiting superior growth traits are not experiencing mitochondrial dysfunction (Chanséaume and Morio, 2009; Vo et al. 2022). Proper mitochondrial function is further supported by several of the predicted upstream regulators as mentioned above (e.g., PPARGC1B in **Figures 2.13, 2.15, and 2.17**) and the extent to which mitochondrial genes appear to underlie superior growth traits seen in these HSB.

The inhibition of the Granzyme A Signaling and Sirtuin Signaling pathways are consistent with the above inflammation response indicated in HSB exhibiting superior growth. Granzyme A is widely known for its role in apoptosis via single-stranded DNA nicking and sirtuins have been well-characterized as NAD⁺-activated signaling proteins in processes such as aging, metabolism, inflammation, DNA repair, and cellular response to stress (Qiu et al., 2010; van Daalen et al., 2020). However, granzymes have recently been found to have a role in mediating inflammation and sirtuins similarly in improvement of metabolic signaling pathways and the reduction of pro-inflammatory pathways (Vachharajani et al., 2016; van Daalen et al.,

2020). Sirtuin expression has been found to increase with calorie restriction and fasting, as this increases carbon oxidation in mitochondria subsequently producing NAD⁺ from NADH and activating sirtuin (Hebert et al., 2013). As such, the inhibition of the sirtuin signaling pathway in fish exhibiting a superior growth phenotype (i.e., those in the TG, LG, and DOM groups) is consistent with the presumption that those fish were not restricted by caloric intake generally or relative to the other HSB exhibiting the poor growth phenotype and/or runt effect. Moreover, sirtuins are known to deacetylate proteins that induce catabolism and inhibit anabolic processes in order to coordinate cellular energy stores (Nogueiras et al., 2012).

In addition to the metabolic processes described above, other activated pathways in TG, LG, and DOM HSB, such as the Oxytocin Signaling Pathway, are indicative of a mosaic of metabolic regulators supporting growth in HSB of a superior phenotype. The Oxytocin Signaling Pathway has been linked to smooth muscle contraction as well as the intracellular release of calcium, the uptake of which through the mitochondrial calcium uptake channel has been linked to control of processes such as glucose oxidation and regulation of myofiber size, and this is perhaps represented by the enrichment of the Calcium Signaling pathway in these fish (**Figures 2.12, 2.14, and 2.16**) (Gimpl and Fahrenholz, 2001; Sartori et al., 2021). The Oxytocin Signaling Pathway is also known to activate AMP-activated protein kinase (AMPK) signaling, which in turn mediates cellular metabolism by stimulating glucose uptake and lipid oxidation (Lee et al., 2008). Moreover, and although the activity for the Iron homeostasis signaling pathway in TG, LG, and DOM HSB was unknown based upon the molecules in the dataset and information in the Ingenuity® Knowledge Base, the five associated genes, (*HBA1*, *HBB*, *HBE1*, *HPX*, and *TF*) with known roles in hemoglobin, oxygen, and/or iron (ferrous and ferric) binding were all up-regulated in TG and LG HSB, as well as HSB strain groups with the larger offspring (*HBA1* was

up-regulated in SC HSB, *HBB*, *HBE1*, and *HPX* were up-regulated in FL HSB, and *TF* was up-regulated in DOM HSB), suggesting activation. Iron is essential for metabolism and other critical cellular processes such as oxygen transport, excess iron has been shown to lead to cirrhosis and cardiomyopathy and iron deficiency has been shown to impair skeletal muscle metabolism (De Domenico et al., 2008; Frise et al., 2022; Ganz et al., 2008). Thus, it is possible that one limiting factor to Runt HSB growth is an insufficient capacity to satisfy the iron requirement for proper erythropoiesis (red blood cell production), oxygen delivery, energy metabolism, and immune response subsequently limiting other critical pathways (e.g., LXR/RXR and FXR/RXR are associated) (Wang et al., 2023).

Conclusions

Overall, the patterns of gene expression identified as underlying superior growth in HSB suggest these fish exhibit optimal functioning of critical metabolic and other developmental processes in skeletal muscle tissue. The distinct expression patterns and association of mitochondrially-encoded genes (maternal) to critical pathways highlighted for TG, LG, and DOM HSB (i.e., those of superior growth performance), including those associated as upstream regulators, indicate that the optimal functioning of these processes in HSB offspring is in part due to a genetic component inherited from the WB parent (maternal). However, further investigation is required to determine the extent to which the effect identified here is driven by these genes alone, is influenced by the inheritance and/or expression of other genes (i.e., epistasis), and/or environmental factors (e.g., genotype-by-environment). Conversely, and although further investigation is required, it may be that Runt HSB that are not consuming enough feed and/or exhibiting optimal functioning of cellular processes are in a catabolic state

due to a potentially heritable metabolic syndrome. Although it is not possible to determine how much feed was consumed by the Runt HSB included in this present study, gene expression patterns indicative of limited metabolic and enzymatic function may be a result of a limited caloric, and therefore nutrient, intake. This catabolic state may be able to be mitigated through diet, as focus on lipoproteins and the regulation thereof is of increasing interest across aquaculture species in the context of immune health and stress tolerance in addition to growth. Specifically, supplementing feed with cholesterol has been shown to improve growth performance of white shrimp (*Litopenaeus vannamei*), nonspecific immune health of rainbow trout (*Oncorhynchus mykiss*), and the digestive health of Japanese flounder (*Paralichthys olivaceus*) (Long et al., 2013; Xu et al., 2018; Yan et al., 2020). Moreover, researchers have shown that the provision of supplemental cholesterol to Nile tilapia (*Oreochromis niloticus*) led to an increase in serum cortisol that facilitated increased Na⁺/K⁺ -ATPase activity of gill epithelial cells and ultimately improved stress response in a hyperosmotic environment (Xu et al., 2018).

Collectively the MFs identified through this study provide a group of biomarkers that should be further evaluated in terms of efficacy in predicting desirable production traits. The mitochondrial genes exclusive to the WB sequence (*MT-ATP6*, *MT-ATP8*, *MT-CO2*, *MT-ND1*, *MT-ND2*, *MT-ND3*, *MT-ND4*, and *MT-ND4L*) are candidates for future research as biomarkers of superior growth performance as they were up-regulated in HSB exhibiting this phenotype. The other twenty-five genes identified through motif-fingerprinting are potentially inherited from both parental fish, as many mapped to sequences present in the translated genome assemblies for both parental fish. Examining the expression of the associated MFs among these fish and/or broadening the scope of this analysis to include additional MFs than those reduced via ML

workflow may allow for a clearer determination of dominance between alleles. Conducting a similar study would allow for the determination of penetrance of these MFs, and therefore genes, between cohorts, as well as the potentially additive effects between mitochondrial genes and alleles associated to specific sire strains. Such a study using domesticated fish may also highlight variation between filial generations of these domesticated fish. It is important to note that although domesticated fish exhibiting superior culture traits such that expression patterns and subsequently identified pathways and upstream regulators are similar between DOM HSB and fish of superior growth performance (TG and LG), it is possible that the Grade and Growth comparisons are biased in favor of the patterns of expression in DOM fish due to their high representation among the sampled population. Additional research to determine specific variation among SB sire strain alleles by examining the MFs that are entirely absent or of highly varied patterns of expression between strains, may point to functional biomarkers representative of specific adaptations of a given strain based upon geographic location of origin. However, investigating this will ultimately be limited unless undetermined MFs can be resolved by alignment to sequence data (e.g., if obtained from wild strains) and identification of true (i.e., persistent) variation among these wild-origin sires. Ultimately, the identification of growth performance biomarkers as described here will greatly reduce the amount of time previously required to identify and select such marks and has facilitated the incorporation of high-throughput phenotype prediction (i.e., phonemics) into selection strategies of the breeding program.

Acknowledgements

We thank Dr. Andrew S. McGinty, Michael S. Hopper, Robert W. Clark, Dr. David Berlinsky, and Dr. Linas W. Kenter for their efforts in producing and rearing these fish, and to Dr. Linas W. Kenter for coordinating the microsatellite genotyping and parental assignment; Artesian Aquafarms LLC (South Mills, NC, USA) for providing white bass females; Dr. Jason Abernathy for contributions towards the *Morone chrysops* (white bass) genome sequence assembly; Royston Carter for helping to coordinate our collaboration with Orion Biosciences Inc.; Dr. David A. Baltzegar for assistance in data acquisition; and Dr. Sarah Rajab for assistance with muscle histology analysis. This work was supported by funding provided from the following sources: the Foundation for Food and Agriculture Research (FFAR) *New Innovator Award*, the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), the National Oceanic and Atmospheric Administration (NOAA) and National Sea Grant (E/2019-AQUA-02, a project to establish a Striped Bass Aquaculture Hub, the *StriperHub*), and North Carolina Sea Grant. Striped bass is a priority species for the USDA National Research Support Project 8 (NRSP-8; *National Animal Genome Research Project*) and funding to support the National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry was provided by the NRSP-8 and the USDA NIFA (Hatch Multistate Project), the USDA Agricultural Research Service (ARS, Harry K. Dupree Stuttgart National Aquaculture Research Center), the North Carolina State University College of Agriculture and Life Sciences and College of Sciences, the North Carolina Agricultural Foundation William White Endowment, and various industry stakeholders including Premex, Clay Chappell, Locals Seafood, and several others who wish to remain anonymous. L.K. Andersen also received support Center for Environmental Farming Systems Graduate Fellowship

program, the North Carolina State University, Raleigh, NC, Biotechnology Program (BIT), and the Coastal Conservation Association of North Carolina David and Ann Speaks Coastal Conservation Association Scholarship. This will be of the publications from to the North Carolina State University Pamlico Aquaculture Field Laboratory.

References

- Abrigo, J., Olguín, H., Gutierrez, D., Tacchi, F., Arrese, M., & Cabrera, D. Valero-Breton, M., Elorza, A.A., Simon, F., & Cabello-Verrugio, C. (2022). Bile Acids Induce Alterations in Mitochondrial Function in Skeletal Muscle Fibers. *Antioxidants*, 11(9), 1706. DOI: 10.3390/antiox11091706.
- Ahmad, S., Ahmad, K., Lee, E., Lee, Y., & Choi, I. (2020). Implications of Insulin-Like Growth Factor-1 in Skeletal Muscle and Various Diseases. *Cells*, 9(8), 1773. DOI: 10.3390/cells9081773.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403-410. DOI: 10.1016/S0022-2836(05)80360-2.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17), pp.3389-3402. DOI: 10.1093/nar/25.17.3389.
- Chanséaume, E., & Morio, B. (2009). Potential Mechanisms of Muscle Mitochondrial Dysfunction in Aging and Obesity and Cellular Consequences. *International Journal Of Molecular Sciences*, 10(1), 306-324. DOI: 10.3390/ijms10010306
- Chen, Y., Wu, X., Lai, J., Liu, Y., Song, M., Li, F., & Gong, Q. (2021). Characterization of two lipid metabolism-associated genes and their expression profiles under different feeding conditions in *Acipenser dabryanus*. *Aquaculture Reports*, 21, 100780. DOI: 10.1016/j.aqrep.2021.100780.
- Couch, C.R., Garber, A.F., Rexroad III, C.E., Abrams, J.M., Stannard, J.A., Westerman, M.E., and Sullivan, C.V. 2006. Isolation and characterization of 149 novel microsatellite DNA

- markers for striped bass, *Morone saxatilis*, and cross-species amplification in white bass, *Morone chrysops*, and their hybrid. *Molecular Ecology Notes*, 6(3):667-669. DOI: 10.1111/j.1471-8286.2006.01292.x.
- D'Abramo, L.R. and M.O. Frinsko. (2008). *Hybrid Striped Bass: Pond Production of Food Fish*. Southern Regional Aquaculture Center. Stoneville, Mississippi. Publication Number 303.
- Douros, J., Baltzegar, D., Reading, B., Seale, A., Lerner, D., Grau, E., & Borski, R. (2018). Leptin Stimulates Cellular Glycolysis Through a STAT3 Dependent Mechanism in Tilapia. *Frontiers In Endocrinology*, 9. DOI: 10.3389/fendo.2018.00465.
- Fojo, SS. and Brewer, H.B. (1992). Hypertriglyceridaemia due to genetic defects in lipoprotein lipase and apolipoprotein C-II. *Journal of internal medicine*, 231(6), pp.669-677. DOI: 10.1111/j.1365-2796.1992.tb01256.x.
- Flower, D. (1996). The lipocalin protein family: structure and function. *Biochemical Journal*, 318(1), 1-14. DOI: 10.1042/bj3180001.
- Frise, M., Holdsworth, D., Johnson, A., Chung, Y., Curtis, M., & Cox, P. et al. (2022). Abnormal whole-body energy metabolism in iron-deficient humans despite preserved skeletal muscle oxidative phosphorylation. *Scientific Reports*, 12(1). DOI: 10.1038/s41598-021-03968-4.
- Gimpl, G., and Fahrenholz, F. 2001. The oxytocin receptor system: structure, function, and regulation. *Physiological Reviews*, 81:629–683. DOI: 10.1152/physrev.2001.81.2.629.
- Hebert, A.S., Dittenhafer-Reed, K.E., Yu, W., Bailey, D.J., Selen, E.S., Boersma, M.D., Carson, J.J., Tonelli, M., Balloon, A.J., Higbee, A.J. and Westphall, M.S. 2013. Calorie restriction and SIRT3 trigger global reprogramming of the mitochondrial protein acetylome. *Molecular cell*, 49(1), pp.186-199. DOI: 10.1016/j.molcel.2012.10.024.

- Heyman, R., Mangelsdorf, D., Dyck, J., Stein, R., Eichele, G., Evans, R., & Thaller, C. (1992). 9-cis retinoic acid is a high affinity ligand for the retinoid X receptor. *Cell*, 68(2), 397-406. DOI: 10.1016/0092-8674(92)90479-v.
- Hodson, R.G. (1990). Hybrid Striped Bass Biology and Life History. Leaflet/Texas Agricultural Extension Service; no. 2416.
- Hodson, R.G., Sullivan, C.V., 1993. Induced maturation and spawning of domestic and wild striped bass, *Morone saxatilis* (Walbaum), broodstock with implanted GnRH analogue and injected hCG. *Aquaculture Research*. 24(3), 389-398. DOI: 10.1111/j.1365-2109.1993.tb00562.x.
- Jones, O.R., Wang, J., 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10(3), 551-555. DOI: 10.1111/j.1755-0998.2009.02787.x
- Kalinowski, S.T., Taper, M.L. and Marshall, T.C., 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular ecology*, 16(5), pp.1099-1106. DOI: 10.1111/j.1365-294X.2007.03089.x.
- Kenter, L.W., Kovach, A.I., Woods III, L.C., Reading, B.J., Berlinsky, D.L., 2018. Strain evaluation of striped bass (*Morone saxatilis*) cultured at different salinities. *Aquaculture*, 492, 215-225. DOI: 10.1016/j.aquaculture.2018.04.017.
- Kim, J., & Lee, J. (2017). Role of transforming growth factor- β in muscle damage and regeneration: focused on eccentric muscle contraction. *Journal Of Exercise Rehabilitation*, 13(6), 621-626. DOI: 10.12965/jer.1735072.536.

- Korzeniewski, B. (2003). Regulation of oxidative phosphorylation in different muscles and various experimental conditions. *Biochemical Journal*, 375(3), 799-804. DOI: 10.1042/bj20030882
- Krämer, A., Green, J., Pollard, Jr., J., and Tugendreich, S. 2014. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*. 30(4):523–30.
- Kressler, D., Schreiber, S.N., Knutti, D. and Kralli, A., 2002. The PGC-1-related protein PERC is a selective coactivator of estrogen receptor α . *Journal of Biological Chemistry*, 277(16), pp.13918-13925. DOI: 10.1074/jbc.M201134200.
- Lee, E., Uhm, K., Lee, Y., Kwon, J., Park, S., & Soo, K. (2008). Oxytocin stimulates glucose uptake in skeletal muscle cells through the calcium–CaMKK–AMPK pathway. *Regulatory Peptides*, 151(1-3), 71-74. DOI: 10.1016/j.regpep.2008.05.001.
- Li, Z., Cogswell, M., Hixson, K., Brooks-Kayal, A., & Russek, S. (2018). Nuclear Respiratory Factor 1 (NRF-1) Controls the Activity Dependent Transcription of the GABA-A Receptor Beta 1 Subunit Gene in Neurons. *Frontiers In Molecular Neuroscience*, 11. DOI: 10.3389/fnmol.2018.00285.
- Ling, C., Poulsen, P., Carlsson, E., Ridderstråle, M., Almgren, P., Wojtaszewski, J., Beck-Nielsen, H., Groop, L. and Vaag, A., 2004. Multiple environmental and genetic factors influence skeletal muscle PGC-1 α and PGC-1 β gene expression in twins. *The Journal of clinical investigation*, 114(10), pp.1518-1526. DOI: 10.1172/JCI21889.
- Long, X., Wang Q., Han X., Zhang X., and Deng J. (2013). Research advances on the mechanism of growth-promoting effects of dietary cholesterol in soybean meal-based diets of fish. *J. Anhui Agr. Sci.* 2954–2955. DOI: 10.3969/j.issn.0517-6611.2013.07.054.

- Mangelsdorf, D., Borgmeyer, U., Heyman, R., Zhou, J., Ong, E., & Oro, A. et al. (1992). Characterization of three RXR genes that mediate the action of 9-cis retinoic acid. *Genes & Development*, 6(3), 329-344. DOI: 10.1101/gad.6.3.329.
- Morales, P.E., Bucarey, J.L. and Espinosa, A. (2017). Muscle lipid metabolism: role of lipid droplets and perilipins. *Journal of diabetes research*. DOI: 10.1155/2017/1789395.
- Moresi, V., Adamo, S., & Berghella, L. (2019). The JAK/STAT Pathway in Skeletal Muscle Pathophysiology. *Frontiers In Physiology*, 10. DOI: 10.3389/fphys.2019.00500.
- Motulsky, H.J. and Brown, R.E., 2006. Detecting outliers when fitting data with nonlinear regression—a new method based on robust nonlinear regression and the false discovery rate. *BMC bioinformatics*, 7(1), pp.1-20. DOI: 10.1186/1471-2105-7-123.
- Mullis, A.W., Smith, J.M., 1990. Artificial spawning and fry production of striped bass and hybrids, in: Harrell, R.M., Kerby, J.H., Minton, R.V. (Eds.) *Culture and Propagation of Striped Bass and Its Hybrids*. American Fisheries Society, Bethesda, Maryland, pp. 7-16.
- National Research Council (NRC), 1996. *Guide for the care and use of laboratory animals*. The National Academies Press. Washington, DC. DOI: 10.17226/5140.
- Nogueiras, R., Habegger, K.M., Chaudhary, N., Finan, B., Banks, A.S., Dietrich, M.O., Horvath, T.L., Sinclair, D.A., Pfluger, P.T. and Tschöp, M.H. (2012). Sirtuin 1 and sirtuin 3: physiological modulators of metabolism. *Physiological reviews*. DOI: 10.1152/physrev.00022.2011.
- Olson, B., Zhu, X., Norgard, M., Levasseur, P., Butler, J., & Buenafe, A. et al. (2021). Lipocalin 2 mediates appetite suppression during pancreatic cancer cachexia. *Nature Communications*, 12(1). DOI: 10.1038/s41467-021-22361-3.

- Poupard, G., André, M., Durliat, M., Ballagny, C., Boeuf, G., & Babin, P. (2000). Apolipoprotein E gene expression correlates with endogenous lipid nutrition and yolk syncytial layer lipoprotein synthesis during fish development. *Cell And Tissue Research*, 300(2), 251-261. DOI: 10.1007/s004419900158.
- Qiu, X., Brown, K., Hirschey, M., Verdin, E., & Chen, D. (2010). Calorie Restriction Reduces Oxidative Stress by SIRT3-Mediated SOD2 Activation. *Cell Metabolism*, 12(6), 662-667. DOI: 10.1016/j.cmet.2010.11.015.
- Rahman, M.A., Bhadra, A., Begum, N., Islam, M.S. & Hussain, M.G. (1995). Production of hybrid vigor through cross breeding between *Clarias batrachus* Lin. and *Clarias gariepinus* Bur. *Aquaculture*, 138(1-4), pp.125-130. DOI: 10.1016/0044-8486(95)01076-9.
- Rahman, M.A., Lee, S.G., Yusoff, F.M., & Rafiquzzaman, S.M. (2018). Hybridization and its application in aquaculture. *Sex control in aquaculture*, 163-178. DOI: 10.1002/9781119127291.ch7.
- Rajab, S.A.S. (2020). An Integrated Metabolomic and Transcriptomic Approach for Understanding White Muscle Growth Regulation in Hybrid Striped Bass Aquaculture. North Carolina State University. Doctoral Dissertation. <https://repository.lib.ncsu.edu/handle/1840.20/38272>.
- Rebalka, I.A., Monaco, C.M., Varah, N.E., Berger, T., D'souza, D.M., Zhou, S., Mak, T.W. and Hawke, T.J., 2018. Loss of the adipokine lipocalin-2 impairs satellite cell activation and skeletal muscle regeneration. *American Journal of Physiology-Cell Physiology*, 315(5), pp.C714-C721. DOI: 10.1152/ajpcell.00195.2017.

- Rees, R.A., Harrell, R.M., 1990. Artificial spawning and fry production of striped bass and in hybrids, in: Harrell, R.M., Kerby, J.H., Minton, R.V. (Eds.) Culture and Propagation of Striped Bass and Its Hybrids. American Fisheries Society, Bethesda, Maryland, pp. 43-72.
- Rexroad, C., Vallejo, R., Coulibaly, I., Couch, C., Garber, A., Westerman, M., and Sullivan, C. 2006. Identification and characterization of microsatellites for striped bass from repeat-enriched libraries. *Conservation Genetics*, 7(6):971-982. DOI: 10.1007/s10592-006-9122-0.
- Rolig, A.S., Sweeney, E.G., Kaye, L.E., DeSantis, M.D., Perkins, A., Banse, A.V., Hamilton, M.K. and Guillemin, K., 2018. A bacterial immunomodulatory protein with lipocalin-like domains facilitates host–bacteria mutualism in larval zebrafish. *Elife*, 7, p.e37172. DOI: 10.7554/eLife.37172.
- Sala, D. & Sacco, A. (2016). Signal transducer and activator of transcription 3 signaling as a potential target to treat muscle wasting diseases. *Current Opinion In Clinical Nutrition And Metabolic Care*, 1. DOI: 10.1097/mco.0000000000000273.
- Sartori, R., Romanello, V., & Sandri, M. (2021). Mechanisms of muscle atrophy and hypertrophy: implications in health and disease. *Nature Communications*, 12(1). DOI: 10.1038/s41467-020-20123-1.
- Seif, F., Khoshmirsafa, M., Aazami, H., Mohsenzadegan, M., Sedighi, G., & Bahar, M. (2017). The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. *Cell Communication And Signaling*, 15(1). DOI: 10.1186/s12964-017-0177-y.

- Spector, A., Mathur, S., & Kaduce, T. (1979). Role of acylcoenzyme A: Cholesterol o-acyltransferase in cholesterol metabolism. *Progress In Lipid Research*, 18(1), 31-53. DOI: 10.1016/0163-7827(79)90003-1.
- Streicher, R., Kotzka, J., Müller-Wieland, D., Siemeister, G., Munck, M., Avci, H. and Krone, W. (1996). SREBP-1 Mediates Activation of the Low Density Lipoprotein Receptor Promoter by Insulin and Insulin-like Growth Factor-I. *Journal of Biological Chemistry*, 271(12), pp.7128-7133. DOI: 10.1074/jbc.271.12.7128.
- Tall, A., & Yvan-Charvet, L. (2015). Cholesterol, inflammation and innate immunity. *Nature Reviews Immunology*, 15(2), 104-116. DOI: 10.1038/nri3793.
- Trauner, M., Claudel, T., Fickert, P., Moustafa, T., & Wagner, M. (2010). Bile Acids as Regulators of Hepatic Lipid and Glucose Metabolism. *Digestive Diseases*, 28(1), 220-224. DOI: 10.1159/000282091.
- Vachharajani, V., Liu, T., Wang, X., Hoth, J., Yoza, B., & McCall, C. (2016). Sirtuins Link Inflammation and Metabolism. *Journal Of Immunology Research*, 2016, 1-10. DOI: 10.1155/2016/8167273
- van Daalen, K., Reijneveld, J., & Bovenschen, N. (2020). Modulation of Inflammation by Extracellular Granzyme A. *Frontiers In Immunology*, 11. DOI: 10.3389/fimmu.2020.00931.
- Van de Vyver, M., Engelbrecht, L., Smith, C. and Myburgh, K.H. (2016). Neutrophil and monocyte responses to downhill running: Intracellular contents of MPO, IL-6, IL-10, pstat3, and SOCS 3. *Scandinavian journal of medicine & science in sports*, 26(6), pp.638-647. DOI: 10.1111/sms.12497.

- Vo, T.A., Galloway, T.F., Arukwe, A., Edvardsen, R.B., Hamre, K., Karlsen, Ø., Rønnestad, I. and Kjørsvik, E. 2022. Effect of diet on molecular relationships between Atlantic cod larval muscle growth dynamics, metabolism, and antioxidant defense system. *Frontiers in marine science*. DOI: 10.3389/fmars.2022.814022.
- Wang, Z., Li, X., Lu, K., Wang, L., Ma, X., Song, K., & Zhang, C. (2023). Effects of dietary iron levels on growth performance, iron metabolism and antioxidant status in spotted seabass (*Lateolabrax maculatus*) reared at two temperatures. *Aquaculture*, 562, 738717. DOI: 10.1016/j.aquaculture.2022.738717.
- Won, E., Douros, J., Hurt, D., & Borski, R. (2016). Leptin stimulates hepatic growth hormone receptor and insulin-like growth factor gene expression in a teleost fish, the hybrid striped bass. *General And Comparative Endocrinology*, 229, 84-91. DOI: 10.1016/j.ygcen.2016.02.003.
- Xu, C., Li, E., Xu, Z., Su, Y., Lu, M., & Qin, J. Chen, L., and Wang, X. (2018). Growth and Stress Axis Responses to Dietary Cholesterol in Nile Tilapia (*Oreochromis niloticus*) in Brackish Water. *Frontiers In Physiology*, 9. DOI: 10.3389/fphys.2018.00254.
- Yan, M., Wang, W., Huang, X., Wang, X., & Wang, Y. (2020). Interactive effects of dietary cholesterol and phospholipids on the growth performance, expression of immune-related genes and resistance against *Vibrio alginolyticus* in white shrimp (*Litopenaeus vannamei*). *Fish & Shellfish Immunology*, 97, 100-107. DOI: 10.1016/j.fsi.2019.11.048.
- You, W., Guo, Q., Fan, F., Ren, P., Luo, X., & Ke, C. (2015). Experimental hybridization and genetic identification of Pacific abalone *Haliotis discus hannai* and green abalone *H. fulgens*. *Aquaculture*, 448, 243-249. DOI: 10.1016/j.aquaculture.2015.05.043.

Zhou, Z., Feng, C., Liu, X., & Liu, S. (2020). 3nLcn2, a teleost lipocalin 2 that possesses antimicrobial activity and inhibits bacterial infection in triploid crucian carp. *Fish & Shellfish Immunology*, 102, 47-55. DOI: 10.1016/j.fsi.2020.04.015.

Table 2.1. Descriptions of comparisons made between hybrid striped bass (HSB). Each comparison is generally referred to as the name listed in the left-most column. Strain refers to the geographic location of origin of the paternal fish (sire). The domestic sires were two years of age and bred at the North Carolina State University Pamlico Aquaculture Field Laboratory (PAFL) in Aurora, NC, USA. Other sires were three years of age and strain is based upon the waters of the states indicated. The sub-population size “(n=#)” of each group for a given comparison, referred to as “classes” in machine learning (ML) analyses, are also provided. The number prior to the forward slash is the number of the seventy-two (72) sacrificed fish, the number following the forward slash is the number of the forty (40) HSB of the sacrificed seventy-two for which sequencing data were generated. A brief description of each group is in the final column.

Comparison	Groups/Classes	Description
Grade *	Top Grade (TG, n=36/20) Runts (n=36/20)	larger grade of HSB smaller grade of HSB
Growth *	Large (n=19/13) Small (n=22/17) Intermediate (n=31/10)	reached/exceeded the mean weight for TG fish below the mean weight for Runt fish between the mean weight for Runt and TG fish
Strain	Domestic (DOM, n=22/12) Florida (FL, n=4/4) South Carolina (SC, n=11/6) Texas (TX, n=15/9) Virginia (VA, n=19/9)	offspring born to domestic sires offspring born to Florida sires offspring born to South Carolina sires offspring born to Texas sires offspring born to Virginia sires

*HSB were graded at two months of age into two groups: Top Grade (TG) and Runts (R), whereby TG fish were larger and R fish were smaller. At fifteen months of age HSB were harvested and sampled, at which time HSB that weighed above the mean weight for the TG group were designated as Large (LG), HSB below the mean weight of R fish were designated as Small (SM), and fish that fell between the two means were designated as Intermediate (IN).

Table 2.2. Outcomes of negative control machine learning (ML) analyses of hybrid striped bass (HSB) expression data examined through two comparisons of growth performance at two months of age (Grade) and final harvest at fifteen months of age (Growth), and geographic origin of paternal parent, or sire strain (Strain). Each negative control was conducted by processing the optimal dataset* with randomized labels (group identifiers for a given comparison) associated with expression data ten separate times with each of the eight cross-validated algorithms (four algorithms x two cross-validated strategies). True learning by algorithms is said to occur if the mean percent correct classification of each negative control run is approximately what can be yielded through random chance. The number of attributes included in the optimal dataset and classes (groups) for each comparison are indicated followed by the percent correct classification that can be predicted under the assumption of random probability based on the number of classes (possible outcomes). The grand mean \pm standard deviation (SD) percent correct classification of all cross-validated ML algorithms on randomized optimal datasets for a given comparison is reported in the right-most column.

Comparison	Optimal Attributes	Classes	Correct Classification Assuming Random Probability (%)	Actual Correct Classification (Mean \pm SD %)
Grade	200	2	50.00	49.87 \pm 2.21
Growth	100	3	33.33	39.23 \pm 3.67
Strain	150	5	20.00	24.75 \pm 3.12

*To establish the optimal dataset for a given comparison, values based on the information gained by the algorithm from the consideration of each attribute (i.e., gene transcript) are calculated and attributes are subsequently assigned a rank from low entropy (high information) to high entropy (little or no information). The lower-ranking of the ranked attributes (i.e., those assigned an information gain value greater than 0.00) are then removed (approximately 50 at a time) from the dataset used for training and testing by all ML algorithms. Algorithm performance measured as percent correct classification is then plotted against the number of input top-ranked attributes to determine points of model performance improvement or deterioration with the iterative exclusion of attributes. The number of top-ranked attributes included in input and yielding optimal or close-to optimal classification performance across multiple cross-validated algorithms is designated as the optimal dataset.

Table 2.3. Mean \pm standard deviation weight (g) and total length (TL, mm) of a study population of aquacultured hybrid striped bass (HSB) and subpopulations thereof grouped for additional analyses and comparisons. The “Grade” comparison describes Top Grade (TG) and Runt HSB who at two months of age were sorted by size if expected to reach or exceed market size of ~680 g (1.5 lbs) by final harvest or not, respectively. HSB were grouped by weight at final harvest (fifteen months) as Large (LG) if above the mean TG weight, Small (SM) if below the mean Runt weight, and Intermediate (IN) if between the two means for the “Growth” comparison. Group sizes are provided in parentheses following the weight values for each comparison and subpopulation. Column “All (N=752)” represents the entire study population, “Sampled (n=72)” are HSB of this population sacrificed for white muscle tissue analyses, specifically, gene sequencing (“Sequenced (n=40)”) and histology (“Histology (n=18)”). Student’s *t*-Test (Grade) or one-way ANOVA and Tukey’s HSD post-hoc test (Growth, see: differentiating letters) were used to determine statistically significant differences between groups, the greatest *p*-value calculated for a comparison (i.e., between pairs) is provided in italic or as four asterisks (****) if $p \leq 0.0001$.

Metric	Study (N=752)	Sampled (n=72)	Sequenced (n=40)	Histology (n=18)
Grade Weight (g)	****	****	****	<i>p=0.0013</i>
Top Grade	633.45 \pm 107.35 (377)	633.22 \pm 106.12 (36)	658.65 \pm 122.83 (20)	589.22 \pm 84.81 (9)
Runt	417.95 \pm 86.65 (375)	428.58 \pm 84.10 (36)	392.15 \pm 80.56 (20)	442.89 \pm 74.018 (9)
Total Length (mm)	****	****	****	****
Top Grade	346.73 \pm 17.11	345.33 \pm 16.20	347.90 \pm 19.10	338.33 \pm 12.042
Runt	301.62 \pm 19.072	301.81 \pm 18.80	293.15 \pm 18.41	304.00 \pm 13.31
Growth Weight (g)	****	****	****	<i>p=0.0003</i>
Large	718.67 \pm 71.52 ^A (187)	713.16 \pm 69.27 ^A (19)	732.00 \pm 75.87 ^A (13)	683.67 \pm 36.96 ^A (3)
Intermediate	529.33 \pm 59.48 ^B (341)	528.29 \pm 58.16 ^B (31)	525.00 \pm 51.50 ^B (10)	538.33 \pm 51.10 ^B (9)
Small	360.42 \pm 53.97 ^C (224)	377.18 \pm 51.83 ^C (22)	367.65 \pm 55.51 ^C (17)	398.83 \pm 24.14 ^C (6)
Total Length (mm)	****	****	****	<i>p=0.0159</i>
Large	359.04 \pm 10.88 ^A	356.74 \pm 11.035 ^A	358.77 \pm 11.78 ^A	349.67 \pm 8.96 ^A
Intermediate	327.59 \pm 14.19 ^B	325.77 \pm 13.52 ^B	324.70 \pm 12.51 ^B	327.67 \pm 12.19 ^B
Small	290.14 \pm 14.69 ^C	291.82 \pm 15.039 ^C	288.82 \pm 15.81 ^C	297.17 \pm 7.22 ^C

Table 2.4. Representation of striped bass (SB) sires among hybrid striped bass (HSB) offspring (N=72). The strain, or geographic location of sire origin is as follows: “DOM” are two year old, 5th generation domestic SB from the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA; “FL”, “SC”, “TX”, and “VA” refer to the three year old SB produced in hatcheries from SB caught in the waters of Florida, South Carolina, Texas, and Virginia, respectively; “N/A” represents the unknown sire of a HSB offspring that could not be matched via microsatellite genotyping. The number of sires represented among sacrificed HSB is listed in the “No.” column followed by the average \pm standard deviation sire weight (kg); DOM and VA sires were the only groups to significantly differ in weight (one-way ANOVA, Tukey’s HSD, $p=0.0061$) indicated by a double asterisk (**). The number of sacrificed HSB offspring sires of each strain produced is provided in the Progeny column and expressed as a percentage of the subpopulation in parentheses. The number of these sacrificed HSB that were included in sequencing analysis ($n=40$ of 72) are included in *italic*. The offspring that fell into Top Grade (TG) and Runt (R) groups for the Grade* comparison as well as Large (LG), Intermediate (IN), and Small (SM) groups for the Growth* comparison are provided in the rightmost columns.

Strain	No.	Weight (kg)	Progeny	Grade		Growth		
				TG	R	LG	IN	SM
DOM	4	2.76 \pm 0.16**	22 (30.56 %)	16	6	12	7	3
			<i>12</i>	8	4	8	<i>1</i>	3
FL	1	3.26	4 (5.56 %)	4	0	3	1	0
			<i>4</i>	4	-	3	<i>1</i>	-
SC	3	3.94 \pm 0.38	11 (15.28 %)	3	8	1	5	5
			<i>6</i>	2	4	<i>1</i>	2	3
TX	4	3.95 \pm 0.77	15 (20.83 %)	4	11	1	7	7
			<i>9</i>	2	7	0	3	6
VA	4	4.46 \pm 0.57**	19 (26.39 %)	9	10	2	10	7
			<i>9</i>	4	5	<i>1</i>	3	5
N/A	-	-	1 (1.39 %)	0	1	0	0	0
			<i>0</i>	-	-	-	-	-

*HSB were graded at two months of age into two groups: Top Grade (TG) and Runts (R), whereby TG fish were larger and R fish were smaller. At fifteen months of age HSB were harvested and sampled, at which time HSB that weighed above the mean weight for the TG group were designated as Large (LG), HSB below the mean weight of R fish were designated as Small (SM), and fish that fell between the two means were designated as Intermediate (IN).

Table 2.5. Mean \pm standard deviation weight (g) and total length (TL, mm) of hybrid striped bass (HSB, N=72) produced from different strains (i.e., geographic location of origin) of SB sires. Specifically, “DOM” are two-year-old, 5th generation domestic SB from the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA; “FL”, “SC”, “TX”, and “VA” refer to the three year old SB produced in hatcheries from SB caught in the waters of Florida, South Carolina, Texas, and Virginia, respectively. At two months of age HSB were sorted by size into Top Grade (TG) or Runt groups based upon expectation to reach or exceed market size of ~680 g (1.5 lbs) by final harvest or not, respectively. The number of HSB (of 72 total) produced from sires of each strain is provided as “N=#” followed by the number of offspring belonging to the TG and Runt groups presented as “(# TG/#Runt)”. Grand means of weight and TL for all offspring of a given strain group are in boldface and means for TG and Runt HSB of a specific strain group are listed below grand means. Differentiating letters between grand means indicate significant differences between strain groups (one-way ANOVA and Tukey’s HSD post hoc test for weight: DOM vs SC $p=0.0070$, DOM vs TX $p=0.0063$, DOM vs VA $p=0.0266$ and TL: DOM vs SC $p=0.0240$). The p -values for Student’s t -Tests comparing weight and TL between TG and Runt fish of each strain are provided below the respective values in italic font. One HSB of the sample population described here is not represented as parentage could not be assigned, this fish weighed 441.00 g, was 310.00 mm in TL and had been in the Runt group.

	DOM N=22 (16/6)	FL N=4 (4/0)	SC N=11 (3/8)	TX N=15 (4/11)	VA N=19 (9/10)
Weight (g)	618.05 \pm 145.99^A	656.25 \pm 71.014^{AB}	456.64 \pm 128.48^B	470.20 \pm 115.24^B	499.26 \pm 107.70^B
Top Grade	684.19 \pm 108.65	<i>see above</i>	589.33 \pm 136.66	557.50 \pm 87.90	580.67 \pm 74.28
Runts	441.67 \pm 52.68	-	406.88 \pm 88.71	438.45 \pm 110.087	426.00 \pm 75.51
	<i>p<0.0001</i>		<i>p=0.0261</i>	<i>p=0.0751</i>	<i>p=0.0003</i>
Total Length (mm)	338.14 \pm 27.28^A	340.75 \pm 11.64^{AB}	308.27 \pm 28.29^B	313.53 \pm 26.80^{AB}	320.58 \pm 25.027^{AB}
Top Grade	352.19 \pm 15.77	<i>see above</i>	335.33 \pm 26.50	341.75 \pm 14.36	340.11 \pm 14.12
Runts	300.67 \pm 7.92	-	298.13 \pm 22.61	303.27 \pm 22.57	303.00 \pm 18.73
	<i>p<0.0001</i>		<i>p=0.0443</i>	<i>p=0.0078</i>	<i>p=0.0002</i>

Table 2.6. Histological analysis of hybrid striped bass (HSB, N=18) white muscle tissue fibers between groups of projected growth performance (Grade). HSB were graded at two months of age into two groups, Top Grade and Runts, based on anticipated growth to meet or exceed market size (~680 g, or 1.5 lbs) or not, respectively, by final harvest at fifteen months of age. ImageJ software (Fiji, v.1.52a, National Institutes of Health, NIH, Bethesda, USA) was used to count and determine the area (μm) of all fibers entirely within the field of view (i.e. not marginated) of each image collected in triplicate for all individuals to evaluate hyperplastic (fiber amount) and hypertrophic (fiber area) muscle growth. The diameter of each scored fiber was calculated as a geometric derivative of its area to compute the average fiber diameter and the frequency of scored fibers of a certain diameter range for each slide image. Triplicate images were scored duplicate and the average \pm standard deviation values here are grand means of the average number and diameter calculated for each individual (i.e., slide) belonging to a given group by both scorers. A Student's *t*-Test was used to compare values between groups and an asterisk is beside p-values indicating a statistically significant difference ($\alpha=0.05$).

	Top Grade (n=9)	Runts (n=9)	p-value
Weight (g)	589.22 \pm 84.81	442.89 \pm 74.018	0.0013*
Total Length (mm)	338.33 \pm 12.04	304.00 \pm 13.31	0.0804
Number of Fibers Scored	114.44 \pm 26.67	115.22 \pm 25.95	0.9508
Fiber Diameter (μm)	63.73 \pm 6.84	63.58 \pm 9.36	0.9684
Number of 10-20 μm Fibers	19.72 \pm 12.27	39.11 \pm 23.90	0.0513
Number of 10-25 μm Fibers	30.11 \pm 16.39	56.78 \pm 29.11	0.0329*
Number of 10-50 μm Fibers	148.56 \pm 67.63	139.94 \pm 58.71	0.7768
Number of 10-75 μm Fibers	233.06 \pm 76.31	232.33 \pm 74.48	0.9840
Number of 10-100 μm Fibers	291.94 \pm 93.63	305.39 \pm 94.34	0.7655
Number of Fibers >75 μm	110.28 \pm 12.71	113.33 \pm 16.90	0.6709
Number of Fibers >100 μm	51.39 \pm 16.10	40.28 \pm 19.08	0.2011
Number of Fibers >150 μm	0.33 \pm 0.71	1.50 \pm 2.41	0.1958

Table 2.7. Official symbols and names of human (*Homo sapiens*) orthologs to the thirty-three (33) genes identified as those encoding the striped bass (*Morone saxatilis*, SB) and white bass (*M. chrysops*, WB) proteins mapped to from unique, twelve amino acid-long motif fingerprints (MFs) identified as informative to the differentiation of fish by growth performance (Grade and Growth comparisons) and/or paternal geographic location of origin (Strain comparison) via application of a machine learning (ML) workflow. The ML workflow reduced dimensionality of a 15,000 MFs dataset by identifying the MFs that yielded optimum classification performance (i.e., most individuals, or instances, correctly assigned comparison groups, or classes) by four distinct, cross-validated ML algorithms. There were 821 unique MFs among the 500 most-informative for each comparison. The number of MFs mapping to each gene is provided in the MFs column and the percentage of all 821 MFs is provided in parentheses. MFs were concatenated into longer amino acid sequences and mapped to complete proteins in both the translated SB and WB genome assemblies (“Both” column), only the SB assembly (“SB” column), only the WB assembly (“WB” column), or were not able to be mapped completely (i.e., a 12 of 12 amino acid alignment) to either and are therefore considered undetermined (“Undetm.” column). MFs counted as “Both” represent those that either completely aligned to both the SB and WB assemblies.

Symbol	Name	MFs (% Total)	Both	SB	WB	Undetm.
<i>ACAN</i> ^S	aggrecan	1 (0.12 %)	1	0	0	0
<i>ALDOA</i>	aldolase, fructose-bisphosphate A	18 (2.19 %)	18	0	0	0
<i>APOA1</i> [*]	apolipoprotein A-I	168 (20.46 %)	162	0	0	6
<i>APOE</i> ^{Gr}	apolipoprotein E	7 (0.85 %)	7	0	0	0
<i>ATP2A1</i> ^S	ATPase sarcoplasmic/endoplasmic reticulum Ca ²⁺ transporting 1	17 (2.07 %)	16	0	0	1
<i>CD74</i> ^S	CD74 molecule	3 (0.37 %)	2	0	0	1
<i>CELA1</i> [*]	chymotrypsin-like elastase 1	132 (16.08 %)	108	0	4	20
<i>CELA2A</i> [*]	chymotrypsin-like elastase 2A	56 (6.82 %)	39	0	6	11
<i>CPB1</i> [*]	carboxypeptidase B1	57 (6.94 %)	43	0	0	14
<i>DRI</i> ^S	down-regulator of transcription 1	2 (0.24 %)	0	0	0	2
<i>GPD1</i> ^S	glycerol-3-phosphate dehydrogenase 1	2 (0.24 %)	2	0	0	0
<i>HBA1</i> [*]	hemoglobin subunit alpha 1	46 (5.6 %)	0	0	37	9
<i>HBB</i> [*]	hemoglobin subunit beta	49 (5.97 %)	0	13	13	23

Table 2.7. (continued).

<i>HBE1</i>	hemoglobin subunit epsilon 1	1	(0.12 %)	0	0	0	1
<i>HPX</i>	hemopexin	3	(0.37 %)	1	0	0	2
<i>LCN2</i> ^{Gr}	lipocalin 2	6	(0.73 %)	4	0	0	2
<i>MT-ATP6</i>	ATP synthase membrane subunit 6	14	(1.71 %)	0	0	14	0
<i>MT-ATP8</i> [*]	ATP synthase membrane subunit 8	10	(1.22 %)	0	0	10	0
<i>MT-CO2</i>	cytochrome C oxidase II	5	(0.61 %)	0	0	5	0
<i>MT-ND1</i> ^G	NADH:ubiquinone oxidoreductase core subunit 1	14	(1.71 %)	0	0	14	0
<i>MT-ND2</i> ^G	NADH:ubiquinone oxidoreductase core subunit 2	13	(1.58 %)	0	0	13	0
<i>MT-ND3</i> ^G	NADH:ubiquinone oxidoreductase core subunit 3	10	(1.22 %)	0	0	10	0
<i>MT-ND4</i> ^G	NADH:ubiquinone oxidoreductase core subunit 4	23	(2.8 %)	0	0	23	0
<i>MT-ND4L</i> ^G	NADH:ubiquinone oxidoreductase core subunit 4L	17	(2.07 %)	0	0	17	0
<i>MYBPC2</i>	myosin binding protein C2	27	(3.29 %)	20	0	2	5
<i>MYH1</i>	myosin heavy chain 1	16	(1.95 %)	4	0	3	9
<i>MYH2</i>	myosin heavy chain 2	8	(0.97 %)	2	0	6	0
<i>MYH4</i> ^S	myosin heavy chain 4	6	(0.73 %)	0	0	0	6
<i>MYH7</i>	myosin heavy chain 7	1	(0.12 %)	1	0	0	0
<i>MYH8</i>	myosin heavy chain 8	19	(2.31 %)	6	0	11	2
<i>RBP4</i> ^G	retinol binding protein 4	7	(0.85 %)	3	2	0	2
<i>TF</i> [*]	transferrin	45	(5.48 %)	23	5	3	14
<i>TNNT2</i> ^S	troponin T2, cardiac type	18	(2.19 %)	15	0	0	3

“MT-” indicates gene is encoded by mitochondrial DNA and may be listed with “Mitochondrially Encoded” at start of official name. Superscripts indicate all MFs were exclusively among the 500 most informative for a comparison Grade (^G), Growth (^{Gr}), or Strain (^S). An asterisk (*) indicates genes were mapped to by MFs shared among the top 500 for all three comparisons (223 MFs were shared).

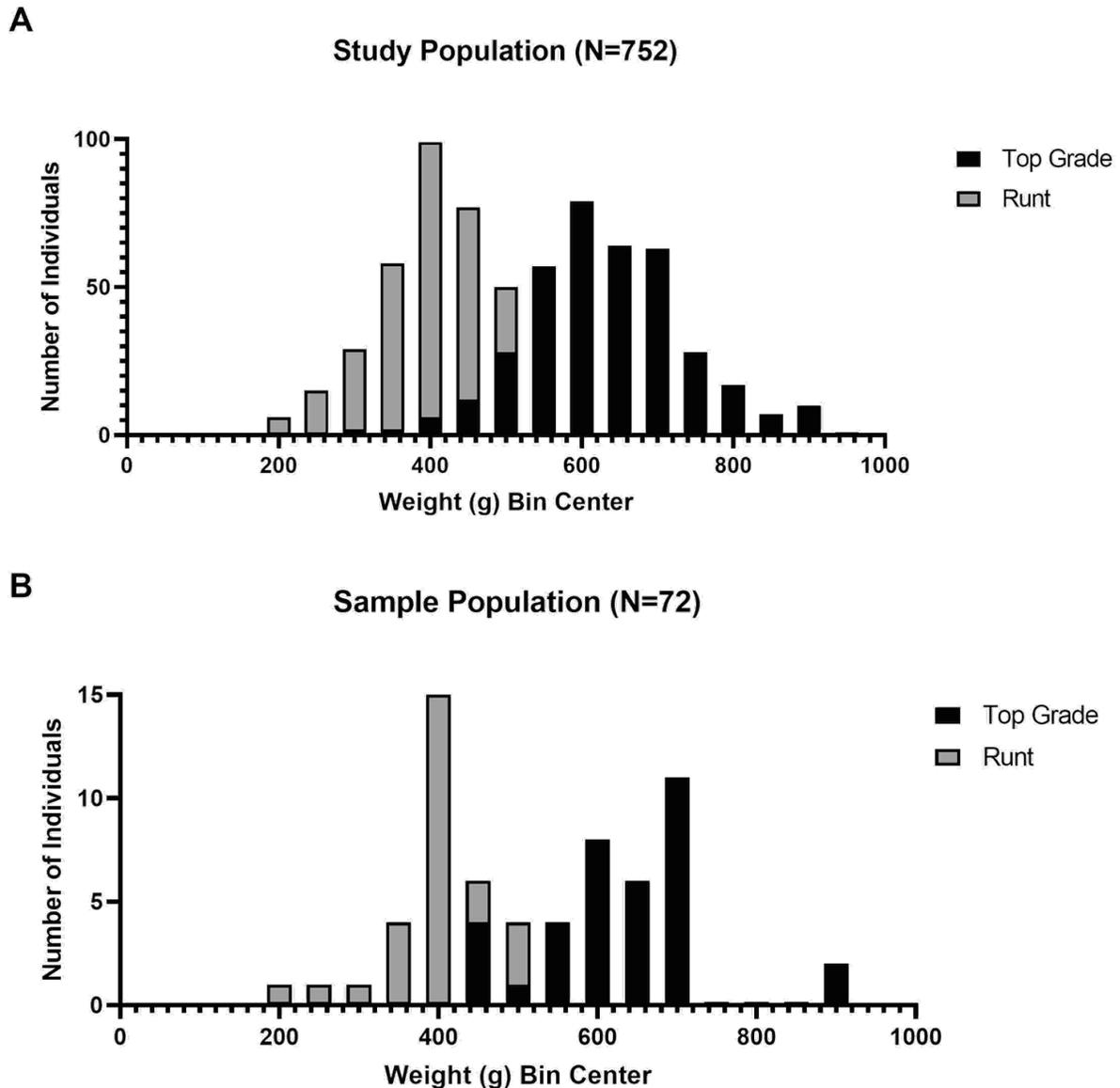


Figure 2.1. Frequency distribution of hybrid striped bass (HSB) weight (g) measured from (A) the complete study population (N=752 HSB) and (B) the individuals sacrificed for additional analyses (n=72). HSB were produced and raised at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA and graded at two months of age into two groups, Top Grade and Runts, based on anticipated growth to reach or exceed market size for these fish (~680 g), or not, respectively. The weights shown were recorded at final harvest (fifteen months of age). The mean \pm standard deviation weight of the entire study population was 525.84 ± 145.36 g (sample population was 530.90 ± 140.19 g). Top Grade HSB from the study population (n=377) weighed 633.45 ± 107.35 g on average which significantly differed from the Runt HSB (n=375, 417.95 ± 86.65 g) and this difference was similarly observed between those in the sample population (Top Grade, n=36, 633.22 ± 106.12 g; Runt, n=36, 428.58 ± 84.10 g) (Student's *t*-Test, $p < 0.0001$).

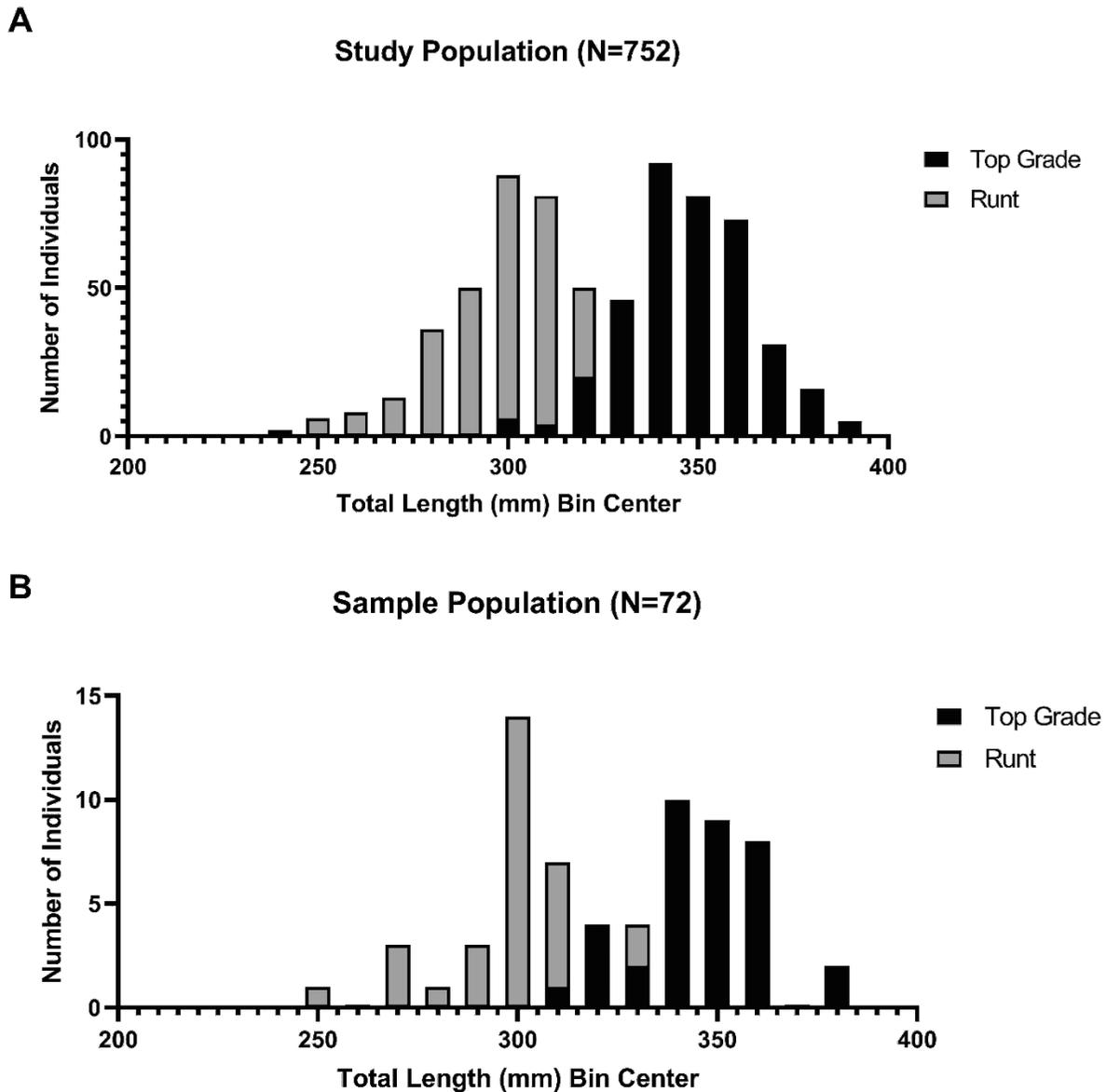


Figure 2.2. Frequency distribution of hybrid striped bass (HSB) total length (TL, mm) measured from (A) the complete study population (N=752 HSB) and (B) the individuals sacrificed for additional analyses (n=72). HSB were produced and raised at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL) in Aurora, NC, USA and graded at two months of age into two groups, Top Grade and Runts, based on anticipated growth to reach or exceed market size for these fish (~680 g), or not, respectively. The TLs shown were recorded at final harvest (fifteen months of age). The mean \pm standard deviation TL of the entire study population was 324.20 ± 28.93 mm (sample population was 323.57 ± 28.00 mm). Top Grade HSB from the study population (n=377) were 346.73 ± 17.11 mm on average which significantly differed from the Runt HSB (n=375, 301.62 ± 19.072 mm) and this difference was similarly observed between those in the sample population (Top Grade, n=36, 345.33 ± 16.20 mm; Runt, n=36, 301.81 ± 18.80 mm) (Student's *t*-Test, $p < 0.0001$).

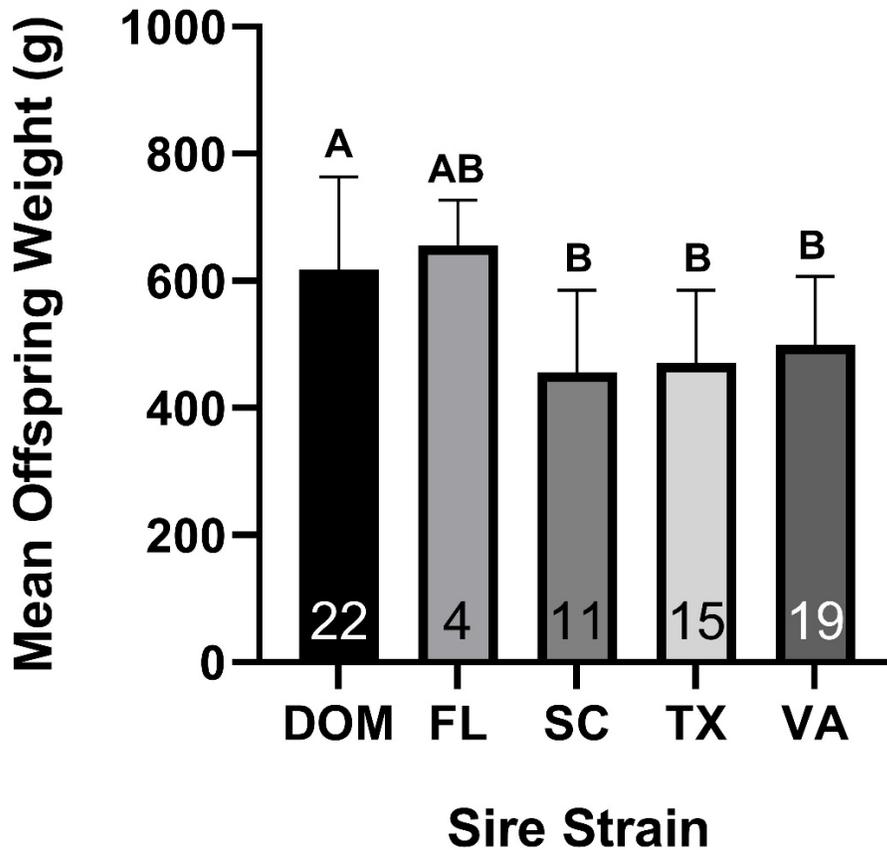


Figure 2.3. The average weight of hybrid striped bass (HSB) offspring produced using male striped bass (SB) of different geographic origin (strain). The “DOM” strain represents domestic SB males raised in captivity at the North Carolina State University’s Pamlico Aquaculture Field Lab (NCSU PAFL) in Aurora, NC, USA. The remaining strains were produced in hatcheries from wild-caught SB from the waters of Florida (FL), South Carolina (SC), Texas (TX), and Virginia (VA). A one-way ANOVA test was used to determine if the weight of HSB produced from sires of different strains significantly differed ($p < 0.0001$). The value within each bar towards the x-axis represents the number of HSB in each group. Specific differences between the weights of offspring grouped by sire strain were identified via Tukey’s HSD post-hoc test, the results of which are indicated in the figure by differentiating letters and written here: The mean \pm standard deviation (SD) weight for HSB produced from DOM sires was 618.05 ± 145.99 g and did not significantly differ from the FL offspring (weight was 656.25 ± 71.01 g). DOM offspring were significantly greater in weight than HSB produced by SC sires (offspring weight was 456.64 ± 128.48 g; $p = 0.0070$), TX sires (offspring weight was 470.20 ± 115.24 g; $p = 0.0063$), and VA sires (offspring weight was 499.26 ± 107.70 g; $p = 0.0266$).

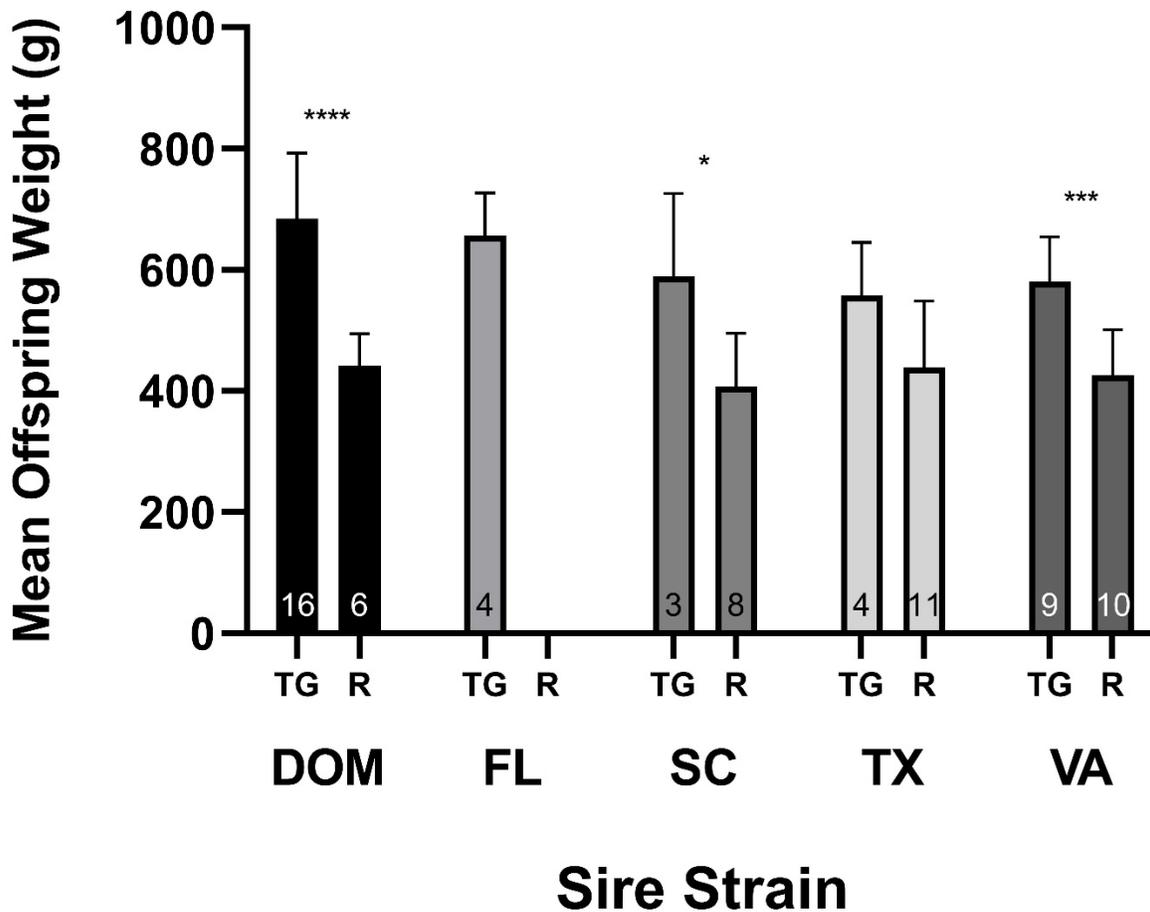
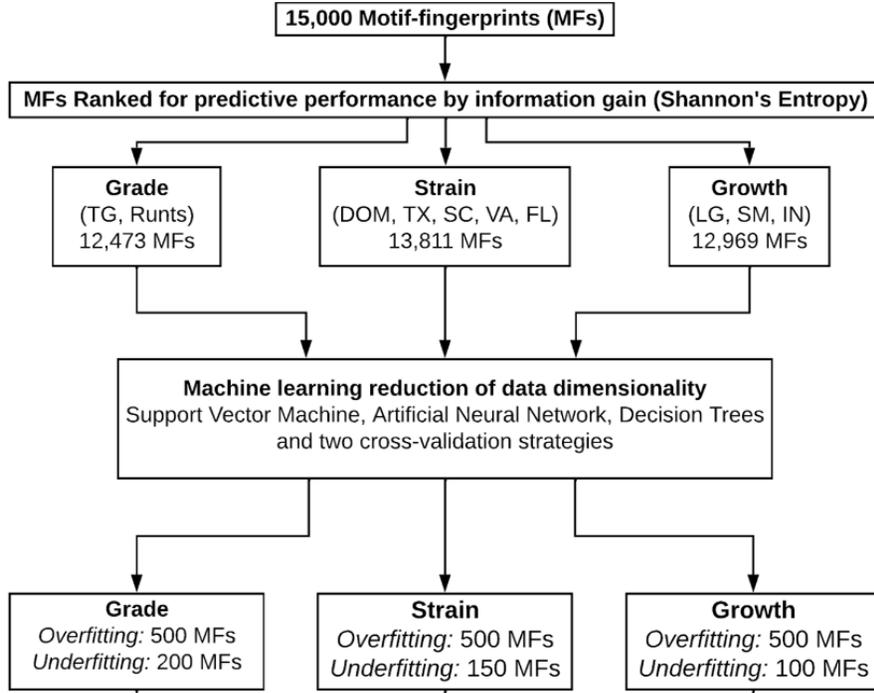


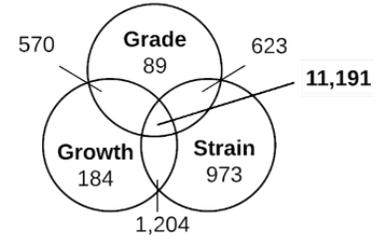
Figure 2.4. A comparison of the average weight of hybrid striped bass (HSB) offspring produced using male striped bass (SB) of different geographic origin (sires) and belonging to one of two grade groups: Top Grade (TG) or Runts (R). HSB were graded at two months of age as projected to grow to or exceed market size (~680 g, or 1.5 lbs) by final harvest (TG fish) or not (R fish). Grade groups were reared separately through final harvest at fifteen months of age. The “DOM” strain represents domestic SB males raised in captivity at the North Carolina State University’s Pamlico Aquaculture Field Lab (NCSU PAFL) in Aurora, NC, USA. The remaining strains were produced in hatcheries from wild-caught SB from the waters of Florida (FL), South Carolina (SC), Texas (TX), and Virginia (VA). Student’s *t*-Tests were completed to compare the average weight of the hybrids in the TG and R HSB within each strain group. The value within each bar towards the x-axis represents the number of HSB in a given group. Significance is denoted in the figure as follows: (*) indicates $p < 0.05$; (***) indicates $p < 0.001$; (****) indicates $p < 0.0001$. The mean \pm standard deviation weight (g) of the DOM TG offspring ($n=16$, 684.19 ± 108.65 g) were significantly larger than the DOM Runt HSB ($n=6$, 441.67 ± 52.68 g) ($p < 0.0001$). The same trend was observed between SC TG offspring ($n=3$, 589.33 ± 136.66 g) and SC Runt HSB ($n=8$, 406.88 ± 88.71 g) offspring ($p=0.0261$), and VA TG ($n=9$, 580.67 ± 74.28 g) and VA Runt ($n=10$, 426.00 ± 75.51 g) offspring ($p=0.0003$). No significant difference was observed between TX TG offspring ($n=4$, 557.50 ± 87.90 g) and TX Runt HSB ($n=11$, 438.45 ± 110.087 g) offspring ($p=0.0751$). A comparison could not be completed for FL offspring ($n=4$), as all HSB were in the TG group (656.25 ± 71.014 g).

Figure 2.5. A flow-chart of the machine learning (ML) analysis used to reduce data dimensionality and identify the unique, 12 amino acid-long motif-fingerprint (MF) sequences that are differentially expressed between hybrid striped bass (HSB) in groups corresponding to three comparisons: (1) Grade: HSB were graded at two months of age based upon projected growth by the time of harvest where Top Grade (TG) fish were anticipated to reach or exceed market size and Runts not; (2) Strain: HSB produced from sires of five different geographic origins, Domestic (DOM) sires were produced and reared at the North Carolina State University Pamlico Aquaculture Field Lab (NCSU PAFL, Aurora, NC), and sires from Texas (TX), South Carolina (SC), Virginia (VA), and Florida (FL) were of hatchery origin and caught in the surrounding respective waters; and (3) Growth: HSB were categorized at harvest based upon realized growth performance as Large (LG), HSB that exceeded the average weight of fish in the TG group (≥ 633.45 g), Small (SM), HSB that did not reach the average weight of the fish in the Runts group (≤ 417.95 g), and Intermediate (IN), HSB whose weight at the time of harvest fell between these two means. **(A)** The reduction of 15,000 MFs (attributes) identified among translated HSB sequencing data via Shannon's Entropy whereby MFs were assigned a weight corresponding to the amount of information it provides for decision making. The numbers under each comparison and groups (classes) thereof are the number of 15,000 MFs that were assigned an information gain value above 0.00 and therefore provide information to a given comparison. The MFs assigned information gain values above 0.00 were further processed by four ML algorithms each twice cross-validated (66.0 % split holdout method and 20-fold cross). Fewer and fewer top-ranking (i.e., highest information gain value) MFs were included in the input to identify points of model improvement and degradation with the inclusion or exclusion of MFs. **(B)** A Venn diagram of the shared and unique MFs among those ranked for each of the three comparisons. **(C)** Plots of ML model performance measured as percent correct classification of HSB into respective groups based upon included MF expression data averaged between cross-validation strategies for each of the four algorithms. Plots were used to identify points at which models were built on too many attributes (overfitting) or too few attributes (underfitting) based upon changes to performance and the occurrence of agreement between models. The shaded area indicates thresholds of optimum model performance based upon underfitting and overfitting. **(D)** A Venn diagram of the shared and unique MFs among those determined to be important for avoiding model overfitting (Top 500 ranked MFs for all three comparisons). **(E)** A Venn diagram of the shared and unique MFs among those determined to be important for avoiding model underfitting (Top 200 for Grade, 150 for Strain, and 100 for Growth).

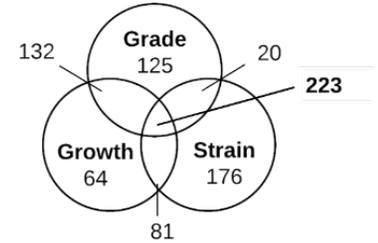
A



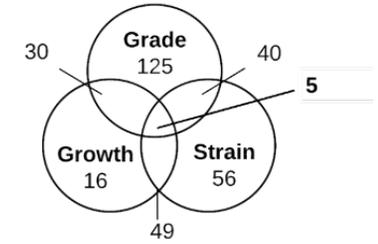
B



D



E



C

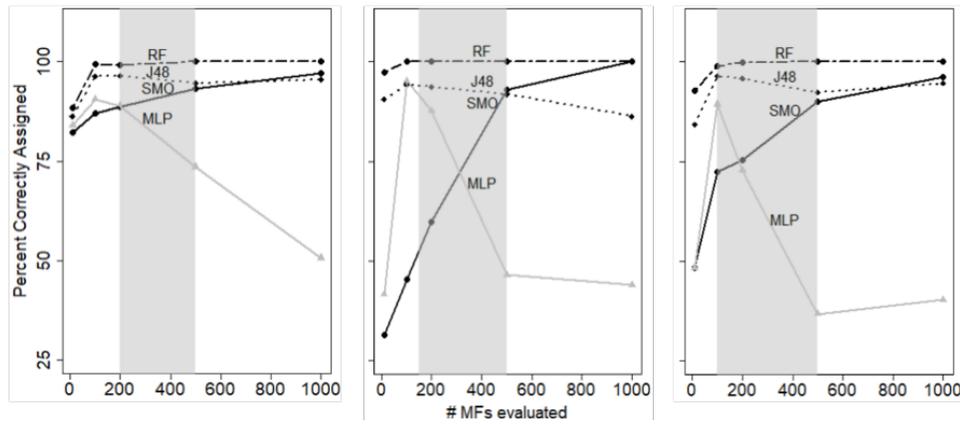


Figure 2.5.

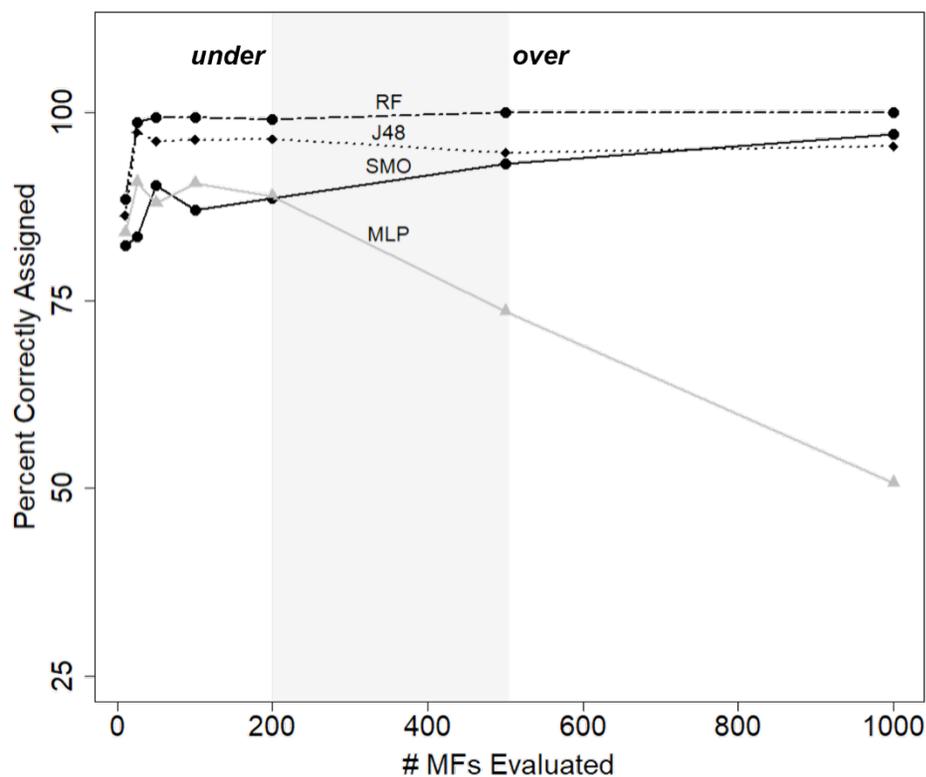


Figure 2.6. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (%) hybrid striped bass (HSB) into classes (groups) for the Grade Comparison (*see*: **Table 2.1**) based upon expression patterns of unique, 12 amino acid-long motif fingerprints (MFs; attributes) quantitated from translated RNA-Seq data generated from these fish and identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 15,000 MFs measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (20-folds) were the cross-validation strategies applied with each run of ML algorithm. Percent correct classification is plotted as an average calculated between the outcomes of each cross-validated algorithm run on subsets of the ranked attributes for this comparison. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance, “under”) and overfitting (i.e., too many highly-ranked attributes included, “over”) for this comparison, which were determined to be 200 and 500 attributes, respectively. This figure corresponds to the plot shown in **Figure 2.5(B)**.

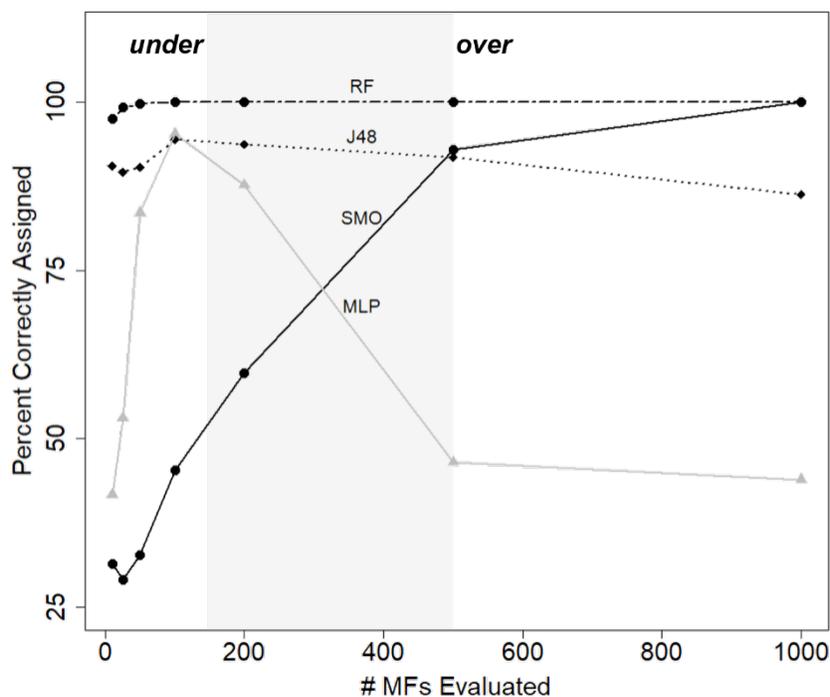


Figure 2.7. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) hybrid striped bass (HSB) into classes (groups) for the Strain Comparison (*see*: **Table 2.1**) based upon expression patterns of unique, 12 amino acid-long motif fingerprints (MFs; attributes) quantitated from translated RNA-Seq data generated from these fish and identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 15,000 MFs measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (20 folds) were the cross-validation strategies applied with each run of ML algorithm. Percent correct classification is plotted as an average calculated between the outcomes of each cross-validated algorithm run on subsets of the ranked attributes for this comparison. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance, “under”) and overfitting (i.e., too many highly-ranked attributes included, “over”) for this comparison, which were determined to be 150 and 500 attributes, respectively. This figure corresponds to the plot shown in **Figure 2.5(B)**.

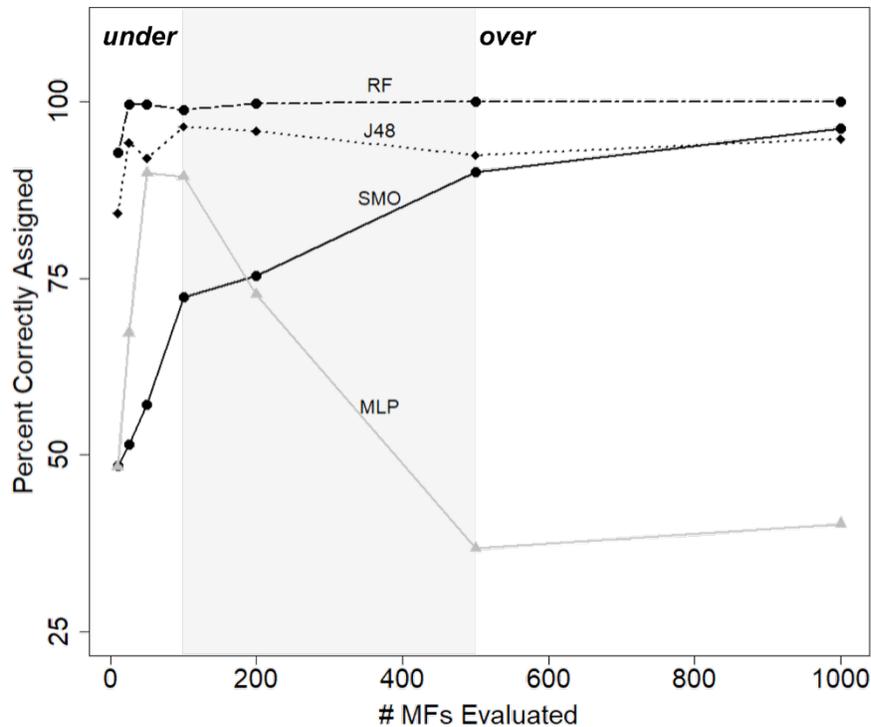


Figure 2.8. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (%) hybrid striped bass (HSB) into classes (groups) for the Growth Comparison (*see*: **Table 2.1**) based upon expression patterns of unique, 12 amino acid-long motif fingerprints (MFs; attributes) quantitated from translated RNA-Seq data generated from these fish and identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 15,000 MFs measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified *K*-fold cross-validation (20-folds) were the cross-validation strategies applied with each run of ML algorithm. Percent correct classification is plotted as an average calculated between the outcomes of each cross-validated algorithm run on subsets of the ranked attributes for this comparison. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance, “under”) and overfitting (i.e., too many highly-ranked attributes included, “over”) for this comparison, which were determined to be 100 and 500 attributes, respectively. This figure corresponds to the plot shown in **Figure 2.5(B)**.

Figure 2.9. The number of unique, twelve-amino acid long motif fingerprints (MFs) among those identified as highly informative in the differentiation of fish by growth performance (Grade and Growth comparisons) and/or paternal geographic location of origin (Strain comparison) (821 MFs total) displayed in a similar fashion to absolute abundance. The MFs mapped to thirty-three genes, the human (*Homo sapiens*) orthologue gene symbols are provided along the x-axis. These MFs were identified through application of a machine learning (ML) workflow to reduce dimensionality from 15,000 MFs to only those that yielded optimum classification performance (i.e., most individuals, or instances, correctly assigned comparison groups, or classes) by four distinct, cross-validated ML algorithms. The 500 most-informative MFs for each of the three comparisons were considered together and reduced to a total of 821 unique MFs. The MFs were mapped to the translated genome assemblies of the parental fish, striped bass (SB, paternal) and white bass (WB, maternal) and categorized as either “Both”, whereby complete homology (i.e., twelve-out-of-twelve amino acid match) was identified among the translated SB and WB sequences; “Undetm” or undetermined, whereby complete homology was not identified among either SB or WB sequences; “WB” if complete homology was exclusively identified among the WB sequence, and “SB” if complete homology was exclusively identified among the SB sequence.

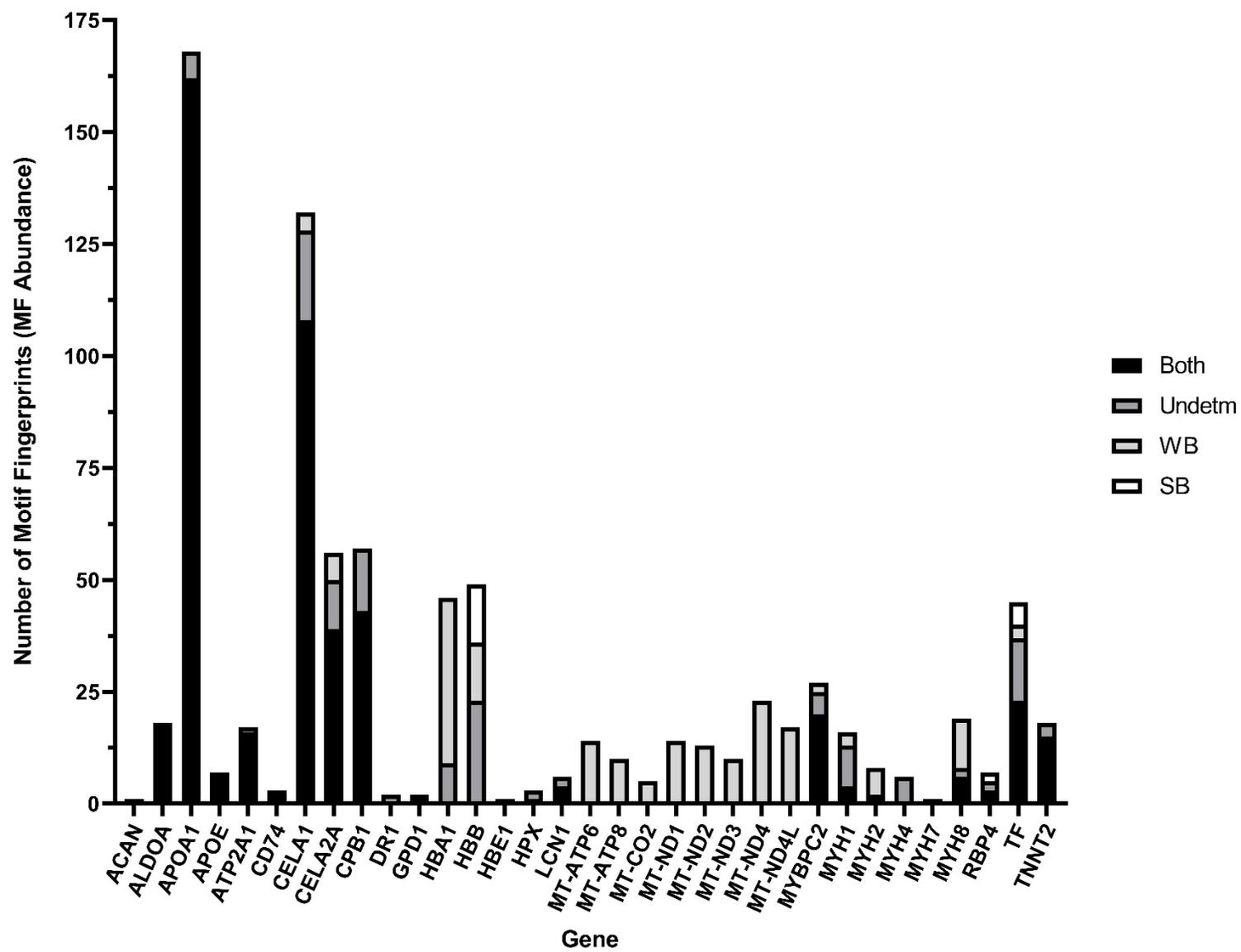


Figure 2.9.

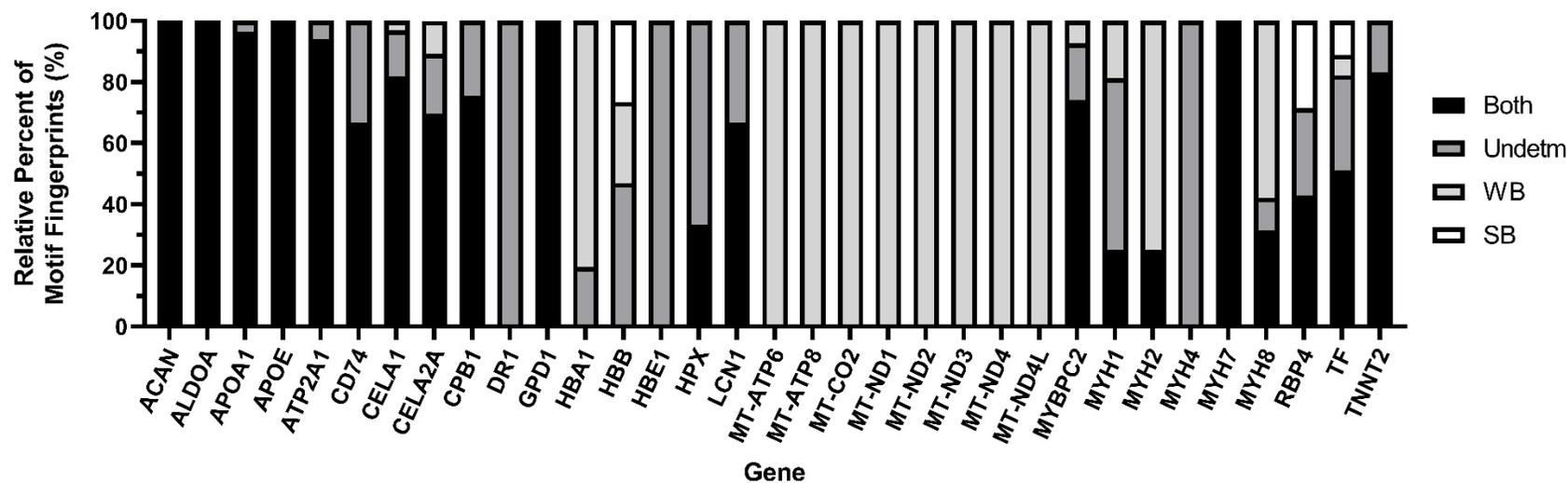
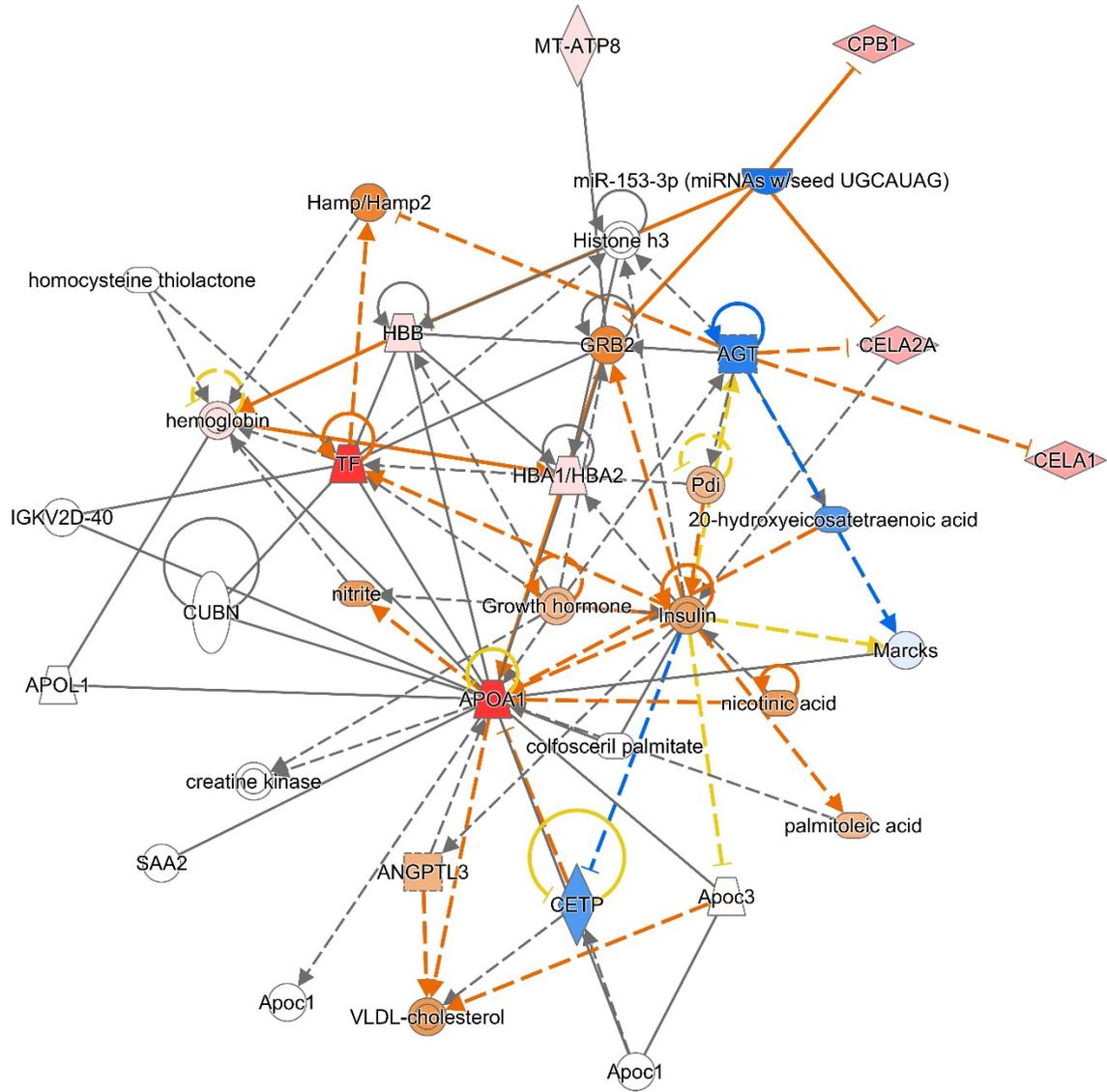


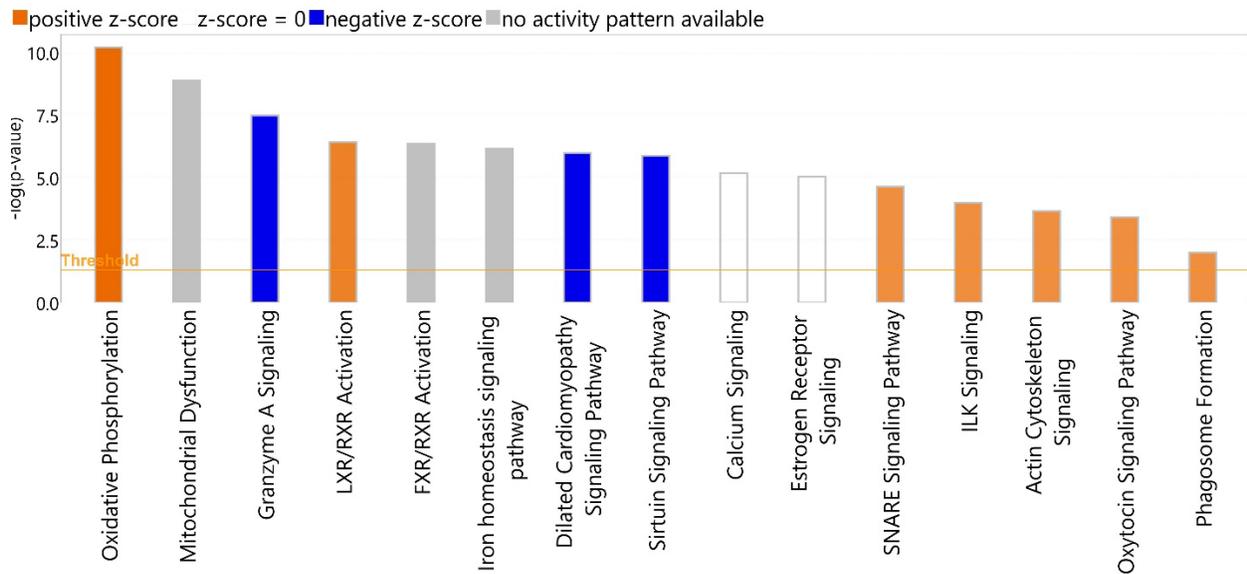
Figure 2.10. The amount of unique, twelve-amino acid long motif fingerprints (MFs) annotating back to each gene (x-axis) that were identified as having complete homology (i.e., twelve-out-of-twelve amino acid match) to “Both” the translated striped bass (SB, paternal) and white bass (WB, maternal) genome sequence assemblies; neither the SB or WB sequences and are therefore considered undetermined or “Undetm”, exclusively the WB sequence, or exclusively the SB sequence expressed as a relative percent (%) of all MFs annotating to a given gene. The MFs and subsequent genes examined here are those identified through the application of a machine learning (ML) workflow whereby 15,000 MFs were reduced in dimensionality to only those yielding the optimum classification performance of hybrid striped bass (HSB) offspring into comparison groups based on growth performance (Grade and Growth comparisons) and/or paternal geographic locations of origin (Strain comparison) by four cross-validated ML algorithms. The 500 most-informative MFs for each of the three comparisons were considered together and reduced to a total of 821 unique MFs, which were subsequently concatenated and aligned to protein sequences of the translated SB and/or WB genome assemblies. Human (*Homo sapiens*) orthologous were identified for subsequent pathway analyses and these gene symbols are provided here.

Figure 2.11. Network of upstream regulatory molecules, causal networks, and downstream effects predicted to be associated to the input genes identified among hybrid striped bass (HSB). A subsample of the HSB study population (n=40) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups based on growth performance and sire strain (i.e., geographic location of origin). These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. There were 223 MFs that were among the most important for each comparison and these MFs concatenated to eight genes, which are included in the network here: apolipoprotein A-I (*APOA1*), carboxypeptidase B1 (*CPB1*), chymotrypsin-like elastase 1 (*CELA1*), chymotrypsin-like elastase 2A (*CELA2A*), hemoglobin subunit alpha 1 (*HBA1*), hemoglobin subunit beta (*HBB*), mitochondrially encoded ATP synthase membrane subunit 8 (*MT-ATP8*), and transferrin (*TF*). Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: complexes or groups as double circles (e.g., Pdi), chemicals or drugs as horizontal ovals (e.g., nicotinic acid), cytokines as squares (e.g., AGT), transmembrane receptors as vertical ovals (e.g., CUBN), peptidases as horizontal diamonds (e.g., CELA2A), transporters as irregular polygons (*APOA1*), mature microRNAs as semi-circles (e.g., miR-153-3p), and elements classified as “other” are displayed as circles (e.g., Marcks). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Arrows curving from a network element back to itself indicate a molecule interacts with itself (e.g., autophosphorylation). Predicted activity is indicated by orange for activation, blue for inhibition, pink or red for increased measurement, yellow for findings that are inconsistent with the up- or down-regulation of a downstream molecule, and grey indicates that an effect is not predicted. Color intensity indicates that it is more extreme in the dataset.



© 2000-2022 QIAGEN. All rights reserved.

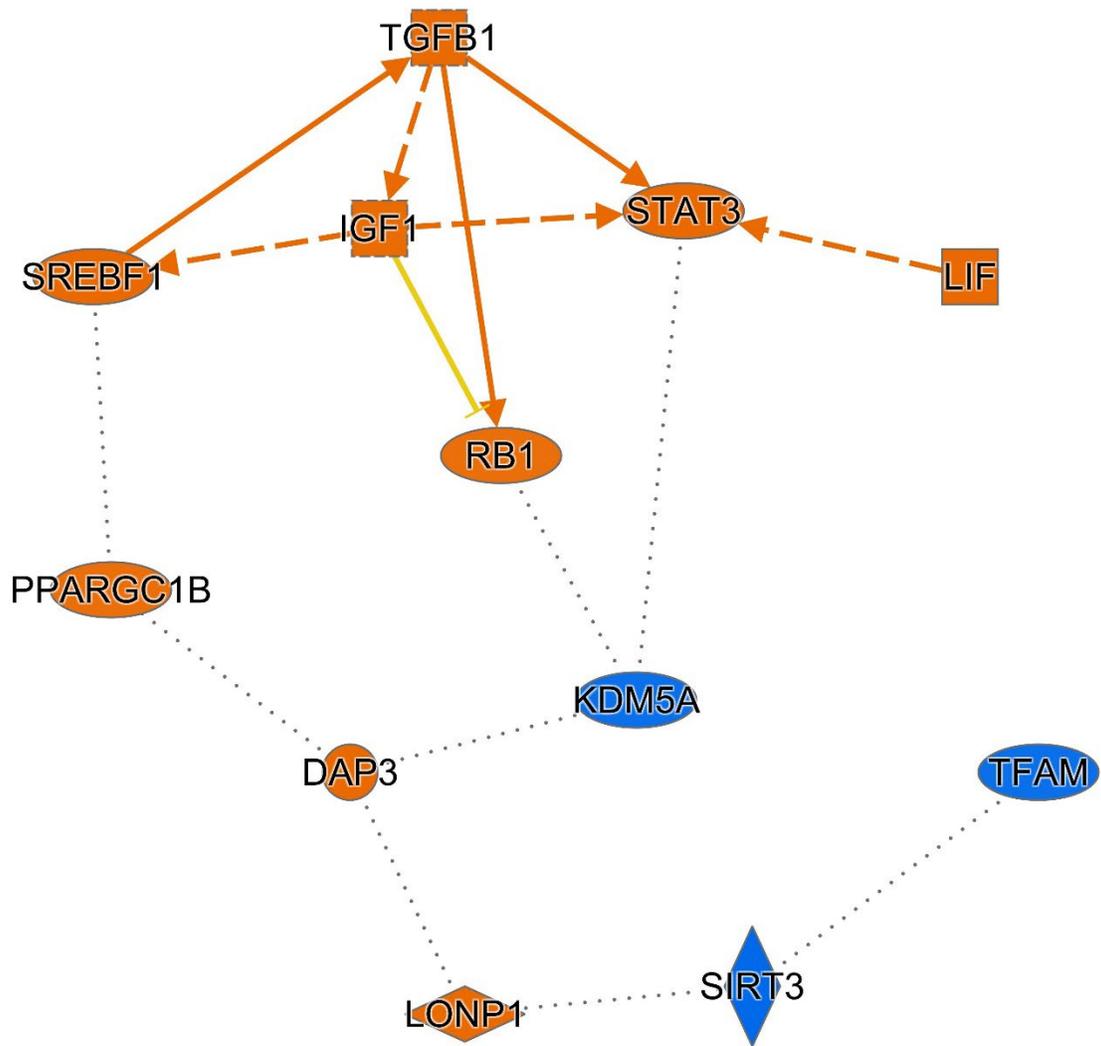
Figure 2.11.



© 2000-2022 QIAGEN. All rights reserved.

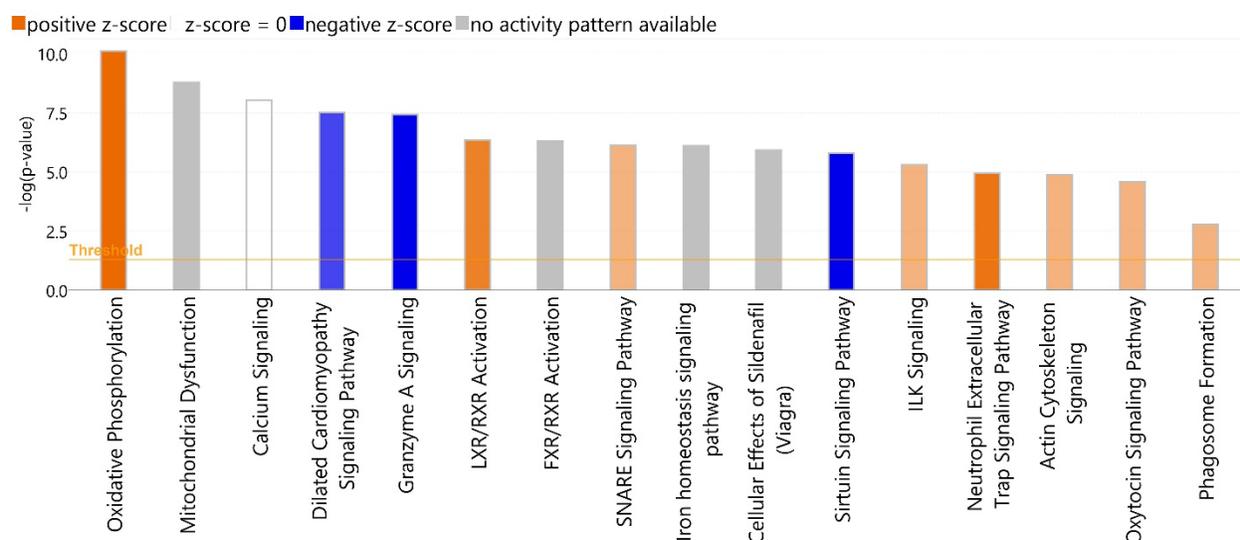
Figure 2.12. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in “Top Grade” (TG) hybrid striped bass (HSB), whereby TG is used to describe HSB that at two months of age were graded as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age. A subsample of the HSB study population (n=40, 20 HSB in TG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p=0.05$ indicated by “Threshold” line parallel to x-axis). Blue bars represent inhibited pathways, from left: Granzyme A Signaling, Dilated Cardiomyopathy Signaling Pathway, and Sirtuin Signaling Pathway. Orange bars represent activated pathways, from left: Oxidative Phosphorylation, LXR/RXR Activation, SNARE Signaling Pathway, ILK Signaling, Actin Cytoskeleton Signaling, Oxytocin Signaling Pathway, and Phagosome Formation. Grey bars indicate that activity prediction cannot be made based upon the specific pathway construction and associated molecules, from left: Mitochondrial Dysfunction, FXR/RXR Activation, and Iron homeostasis signaling pathway. White bars have a z-score of zero indicating that the evidence for activation and inhibition are equal preventing a prediction from being made: Calcium Signaling and Estrogen Receptor Signaling.

Figure 2.13. Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among “Top Grade” (TG) hybrid striped bass (HSB), whereby TG is used to describe HSB that at two months of age were graded as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age. A subsample of the HSB study population (n=40, 20 HSB in TG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., TGFB1), transcription regulators are displayed as horizontal ovals (e.g., STAT3), peptidases are displayed as horizontal diamond (e.g., LONP1), enzymes are displayed as vertical diamonds (e.g., SIRT3), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, yellow for findings that are inconsistent with the up- or down-regulation of a downstream molecule, and grey indicates that an effect is not predicted.



© 2000-2022 QIAGEN. All rights reserved.

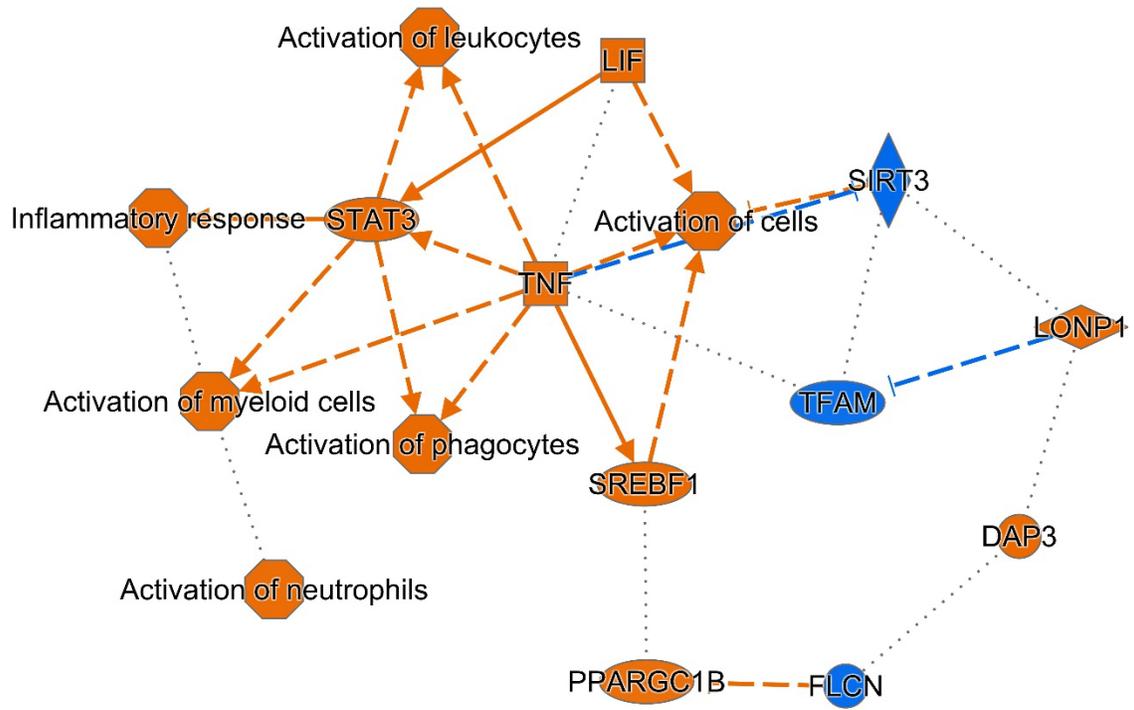
Figure 2.13.



© 2000-2022 QIAGEN. All rights reserved.

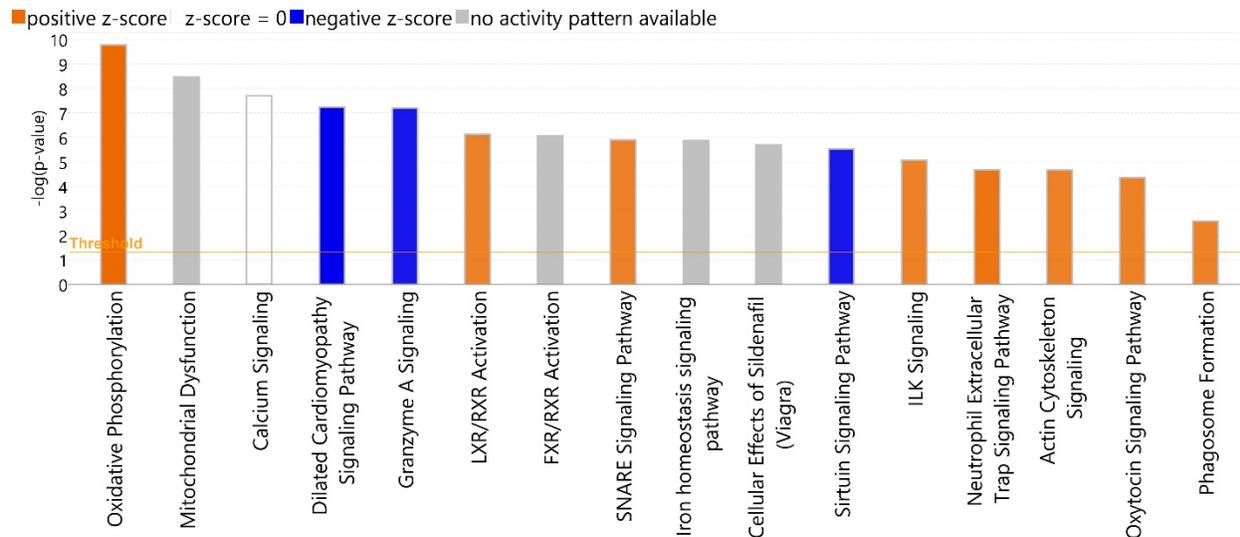
Figure 2.14. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in “Large” (LG) hybrid striped bass (HSB). HSB were graded at two months of age as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age (referred to as Top Grade, TG), or not (referred to as Runts). LG is used to describe HSB that reached or exceeded the mean weight of HSB in the TG group. A subsample of the HSB study population (n=40, 13 HSB in LG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into growth performance groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p=0.05$ indicated by “Threshold” line parallel to x-axis). Blue bars represent inhibited pathways, from left: Dilated Cardiomyopathy Signaling Pathway, Granzyme A Signaling, and Sirtuin Signaling Pathway. Orange bars represent activated pathways, from left: Oxidative Phosphorylation, LXR/RXR Activation, SNARE Signaling Pathway, ILK Signaling, Neutrophil Extracellular Trap Signaling Pathway, Actin Cytoskeleton Signaling, Oxytocin Signaling Pathway, and Phagosome Formation. Grey bars indicate that activity prediction cannot be made based upon the specific pathway construction and associated molecules, from left: Mitochondrial Dysfunction, FXR/RXR Activation, Iron homeostasis signaling pathway, and Cellular Effects of Sildenafil (Viagra). White bars have a z-score of zero indicating that the evidence for activation and inhibition are equal preventing a prediction from being made: Calcium Signaling.

Figure 2.15. Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among “Large” (LG) hybrid striped bass (HSB). HSB were graded at two months of age as being expected to reach or exceed market size (~680 g, or 1.5 lbs) by final harvest at fifteen months of age (referred to as Top Grade, TG), or not (referred to as Runts). LG is used to describe HSB that reached or exceeded the mean weight of HSB in the TG group. A subsample of the HSB study population (n=40, 13 HSB in LG group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal striped bass (SB, *M. saxatilis*) fish, and examined further via pathway analysis. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., TNF), transcription regulators are displayed as horizontal ovals (e.g., STAT3), peptidases are displayed as horizontal diamond (e.g., LONP1), enzymes are displayed as vertical diamonds (e.g., SIRT3), functions are displayed as octagons (e.g., Activation of neutrophils), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted.



© 2000-2022 QIAGEN. All rights reserved.

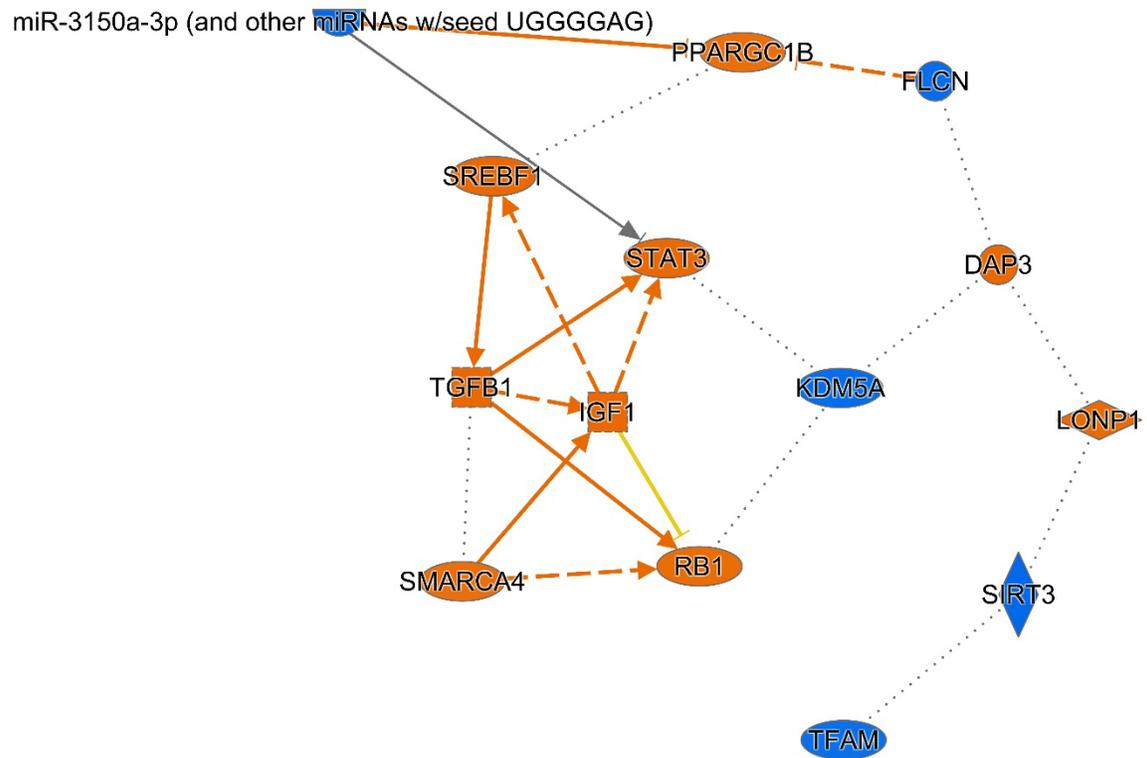
Figure 2.15.



© 2000-2022 QIAGEN. All rights reserved.

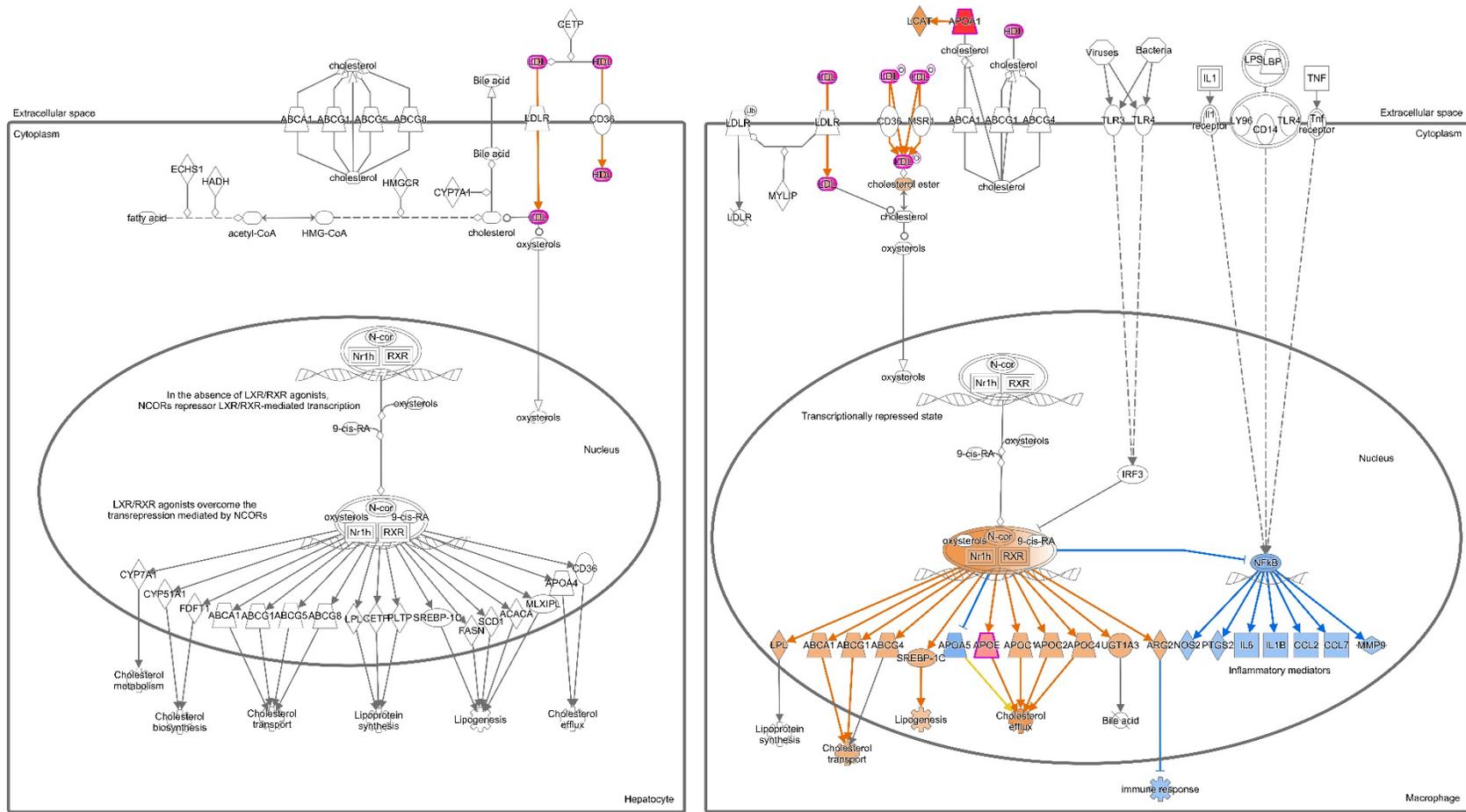
Figure 2.16. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in HSB produced from domestic (DOM) striped bass (SB) sires that were produced and reared at the North Carolina State University Pamlico Aquaculture Field Lab (NCSU PAFL, Aurora, NC). A subsample of the HSB study population (n=40, 12 HSB in DOM group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into growth performance groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal SB (*M. saxatilis*) fish, and examined further via pathway analysis. Pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher's Exact Test ($p=0.05$ indicated by "Threshold" line parallel to x-axis). Blue bars represent inhibited pathways, from left: Dilated Cardiomyopathy Signaling Pathway, Granzyme A Signaling, and Sirtuin Signaling Pathway. Orange bars represent activated pathways, from left: Oxidative Phosphorylation, LXR/RXR Activation, SNARE Signaling Pathway, ILK Signaling, Neutrophil Extracellular Trap Signaling Pathway, Actin Cytoskeleton Signaling, Oxytocin Signaling Pathway, and Phagosome Formation. Grey bars indicate that activity prediction cannot be made based upon the specific pathway construction and associated molecules, from left: Mitochondrial Dysfunction, FXR/RXR Activation, Iron homeostasis signaling pathway, and Cellular Effects of Sildenafil (Viagra). White bars have a z-score of zero indicating that the evidence for activation and inhibition are equal preventing a prediction from being made: Calcium Signaling.

Figure 2.17. Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among HSB produced from domestic (DOM) striped bass (SB) sires that were produced and reared at the North Carolina State University Pamlico Aquaculture Field Lab (NCSU PAFL, Aurora, NC). A subsample of the HSB study population (n=40, 12 HSB in DOM group) were sacrificed for sequencing and motif fingerprint (MF) analysis whereby unique, twelve amino acid long sequences found to highly vary between the translated sequence data generated from white muscle tissue of these HSB were further reduced through application of a machine learning (ML) workflow to determine those most important to the correct classification of HSB into grade groups. These MFs were then concatenated, if possible, mapped to the translated reference genome sequence assemblies of the maternal white bass (WB, *Morone chrysops*) and paternal SB (*M. saxatilis*) fish, and examined further via pathway analysis. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., TGFβ1), transcription regulators are displayed as horizontal ovals (e.g., STAT3), peptidases are displayed as horizontal diamond (e.g., LONP1), enzymes are displayed as vertical diamonds (e.g., SIRT3), mature microRNAs are displayed as half circles (e.g., miR-3150a-3p), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, yellow for findings that are inconsistent with the up- or down-regulation of a downstream molecule, and grey indicates that an effect is not predicted.



© 2000-2022 QIAGEN. All rights reserved.

Figure 2.17.

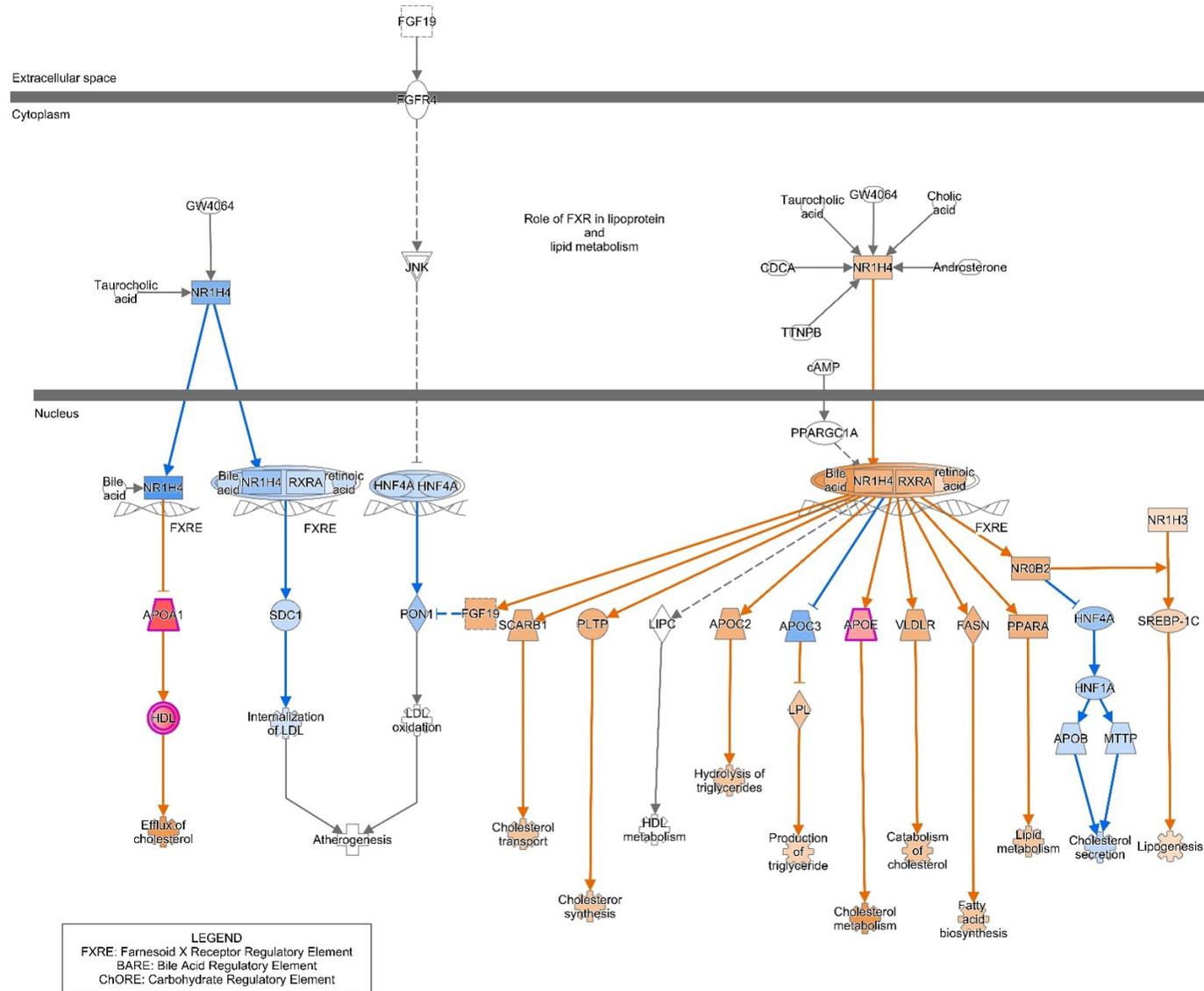


© 2003 2002 QIAGEN. All rights reserved.

Figure 2.18. The LXR/RXR Activation pathway (liver X receptor/retinoid X receptor) enriched from genes identified among hybrid striped bass (HSB) exhibiting superior growth and/or of domestic sire strain (i.e., male striped bass were not of wild-origin). Specifically, apolipoprotein a-I (*APOA1*) and apolipoprotein E (*APOE*) are highlighted in red and are up-regulated in the dataset. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of other interacting elements indicated as being activated if in orange or inhibited in blue.

Figure 2.19. The FXR/RXR Activation pathway (farnesoid X receptor/retinoid X receptor) enriched from genes identified among hybrid striped bass (HSB) exhibiting superior growth and/or of domestic sire strain (i.e., male striped bass were not from wild-origin). Specifically, apolipoprotein a-I (*APOA1*) apolipoprotein E (*APOE*), shown in red to indicate their up-regulation in the dataset. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of other interacting elements indicated as being activated if in orange or inhibited in blue.

FXR/RXR Activation



© 2000-2022 QIAGEN. All rights reserved.

Figure 2.19.

CHAPTER 3. TRANSCRIPTOMICS ANALYSIS OF STRIPED BASS SKELETAL MUSCLE REVEALS DISTINCT METABOLIC DIFFERENCES BETWEEN FISH EXHIBITING INFERIOR AND SUPERIOR GROWTH

Abstract

Distinct metabolic response pathways of striped bass (SB, *Morone saxatilis*) exhibiting superior and inferior growth traits were identified via machine learning (ML) analysis of whole transcriptomes generated from white, fast-twitch skeletal muscle tissue. Half-sibling families of fifth generation domestic striped bass were produced by crossing two female ('dam') SB each with six male ('sire') SB each. The six sires crossed with each female had been categorized as "Large" (by weight and length) or "Small", such that there were three in each group to enable a comparison of gene expression profiles of SB exhibiting superior growth traits as well as those bred from sires of a distinct growth phenotype. The ML pipeline identified between 35–300 unique gene transcripts as determinant of growth performance, dam, sire, or sire size and a pathway analysis of these genes revealed distinct differences in genes underlying critical metabolic pathways. Specifically, genes up-regulated in SB exhibiting inferior growth suggest that protein ubiquitination and skeletal muscle degeneration is active in these fish, rather than critical signaling pathways that regulate genes involved in growth (HIF-1 α Signaling, JAK/STAT) and muscle homeostasis. These findings are suggestive of a metabolic dysfunction in these inferior growth fish that can potentially be targeted via selective breeding or biotechnological interventions.

Introduction

A domesticated line of striped bass (SB, *Morone saxatilis*) has been bred in captivity as part of the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* for several generations (Garber & Sullivan, 2006; Hodson et al., 1999; Reading, McGinty, et al., 2018; Woods III, 2001). It is estimated that as much as 90.0 % of the sunshine hybrid striped bass (HSB, male SB) raised in the U.S. from 2015–2019 were produced using domesticated male SB broodstock from this program, which is housed at the North Carolina State University Pamlico Aquaculture Field Laboratory (NCSU PAFL, Aurora, NC, USA). The breeding of these fish in a culture environment has led to their acclimation to aquaculture conditions such that their offspring demonstrate greater hardiness than those produced from wild counterparts in the same culture conditions. For example, at present the breeding program provides SB that have been selectively bred without the exogenous hormone compounds previously thought of as necessary for spawning these fish in captivity (Andersen et al., 2021). Despite great advancements in the domesticated SB broodstock, it is understood that the integration of wild-origin striped bass stocks may be critical for offshore culture in some regions due to escapement concerns. Nevertheless, the establishment and continued maintenance of a domesticated broodstock that shows genetic gains over each filial generation is necessary for the success of the aquaculture industry as a whole.

In addition to desirable gains in culture traits seen as a result of selective breeding, the SB is a priority species for the United States Department of Agriculture (USDA) National Animal Genome Research Support Program (NRSP-8) and a considerable number of genomic resources have been developed for these fish. The SB genome assembly (NCSU_SB_2.0, GenBank® accession: GCA_004916995.1) was recently updated (2019) through a combinatorial approach

of short-read sequencing (Illumina, San Diego, CA), long-read sequencing (Pacific Biosciences, Menlo Park, CA), and Chicago® and Dovetail™ Hi-C + HiRise™ scaffolding (Dovetail Genomics, Scotts Valley, CA). Additionally, other available genomic resources for SB include a medium-density genetic linkage map of DNA markers (Liu et al., 2012); a multi-tissue transcriptome (Li et al., 2014; GenBank accession GBAA000000000); an ovary transcriptome representative of all oocyte growth and maturation stages (Reading et al., 2012; GenBank accession SRX007394); several microsatellite DNA marker panels (Brown et al., 2005; Couch et al., 2006; Han et al., 2000; Rexroad et al., 2006; Skalski et al., 2006); and an epigenetic profile of sperm methylation as it correlates to fertility (Woods III et al., 2018).

Despite this great amount of available omics data available for SB, breeding strategies for SB are currently limited overall in terms of ready-to-use molecular tools that can be integrated into the breeding program. Herein describes the results of a study of SB reared until eighteen months of age with the final twelve months of age spent in an indoor recirculating aquaculture system. The results of a machine learning (ML) workflow are presented whereby the genes most important to the correct classification of SB into groups based upon growth performance or parentage group were identified and incorporated into pathway analysis. The pathway analysis of these fish revealed distinct immune responses underlying metabolic processes, suggestive of protein ubiquitination and skeletal muscle degeneration being activated in SB exhibiting inferior growth traits. The findings here represent targets for high-throughput phenotyping, selective breeding, and/or biotechnological targeting to support the development of these domesticated broodfish and the SB aquaculture industry.

Materials and Methods

All research was performed under the protocol approved by the Institutional Animal Care and Use Committee of North Carolina State University (Protocol number 10-042-A).

Furthermore, this study was carried out in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health (National Research Council, 1996).

Experimental Animals and Tissue Collection

Fifth (5th) generation domestic SB females (4 years old; N=2: Dam 'A' was 5.68 kg in weight and 722.00 mm in length, Dam 'B' was 6.15 kg in weight and 703.00 mm in length) and males (N=12, 3 years old, mean \pm standard deviation, SD, weight was 2.26 + 0.55 kg) from the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* were raised and tank spawned at the North Carolina State University Pamlico Aquaculture Field Laboratory (PAFL, Aurora, NC, USA) to create the SB offspring (6th generation) used for the research described herein. Beginning in the fall (October) prior to the spawning season, females were fed a commercial broodstock diet (BioBrood from Bio-Oregon, Westbrook, ME, USA). In the late winter (February), broodstock were moved from outdoor tanks (38,433 L) with flow-through well water to indoor tanks (31,139 L) on recirculating aquaculture systems maintained at 10.0–12.0 °C, 10.0–12.0 ppt salinity, and simulated ambient photoperiod (provided by overhead fluorescent lighting). Beginning in late April, the fish were gradually warmed to spawning temperature (18.0–20.0 °C) and salinity was reduced to <1.0 ppt (approximately 1.0 °C increase and 1.0 ppt salinity decrease per day) (Hodson and Sullivan, 1993). Prior to the spawning trials, males were anesthetized (MS-222, Sigma-Aldrich, St. Louis,

MO, USA), verified for spermiation by applying gentle pressure to their ventral surfaces, and treated with 165 IU human chorionic gonadotropin (hCG)/kg body weight (Chorulon®, Merck Animal Health, Kenilworth, NJ, USA) via injection into the dorsal lymphatic sinus. Male sires were divided into two equal groups of “Large” (mean \pm SD weight was 2.76 + 0.15 kg, TL was 570.67 \pm 7.84 mm) and “Small” (mean \pm SD weight was 1.75 + 0.15 kg, TL was 490.00 \pm 21.14 mm) that significantly differed in weight and TL ($p < 0.0001$, Student’s *t*-Test, JMP® Pro v.14.0.0, SAS Institute, Cary, NC, USA). The procedure of verifying spermiation was repeated at the conclusion of each spawning trial to confirm fertile male participation. Females were anesthetized and oocyte samples were collected by ovarian biopsy as described in Rees and Harrell (1990). The oocytes were examined with a dissecting microscope to determine the eligibility of the female fish to participate in spawning (i.e., fully vitellogenic with ovary diameter $> 1000 \mu\text{m}$, stage < 15 hours (h), and non-atretic oocytes based upon the staging criteria of Bayless (1972) and methods described in Sullivan et al. (2003). Eligible females were treated with 330 IU hCG/kg prior to tank spawning. Caudal fin clips from all broodfish were collected and stored in 95.0 % ethanol for microsatellite genotyping.

Females were put into separate individual indoor spawning tanks (2,403 L) equipped with external egg collectors and supplied with flow-through well water (< 0.5 ppt) (Smith and Whitehurst, 1990). Three sires from each of the two size groups (Large and Small) were placed into the two spawning tanks with the females. The fish were left undisturbed for up to six days or until spawns were observed and egg collectors were checked at approximately 2 h intervals for the presence of fertilized and unfertilized eggs. Upon conclusion of the trials, female spawning was confirmed by applying gentle pressure to ventral surfaces and/or performing ovarian biopsies if eggs were still present in the ovary. Eggs from each spawn were harvested from the

collectors to determine fertilization success. Fertilization was calculated using triplicate random sub-samples of eggs at 4 to 6 h post-fertilization (1.0 mL each; Rees and Harrell, 1990).

Spawning time was verified by the degree of embryo development, using the criteria of Rees and Harrell (1990). Eggs from each spawn were incubated separately in McDonald hatching jars according to standard procedures (Mullis and Smith, 1990) and larvae (fry) were collected into separate aquaria upon hatching. The number of fry produced was enumerated by taking triplicate samples (40.0–60.0 mL per sample) from each aquarium and counting the amount of fry in each sample, then multiplying the average count by the known volume of each aquarium. Hatchery water ranged from 18.0–2.0 °C and 0.0 ppt salinity during spawning, egg incubation, and fry incubation. Fertilized eggs were incubated in McDonald hatch jars and maintained in aquaria for four days post hatch until being pooled (~20,000 fry from each group of fry offspring) and stocked across two 0.1 hectare (ha) ponds (Mullis and Smith, 1990).

At four months of age, a subpopulation of the offspring produced (n=194) were brought to Grinnell's Animal Health Laboratories building on the NCSU main campus (Raleigh, North Carolina, USA) and split randomly into three, indoor recirculating aquaculture system tanks (908 L). Fish were fed a measured amount of 3.0 MM, 1/8" Zeigler Silver Floating feed (40.0 % protein, 10.0 % lipid; Zeigler Bros., Inc., Gardners, PA, USA) once per day *ad libitum* (i.e., to satiation) five consecutive days each week (generally Monday through Friday). At five months of age all individuals were partially sedated with AQUI-S® 20E (eugenol; Merck Animal Health), measured (weight and total length, TL), and sampled for caudal fin clips. Fin clips were stored in 95.0 % ethanol (EtOH) at 4.0 °C. This sampling procedure was repeated in a similar fashion every three months thereafter for twelve months (i.e., until eighteen months of age). At eight months of age the entire study population was moved into larger tanks (2,006 L) and

transitioned to a larger feed (Zeigler Silver Floating feed, 5.0 MM, 3/16", 40.0 % protein, 10.0 % lipid) to accommodate growth. The *ad libitum* feeding regime and recording of the total amount fed per tank continued through the duration of the study. Tank-groups were maintained throughout the study and effort was made to rotate the groups into tanks differing from that of origin at each sampling event to reduce potential tank effects based upon light exposure and similar potential factors. Twenty-one SB were lost throughout the study (89.18 % survival); all lost fish were found to have jumped out of the tank (i.e., death was not due to disease) and biomass loss was recorded on the date of first notice. The one-way ANOVA and Tukey's Honestly Significant Difference (HSD) post-hoc test, if appropriate, were used to identify any differences in growth (size) between tanks at each sampling point (alpha=0.05; JMP® Pro v.14.0.0). The feed conversion ratio (FCR) for each tank group was calculated after sampling events at eight, eleven, fourteen, and eighteen months and overall (i.e., for the duration of the study from five to eighteen months) as:

$$\text{FCR} = \frac{\text{sum feed offered (g)}}{\Delta \text{ biomass (g)}}$$

Where Δ (Delta) biomass is calculated by subtracting the sum weight of a tank calculated at the previous sampling from the current sum weight of the tank. For the overall FCR calculation, Δ biomass was calculated by subtracting the initial biomass (sum at 5 months) from the final biomass (sum at eighteen months).

At the final sampling event, twenty-four fish from each tank (n=72 of 173) were randomly selected to be sacrificed via spring-loaded captive bolt (MS Schippers, The Netherlands) to collect muscle tissue samples for histological and transcriptomic analyses. The number of sacrificed fish was determined to increase the probability of sampling a representative

subpopulation of the twelve half-sibling families of fish created and reared for this study. Duplicate muscle tissue samples from sacrificed individuals were fixed in 10.0 % neutral buffered formalin (NBF, Sigma-Aldrich) for histology or submerged in RNAlater™ (Life Technologies, CA, USA) for 24 h prior to removing the solution, storing at -80.0 °C for transcriptomics analysis.

All sacrificed fish were designated as belonging to one of two groups based upon growth performance: Superior and Inferior, defined as follows: the Superior group are those that are the largest thirty-six fish by weight and the Inferior group are those that are the smallest thirty-six by weight. SB were further divided within these groups into the fish that reached or exceeded market size (1.36 kg, or 3.0 lbs) by the time of sampling (n=12 from the Superior group, “Market Size”, mean \pm SD weight was 1.52 ± 0.13 kg) and the fish that exhibited the poorest growth (n=12 from the Inferior group, “Under Size”, (mean \pm SD weight was 0.72 ± 0.060 kg). Additional comparisons between the Market Size, Under Size, and remaining intermediate size SB from both the Superior and Inferior groups (referred to as Superior Other and Inferior Other, respectively) enabled deeper analysis of the sub-groups representing the extremes of the sample population distribution and differences thereof. Statistical comparisons of morphometric differences (weight, TL) between groups of SB offspring based on size or parentage (Dams, Sires, and Sire size group) were conducted using Student’s *t*-Test, one-way ANOVA, Tukey’s Honestly Significant Difference (HSD) post-hoc test, and/or simple linear regression as appropriate (alpha=0.05; JMP® Pro v.14.0.0). A brief description of the groupings for each comparison is provided in **Table 3.1**.

Microsatellite Genotyping and Parentage Assignment

Fin clips from broodfish (N=14, 2 dams and 12 sires) and sacrificed offspring (n=72) were processed at the Center for Aquaculture Technologies (CAT, San Diego, CA, USA) and genotyped using two multiplexed microsatellite genotyping panels of eleven established markers: MSM 1144, MSM 1095, MSM 1096, MSM 1067, MSM 1094, MSM 1168, MSM 1208 and MSM 1243 of Couch et al. (2006) and MSM 1526, MSM 1592 and MSM 1357 of Rexroad et al. (2006). The CAT-established protocols are briefly as follows: a series of multiplex PCR experiments were first performed (using varied conditions) on a test panel of gDNAs to maximize the number of scorable loci for the two multiplexes. Once optimized panels were developed, all samples were genotyped using established protocols. Two multiplex PCR reactions were performed on each gDNA sample using Type-It Master Mix (Qiagen Inc., Hilden, Germany), in conjunction with defined primer concentrations and thermal cycling conditions. Resulting PCR products for each multiplex pool were subjected to capillary electrophoresis on an Applied BioSystems 3730xl Electropherograms (Thermo Fisher Scientific Inc., Waltham, MA, USA) were analyzed using GeneMapper™ (Thermo Fisher); and allele calls for up to eleven loci were quality checked and compiled for each sample. Samples were re-extracted, genotyped and re-called if initial results were deemed of poor quality (<50.0 %).

Parentage was assigned using CERVUS software (v.3.0.7, Field Genetics, Marshall et al., 1998). CERVUS calculates the log-likelihood of each candidate parent being the true parent by calculating the allele frequency of each locus in the population and uses simulations to determine the level of confidence in parentage assignment. An error rate of 1.0 % was assumed and the proportion of parents sampled was 100.0 %. Parentage assignment allowed for the distinction of offspring based upon dam (Dam A, Dam B), individual sire (Sires 1–12), and sire size group

(Large, Small, where “Large” are offspring produced from Sires 1–3 crossed with Dam A and Sires 7–9 crossed with Dam B and “Small” are offspring produced from Sires 4–6 crossed with Dam A and Sires 10–12 crossed with Dam B) for comparisons of morphometric measurements and gene expression.

Muscle Histology

Muscle samples were prepared for histological analysis at the NCSU College of Veterinary Medicine (CVM, Raleigh, NC, USA). Briefly, each muscle tissue sample (n=72) was washed in ethanol, embedded tissue in paraffin, sectioned each of the tissue samples into 5 μm cross sections perpendicular to the muscle fiber, stained the tissue with hematoxylin and eosin stain (H&E), and mounted each cross section onto microscope slides for muscle morphometry analysis. Identifying slide labels corresponding to samples were concealed until after enumerations had been made to eliminate potential bias in image processing and analysis.

Three unique images of the muscle fibers on each slide were taken using an Olympus CH microscope (4X) with a Celestron digital microscope imager camera (10X). ImageJ software (Fiji, v.1.52a, National Institutes of Health, NIH, Bethesda, USA) was used to count and determine the area (μm^2) of all fibers entirely within the field of view (i.e. not marginal fibers or those with a limited area due to the margin of view) of each image to evaluate hyperplastic (fiber amount) and hypertrophic (fiber area) muscle growth. The diameter of each scored fiber was calculated as a geometric derivative of its area to compute the average fiber diameter and the frequency of scored fibers of a certain diameter range for each slide image. Specifically, the frequency (i.e., count) of fibers scored on each slide image that fell into <10 μm , 10–20 μm , 20–25 μm , 25–50 μm , 50–100 μm , 100–150 μm , 150–200 μm , and >200 μm diameter bins were

recorded. Image processing and subsequent calculations were completed in duplicate (i.e., by two scorers) and the total number of fibers, average fiber diameters, and number of fibers per diameter bin were pooled prior to analysis to account for any scorer-error. The ROUT outlier test was used to identify and subsequently remove outliers (Motulsky and Brown, 2006). The total number of fibers, average fiber diameter, and number of fibers <20 μm were compared between groups of growth groups of SB offspring (Superior v. Inferior groups) using Student's *t*-Tests ($\alpha=0.05$; JMP® Pro v.14.0.0). A simple linear regression was used to assess the relationship between the average total fiber number and average fiber diameter to the weight and TL of offspring (JMP® Pro v.14.0.0).

RNA-Sequencing and Quantitative Analysis of Gene Expression

Extraction, library preparation, and next-generation sequencing of RNA from muscle tissue samples ($n=72$) was performed by the NCSU Genomic Sciences Laboratory (GSL, Raleigh, NC, USA). Briefly, total RNA extraction was performed using the RNeasy Fibrous Tissue Mini Kit (Qiagen Inc.). Sample quality and concentration was evaluated using an Agilent Bioanalyzer 2100 with an RNA 6000 Nano chip (Agilent, Santa Clara, CA, USA). The NEBNext Poly(A) mRNA Magnetic Isolation Module oligo-dT beads (New England Biolabs, NEB, Ipswich, MA, USA) were used for messenger (mRNA) purification. Complementary DNA (cDNA) libraries were constructed using the NEBNext Ultra Directional RNA Library Prep Kit and NEBNext Multiplex Oligos for Illumina sequencing (Illumina Inc., San Diego, CA, USA). Per the manufacturer-specified protocol, mRNA was chemically fragmented and primed with random oligos for first strand cDNA synthesis, dUTPs were incorporated during second-strand cDNA synthesis to accomplish strand-specificity, and cDNA was purified, end repaired, and “a-

tailed” for adaptor ligation. Samples were then selected for a final library size of 400–550 base pair (bp) using sequential AMPure XP bead isolation (Beckman Coulter, Brea, CA, USA). Library enrichment was performed and specific indexes for each sample were added during the protocol-specified PCR amplification. Amplified library fragments were purified and an Agilent 2100 Bioanalyzer with a High Sensitivity DNA chip was used to verify quality and final concentration. The final libraries were pooled in equimolar amounts and sequenced on the Illumina NovaSeq 6000 platform to yield 150 bp paired-end reads.

CLC Genomics Workbench (v.21.0.3, Qiagen Inc.) was used to complete sequencing quality control (QC), read trimming, and gene expression quantitation. A quality score limit of 0.05, maximum of two ambiguous nucleotides, and discard of reads less than fifteen bp in length were set for QC and read trimming. The CLC Genomics Workbench RNA-Seq Analysis tool is based on methods described in Mortazavi et al. (2008) and was used to extract genes (annotated as *gene*) and transcripts (annotated as *mRNA*) from the reference genome published through NIH National Center for Biotechnology Information (NCBI) GenBank® (NCSU_SB_2.0, accession: GCA_004916995.1) and calculate expression values (RPKM, reads per kilobase of exon model per million mapped reads) of each across all individual transcriptomes generated (n=72).

The CLC Genomics Workbench Differential Expression for RNA-Seq tool was used to generate fold change values and perform statistical analysis ($\alpha=0.05$) of gene transcript expression between groups of each comparison. The \log_2 fold change values were incorporated into the subsequent biological pathway analysis. The outcomes of the statistical comparisons, in particular the false discovery rate (FDR) p-value, were considered further alongside those of the machine learning (ML) analysis described below to provide greater insight into the scope (and limitations) of both approaches. Briefly, the Differential Expression for RNA-Seq tool models

each gene transcript with a separate Generalized Linear Model (GLM) to fit curves to expression values and with the assumption that read counts follow a Negative Binomial Distribution. This Negative Binomial Distribution can be considered a mixture of Poisson distributions with a Gamma-distributed Poisson parameter (λ). The Gamma distribution is a function of the dispersion parameter, which is positively correlated with variation in expression such that when dispersion is high variance is high and when dispersion is zero the Negative Binomial Distribution is reduced to a Poisson distribution. The tool employs the Wald test (“All group pairs”) for comparisons of two groups (Growth, Dam, and Sire Size) and the Likelihood ratio test (“Across groups (ANOVA-like)”) for comparisons of more than two groups (Market Size and Sire).

Machine Learning Analysis

The complete dataset of all quantitated gene transcripts was reduced in dimensionality (i.e., number of values) through the application of a supervised ML workflow applied using Weka software (v.3.8, University of Waikato, New Zealand). This workflow was applied separately for each of the five comparisons (Growth, Market Size, Dam, Sire, Sire Size) to identify which gene transcripts (attributes), and subsequently genes, are the most informative to the classification of SB individuals (instances) into the user-defined groups (classes). Four ML algorithms were used for model building and subsequent determination of informative attributes: a support vector machine (SVM; sequential minimal optimization, SMO), an artificial neural network (ANN; multilayer perceptron, MLP), a decision tree (J48), and an ensemble (i.e., combination of models) decision tree (Random Forest). Two cross-validation strategies, the holdout method and the stratified K -fold cross-validation, were applied with each algorithm to

designate the training and test subsets of data necessary for model building and evaluation. The 66.0 % split was used as the holdout method, whereby 66.0 % of data is randomly designated as the training set from which the algorithm learns and the other 34.0 % of data is designated as the test set used to validate the model and measure performance. In the *K*-fold cross-validation, the data was partitioned into subsets of data (“folds”) based on the minimum number of samples in a class for a given comparison. The *K*-fold cross-validation then runs several successive models where each run is the model training on the entire dataset minus a random fold and the random fold that is excluded changes with each run. Three metrics were used to evaluate model performance: (1) percent correct classification, or how well did the model classify each sample to its true grouping based upon the values associated to the provided attributes; (2) Area Under the Receiver Operating Characteristic Curve (AUROC, or ROC Curve), which is the power of the analysis as a function of the Type I Error, or the plot of the true positive rate (TPR) against the false positive rate (FPR); and (3) Kappa Statistic (Cohen’s Kappa Coefficient), a measure of randomness, or the possibility of the model correctness as a function of chance.

Each cross-validated algorithm was first applied to the complete (i.e., all transcripts) and labeled (i.e., identified as belonging to a given class) dataset to establish baseline model performance. Information gain (“Shannon’s Entropy”) was then calculated for each attribute. Information gain is a numerical value representative of the amount of information gained from the inclusion of a given attribute in learning; the lower the calculated entropy value the more information gained from considering a given attribute and associated data. Each attribute was then assigned a rank based on information gain, whereby the top-ranked attribute (i.e., “number 1”) was that with the lowest calculated entropy value. Any attributes with an information gain

value of 0.0 (no information) were removed from subsequent analyses, thus representing the first reduction of data dimensionality.

Data dimensionality was further reduced using recursive elimination by identifying the minimum number of highest-ranking gene transcripts required to maintain optimal classification determined for the four algorithms. Specifically, subsets of the ranked attributes (i.e., information gain weight above 0.0) were created such that each subset included fewer and fewer top-ranked attributes to allow for the evaluation of changes in model performance (model fit). The iterative recursive exclusion of ranked attributes allows for the identification of overfitting and underfitting, or when too many or too few attributes, respectively, are included in the dataset and may negatively impact model performance. For the Growth, Market Size, Dam, and Sire Size comparisons, the subset of ranked attributes was first reduced to the nearest fifty or hundred (the top 900, 250, 250, and 350 attributes, respectively), then by fifty attributes at a time until only the top fifty remained. Subsets were then reduced to the top twenty-five, fifteen, and ten ranked attributes until only the number one attribute remained; the Sire comparison subsets were reduced by five attributes each, as only thirty-five had an information gain value above zero. Model performance (percent correct classification) for each cross-validation algorithm was plotted against the number of attributes included in each iterative run.

The number of gene transcripts required to avoid model underfitting (i.e., too few gene transcripts included in analysis leading to diminished model performance) was determined as the point at which the four models had the greatest agreement of percent correct classification performance with the inclusion of the fewest high-ranking attributes. The number of gene transcripts required to avoid model overfitting (i.e., too many gene transcripts included in analysis leading to diminished model performance) was determined by identifying the point at

which the performance of at least one model began to decline with the inclusion of a greater number of high-ranking attributes. The number of top-ranked gene transcripts identified as those necessary to avoid overfitting for each comparison are referred to as “optimal” for the purpose of subsequent pathway analyses described here (**Table 3.2**).

A negative control for learning was established by running the cross-validated algorithms on randomized versions of the optimal dataset such that the class labels (e.g., Superior or Inferior) are not known to be associated with data truly representative of a given class (Schilling et al., 2014, 2015). This randomization process was repeated ten times for each comparison and the average values of performance metrics between cross-validated algorithms were calculated. Mean values were approximately what can be predicted from random assignment based on the number of classes for the Growth, Market Size, and Sire Size comparisons, thus it was concluded that true learning occurred when using correctly labeled data to build models (**Table 3.2**). The mean percent correct classification of randomized, negative control datasets were greater than those predicted by random chance for the Dam and Sire comparisons (**Table 3.2**), presumed to be in part due to skewed number of offspring born to a single dam or sire over others (i.e., class imbalance) (**Table 3.1**), parental effects of gene expression, and/or a combination thereof that may lead to bias in model training.

Pathway Analysis

Qiagen Ingenuity Pathway Analysis (IPA; Qiagen Inc.) software was used to identify the most enriched (based upon number of associated molecules and calculated p-value) pathways (“Canonical Pathways”), predict what molecules and cellular processes may be underlying observed patterns in gene expression (“Upstream Regulators” and “Causal Networks”), and

identify which genes of those included in the optimal datasets for each comparison are the most informative to pathway and network building (“Top Analysis-Ready Molecules”) for a given group of SB based upon gene expression (intensity and \log_2 ratio of fold change between groups) (Krämer et al., 2014). The significance of pathways identified through IPA is representative of the probability that the molecules within the dataset associate to a given pathway by random chance alone (low significance), or not (high significance). A z-score was calculated from the \log_2 fold change of gene expression values to measure the state of molecule activation or inhibition and ultimately predict what biological mechanisms may underlie differences in gene expression and subsequently phenotype or other variation between groups. The Qiagen Ingenuity® Knowledge Base (i.e., molecule, network, and pathway information) is specific to human (*Homo sapiens*), mouse (*Mus musculus*), and rat (*Rattus norvegicus*) systems, and therefore gene symbols associated to the SB gene transcripts in the five optimal datasets were converted into human gene symbols. The *HUGO Gene Nomenclature Committee (HGNC) at the European Bioinformatics Institute Multi-Symbol Checker tool* (<https://www.genenames.org/tools/multi-symbol-checker/>) was first used to convert any SB symbols to human symbols if possible. In instances where the HGNC tool did not match a human gene symbol to the SB symbol, ortholog information was obtained through the associated NCBI Gene page (i.e., by SB gene ID number). Any SB genes for which the human ortholog was not obtainable (i.e., previously identified) directly through the NCBI page or secondarily based upon information published about Zebrafish (*Danio rerio*) orthologs on NCBI and/or the Zebrafish Information Network (ZFIN) were subject to the NCBI nucleotide BLAST® tool against the Zebrafish genome (GRCz11, GenBank® accession: GCA_000002035.4) and determined based upon sequence homology (percent identity ranged from 74.32–100.00 % and

e-values ranged from 0.00–7.00E-16). Homologs for seven genes (nine gene transcripts, two sets of duplicates) were unable to be confidently retrieved through NCBI nBLAST®, six of these genes (eight gene transcripts) were in the optimal dataset of the Dam comparison and one from the Sire comparison. Additionally, each optimal dataset included at least one set of two transcripts encoded by the same gene and these duplicates were resolved in IPA by consolidating duplicates based upon maximum \log_2 ratio (i.e., the duplicate with the greater \log_2 ratio value was incorporated into pathway analysis).

Results

Experimental Animals Growth, Parentage, and Morphometric Comparisons

A summary of morphometric (weight, TL) and feed (total offered) data throughout the twelve-month duration the SB were raised in the RAS is presented in **Table 3.3**. The mean weight and amount of feed consumed per fish is shown in **Figure 3.1**. The weight of all SB at five months of age was 53.67 ± 1.61 g (range: 28.35–104.89 g). At eleven months of age, fish in two of the three replicate tanks were significantly smaller by weight than the third tank ($p=0.0007$) (**Table 3.3**). One of these two tanks remained significantly different in weight from the third at fourteen months of age ($p=0.0024$) (**Table 3.3**). Fish in two of the tanks significantly differed from fish in the third tank by TL at eight months ($p=0.0083$) and eleven months ($p=0.0002$) (**Table 3.3**). No significant differences of weight or TL were identified between the replicate tanks at other sampling points including at eighteen months of age when this portion of the study concluded (**Table 3.3**). The mean weight (all values presented as mean \pm standard deviation herein) of all SB at eighteen months of age was 1109.00 ± 33.42 g (range: 570–1720 g) and twenty-five SB of the total population ($N=173$, 14.45 %) had reached or exceeded market

size of 1360 g (1.36 kg, or 3.0 lbs) by this time. The mean FCR values calculated between tanks and across all sampling periods ranged from 1.24 ± 0.017 (five to eight months) to 2.17 ± 0.087 (fourteen to eighteen months) (**Table 3.3**). The FCR calculated for the entire population over the course of one year from five to eighteen months of age was 1.68.

The two groups of sacrificed SB, Superior (n=36, 1.32 ± 0.18 kg weight, 449.00 ± 16.54 mm TL) and Inferior (n=36, 0.85 ± 0.12 kg weight and 399.58 ± 15.96 mm TL) significantly differed in weight and length ($p < 0.0001$) (**Table 3.4**). The frequency distribution of weight of sacrificed SB is shown in **Figure 3.2**. The right-most extreme of what is roughly a bimodal distribution represents the SB that reached or exceeded Market Size (n=12 of 72) (**Figure 3.4**). These Market Size fish weighed 1.52 ± 0.13 kg and were 465.67 ± 13.27 mm TL (**Table 3.4**). The Under Size SB that exhibited the poorest growth performance throughout the study (n=12 of 72) weighed 0.72 ± 0.060 kg and were 382.58 ± 8.50 mm TL (**Table 3.4**). The SB in the Market Size and Under Size groups significantly differed from each other as well as those remaining in population of sacrificed SB, the Superior Other group (n=24 of 36 Superior SB; 1.21 ± 0.090 kg weight and 440.67 ± 10.72 mm TL) and the Inferior Other group (n=24 of 36 Inferior SB; 0.91 ± 0.076 kg weight and 408.083 ± 11.27 mm TL) ($p < 0.0001$) (**Table 3.4**).

Fifty-five of the seventy-two (55 of 72, 76.38 %) sacrificed SB offspring were produced from Dam A, and seventeen (17 of 72, 23.61 %) of the sacrificed SB offspring were produced from Dam B. The sacrificed SB did not significantly differ by weight or TL between Dam parentage groups (**Table 3.5**). Twenty-seven of all sacrificed SB offspring from both Dam A and B (27 of 72, 37.50 % Large sires) were produced from sires in the Large group and forty-five from sires in the Small group (45 of 72, 62.50 % Small sires). Sire 6-S from the Small group crossed with Dam A was determined to be the paternal broodfish for over two times the number

of sacrificed offspring than any other of the eleven remaining sires (22 of 72 offspring, other sires produced less than 10 offspring of the 72 each) (**Figure 3.3**). The weight of these offspring significantly differed from the weight of offspring produced from sire 3-L that was also crossed with Dam A and is therefore a half sibling family, and those produced from sire 12-S, which were not of a half-sibling family and did not significantly differ from those produced of sire 3-L ($p \leq 0.0201$; **Figure 3.3**).

Statistically significant correlations between SB offspring weight and TL and SB sire weight and TL were not observed via simple linear regression. The fitted regression models for weight and TL were: SB Offspring Weight = $0.1297 * \text{Sire Weight} + 0.8044$ ($p = 0.0581$, $R^2 = 0.050$) and SB Offspring TL = $0.1042 * \text{Sire TL} + 369.80$ ($p = 0.2562$, $R^2 = 0.018$), respectively. However, the offspring produced from sires in the Large group were found to be significantly larger by weight and TL than those produced from sires in the Small group ($p \leq 0.0376$, **Table 3.5**). Seventeen of the Large sire offspring were of the Superior group (23.61 % all offspring, 62.92 % Large Sire offspring) and ten were of the Inferior group (13.89 % of all offspring, 37.04 % Large Sire offspring). Nineteen of the Small sire offspring were of the Superior group (26.39 % all offspring, 42.22 % Small Sire offspring) and twenty-six of which of the Inferior group (36.11 % of all offspring, 57.78 % Small Sire offspring). Further, nine of the Superior offspring produced from Large sires were in the Market Size group (three from Small Sires) and four of the Inferior offspring in the Under Size group (eight from Small sires).

Muscle Histology

Significant differences in average number of fibers and average fiber diameter were present between growth groupings of SB offspring. SB in the Inferior group had a greater

number of muscle fibers (134.0 ± 25.40 fibers/slide) than those in the Superior growth group (119.1 ± 22.38 fibers/slide, $p=0.0102$) and these fibers were of smaller average diameter (average fiber diameter was 61.56 ± 6.058 μm for Inferior group and 65.82 ± 8.054 μm for Superior group, $p=0.0136$). Of these muscle fibers, there were significantly more that were <20 μm in diameter measured from the Inferior growth group (53.92 ± 26.23 and 37.75 ± 23.07 fibers <20 μm in the Inferior and Superior growth groups, respectively, 0.0070). Statistically significant correlations between SB weight and TL and number of fibers and average fiber diameter were identified via simple linear regression. The fitted regression models for SB weight are: Number of Fibers = $-30.96 \cdot \text{weight} + 160.1$ ($p=0.0027$, $R^2=0.12$) and Average Fiber Diameter = $8.80 \cdot \text{weight} + 54.18$ ($p=0.0041$, $R^2=0.11$). The fitted regression models for SB TL are: Number of Fibers = $-0.29 \cdot \text{TL} + 251.2$ ($p=0.0026$, $R^2=0.12$) and Average Fiber Diameter = $0.081 \cdot \text{TL} + 29.15$ ($p=0.0051$, $R^2=0.11$).

RNA-Sequencing and Quantitative Analysis of Gene Expression

A total of 6.2 billion paired reads from the seventy-two SB muscle tissue samples (mean \pm SD of $86,231,027 \pm 13,805,902$ reads per individual) were subject to further quantitative analysis after quality checking and trimming. An average of 95.35 ± 0.30 % of reads (per individual) were mapped to the SB reference genome (NCSU_SB_2.0) and expression values of 32,018 gene transcripts (22,746 genes) were quantified and statistically compared using CLC Genomics Workbench workflows and tools. The results of the ML, statistical, and pathway analyses for comparisons between SB growth (Growth, Market Size) and parentage (Dam, Sire, Sire Size) are presented together in subsections below. The number of gene transcripts

determined to be important to the five comparisons based upon entropy (information gain) determined through the ML approach and/or statistical analysis are provided in **Table 3.6**.

Striped Bass Gene Expression and Growth

Of the 32,018 gene transcripts measured, 401 significantly differed (FDR $p \leq 0.05$) in expression between Superior and Inferior growth groups of SB and 965 were assigned an information gain value above 0.00 based upon information gain for the Growth comparison (**Table 3.6**). A total of ninety-one gene transcripts of the 965 ranked based upon entropy were also included in the set of transcripts that significantly differed between groups (i.e., were “differentially expressed”) (**Table 3.6**). In total, 1,275 transcripts (1,211 unique genes) have been determined as important to the Growth comparison by either statistical analysis, ML, or both. Optimal learning occurred when the top ten and top 300 ranked attributes were input into cross-validated algorithms and thus these were considered the thresholds for model underfitting and overfitting, respectively (**Figure 3.4**). Thirty-two of the differentially expressed transcripts were also among the top 300 attributes, one of which was in the top ten. The mean \pm SD percent correct classification of all cross-validated algorithms when the top ten ranked attributes were input was 80.56 ± 5.25 % with a range of 70.83–87.50 % (mean \pm SD Kappa statistic and AUROC were 0.61 ± 0.10 and 0.83 ± 0.078 , respectively). The mean \pm SD percent correct classification when the top 300 ranked attributes were input was 78.99 ± 10.16 % with a range of 66.67–94.44 % (Kappa statistic and AUROC were 0.57 ± 0.21 and 0.84 ± 0.10 , respectively). The top 300 ranked attributes (300 gene transcripts) annotated to 294 unique genes and were included in subsequent pathway analysis as the optimal dataset for the Growth comparison (**Appendix A**).

The majority of genes (272 of 294) in the optimal dataset were up-regulated in SB of the Inferior Growth group and the top (i.e., greatest enrichment and/or prediction of activity) canonical pathways identified for SB in both groups are shown in **Figure 3.9**. A graphical summary of upstream regulators and causal networks predicted to underlie the observed patterns of gene expression in the data and that are determined to be significant (Fisher's Exact Test, $p \leq 0.05$) for Inferior fish is shown in **Figure 3.10**. There were not enough connectable entities between the twenty-two genes up-regulated in Superior fish and the information in the Ingenuity® Knowledge Base to generate a network diagram of upstream regulators and causal effects of equal depth as that for Inferior SB.

Striped Bass Gene Expression and Growth to Market Size

For the comparison between Market Size, Superior, Inferior, and Under Size SB, 621 transcripts were differentially expressed (FDR $p \leq 0.05$) between groups and 284 were assigned an information gain value above 0.00 based upon information gain, forty-three of which were among those identified as differentially expressed (**Table 3.6**). In total, 862 transcripts (785 unique genes) have been determined as important to the Market Size growth comparison by either statistical analysis, ML, or both. Including the top seventy-five and all 284 ranked attributes in learning led to optimal model performance (**Figure 3.5**). The mean \pm SD percent correct classification of all cross-validated algorithms when the top seventy-five ranked attributes were input was 54.86 ± 5.85 % with a range of 50.00–66.67 % (mean \pm SD Kappa statistic and AUROC were 0.36 ± 0.10 and 0.75 ± 0.073 , respectively). The mean \pm SD percent correct classification when all 284 ranked attributes were input was 65.63 ± 16.29 % with a range of 40.28–83.33 % (Kappa statistic and AUROC were 0.51 ± 0.24 and 0.81 ± 0.14 ,

respectively). The complete set of ranked attributes (284 transcripts) annotated to 277 unique genes and was included in subsequent pathway analysis as the optimal dataset for the Market Size comparison (**Appendix B**).

Most of the genes were identified as up-regulated in the Under Size SB (194 of 277 genes), and the remaining vary in pattern of up-regulation (25 genes up-regulated in Inferior Other SB, 10 genes in Superior Other SB, and 49 genes in Market Size SB). Six of the seven sets of transcript variants that map to the same gene (“duplicates”) all up-regulated in Under Size SB, however, the one transcript of the seventh set was up-regulated in Under Size SB and the other was up-regulated in Market Size SB. The top canonical pathways identified for all four groups are shown in **Figure 3.11**. A graphical summary of upstream regulators and causal networks predicted to underlie the observed patterns of gene expression in the data and that are determined to be significant (Fisher’s Exact Test, $p \leq 0.05$) for Under Size SB is shown in **Figure 3.12** and **Figure 3.13** for Market Size SB. There were not enough connectable entities between the genes up-regulated in Superior Other or Inferior Other SB and the information in the Ingenuity® Knowledge Base to generate a network diagram of upstream regulators and causal effects of equal depth as that for Market Size and Under Size SB.

Striped Bass Gene Expression and Dam Parentage

The comparison between SB born to Dam A and Dam B identified 102 transcripts as differentially expressed (FDR $p \leq 0.05$) between groups and 257 gene transcripts were assigned an information gain value above 0.00 based upon information gain (**Table 3.6**). Of the 257 transcripts assigned an information gain value above 0.00, thirty-one were also differentially expressed. In total, 328 transcripts (317 unique genes) have been determined as important to the

Dam parentage comparison by either statistical analysis, ML, or both. Including the top twenty-five and top 150 ranked attributes in learning led to optimal model performance (**Figure 3.6**). The mean \pm SD percent correct classification of all cross-validated algorithms when the top twenty-five ranked attributes were input was 94.27 ± 5.91 % with a range of 83.33–100.00 % (mean \pm SD Kappa statistic and AUROC were 0.75 ± 0.35 and 0.89 ± 0.21 , respectively). The mean \pm SD percent correct classification when the top 150 ranked attributes were input was 93.40 ± 9.05 % with a range of 73.61–100.00 % (Kappa statistic and AUROC were 0.77 ± 0.29 and 0.92 ± 0.15 , respectively). The top 150 ranked attributes (150 gene transcripts) annotated to 139 unique genes include in pathway analysis as the optimal dataset for the Dam comparison; paralogs could not be identified for eight transcripts, two of which were variants of the same gene, and there were three additional sets of duplicates that were annotated to paralogous genes (**Appendix C**).

The Dam pathway analysis identified roughly two-thirds of genes (94 of 139, 67.63 %) as being up-regulated in SB born to Dam B. The top canonical pathways identified for Dam A and Dam B offspring are shown in **Figure 3.14**. A graphical summary of upstream regulators and causal networks predicted to underlie the observed patterns of gene expression in the data and that are determined to be significant (Fisher's Exact Test, $p \leq 0.05$) for Dam B offspring is shown in **Figure 3.15**. There were not enough connectable entities between the forty-five genes up-regulated in Dam A offspring and the information in the Ingenuity® Knowledge Base to generate a network diagram of upstream regulators and causal effects of equal depth as that for Dam B offspring.

Striped Bass Gene Expression and Sire Parentage

The comparison between SB offspring born to individual sires (1–12) identified 1,177 transcripts as significantly differing (FDR $p \leq 0.05$) between groups and only thirty-five transcripts as having an information gain value above 0.00 based upon information gain (**Table 3.6**). Twenty-two of these transcripts were shared between sets for the Sire comparison (**Table 3.6**). In total, 1,190 transcripts (987 unique genes) have been determined as important to the Sire parentage comparison by either statistical analysis, ML, or both. Optimal performance was achieved when all thirty-five ranked attributes were included (**Figure 3.7**) and the mean \pm SD percent correct classification of all cross-validated algorithms was 74.13 ± 15.94 % with a range of 45.83–91.67 % (mean \pm SD Kappa statistic and AUROC were 0.68 ± 0.19 and 0.89 ± 0.10 , respectively). The complete set of ranked attributes (35 gene transcripts) annotated to 33 unique genes, one duplicate and one transcript that could not be mapped to a paralog were removed, and these genes were included in subsequent pathway analysis as the optimal dataset for the Sire comparison (**Appendix D**).

Offspring born to Sire 9 (9-L, n=2 of 72 offspring) of the Large size group and crossed with Dam B had the greatest number of up-regulated gene transcripts relative to the other eleven sires (9 of 33, 25.71 %). Offspring born to Sire 2 (2-L, n=5 of 72 offspring) of the Large size group and crossed with Dam A had the greatest number of down-regulated genes relative to the other eleven sires (8 of 33, 22.86 %). The top canonical pathways identified for these genes are shown in **Figure 3.16**.

Striped Bass Gene Expression and Sire Size

The comparison of SB offspring born from sires in the Large or Small groups identified 150 transcripts as differentially expressed (FDR $p \leq 0.05$) and 378 as having an information gain value above 0.00 based upon information gain (**Table 3.6**). Of the ranked transcripts, twenty-eight of the 378 are considered differentially expressed based upon statistical analysis (**Table 3.6**). In total, 500 transcripts (471 unique genes) have been determined as important to the Sire Size parentage comparison by either statistical analysis, ML, or both. Including the top fifteen and top seventy-five ranked attributes in learning led to optimal model performance (**Figure 3.8**). The mean \pm SD percent correct classification of all cross-validated algorithms when the top fifteen ranked attributes were input was 93.06 ± 5.30 % with a range of 84.72–100.00 % (mean \pm SD Kappa statistic and AUROC were 0.85 ± 0.11 and 0.95 ± 0.047 , respectively). The mean \pm SD percent correct classification when the top seventy-five ranked attributes were input was 92.71 ± 5.64 % with a range of 84.72–100.00 % (Kappa statistic and AUROC were 0.84 ± 0.12 and 0.94 ± 0.071 , respectively). The top seventy-five attributes annotated to seventy-one genes that were included in subsequent pathway analysis as the optimal dataset (**Appendix E**). A paralog was not identified for one of the gene transcripts and the other three genes each encoded two transcripts among those in the optimal dataset, one transcript for each gene was up-regulated in either sire size group.

Over half of the genes (42 of 71 genes, 59.15 %) in the Sire Size comparison were up-regulated in SB produced from Large sires and the top canonical pathways identified for SB in both groups are shown in **Figure 3.17**. There were not enough connectable entities between the sets of genes up-regulated in either group of SB produced from Large sires or Small sires and the

information in the Ingenuity® Knowledge Base to generate network diagrams of upstream regulators and causal effects.

The top analysis-ready (i.e., providing the most information for pathway and network building) molecules for all comparisons and groups thereof are listed in **Appendix F**. Similarly, the most important of the upstream regulators and causal networks identified for each comparison and group thereof are listed in **Appendices G and H**, respectively.

Discussion

The mean weight of the SB raised for this study at eighteen months of age is representative of the expected performance (i.e., growth rate) for domestic SB of filial generation (F6) in terms of gains in performance traits compared to previous generations (Andersen et al., 2021). The FCR values calculated between each sampling event and for the duration of the study (FCR range over all sampling intervals was 1.24 to 2.20 and FCR=1.68 overall for the full year in RAS) are within the standard range (1.0–2.4) for aquacultured finfish (**Figure 3.1**). The distinct (i.e., significant) differences between the Superior and Inferior groups of SB, as well as the “extremes” of the population (Market Size and Under Size), indicate the pathway analysis described below provides a representative snapshot of the phenomena underlying differences in the growth phenotype observed among a cohort of aquacultured SB.

Over half of the 294 genes (171 unique genes from 173 transcripts) identified as important to the classification of SB into Growth comparison groups (Superior, Inferior) were the same as those identified for the Market Size comparison (Market Size, Superior Other, Inferior Other, Under Size), indicating that these genes are in fact associated with growth processes. As the majority of these and the other genes identified in the Growth and Market Size comparisons were up-regulated in Inferior SB and Under Size SB, respectively, the canonical

pathways and predicted upstream regulators and downstream effects identified for both appear to primarily reflect the phenomena underlying the effect(s) leading to poor growth in domestic SB (i.e., Inferior SB and Under Size SB) (**Figures 3.9–3.12**). Interestingly, the patterns of gene expression for those identified in the Market Size comparison are such that most genes up-regulated in Under Size SB are down-regulated in Market Size SB and vice versus, subsequently leading to the prediction of similar pathways and processes as relevant to both groups and, in some cases, as having opposite activation states (e.g., inhibited in one and activated in the other). This observation is similar to that of genes related to egg fertility in SB as reported in Sullivan et al. (2015). This distinctly “opposite” pattern of gene expression between groups representative of the extremes of the population by size was not necessarily observed in the intermediate phenotype groups, as growth is a complex polygenic trait expected to be influenced by several factors including genes and the interactions thereof. Moreover, opposite patterns of gene expression between groups designated based upon phenotype are not observed universally. This also was similar to the observation using an evaluation of Shannon’s Entropy of expressed gene transcripts in SB ovary, whereby the highest and lowest fertility groups had a similar organizational states compared to the intermediate phenotypes, which had a lower organization state (i.e., disorganized) (Sullivan et al., 2015).

The genes and subsequent pathway predictions across all comparisons identified critical immune response mechanisms that appear to differentiate SB of superior growth from those of inferior growth traits. One cluster of genes up-regulated in Inferior and Under Size SB and are associated with a number of the top canonical pathways identified for these fish encode members of the 26S proteasome, a proteinase complex comprised of a 20S core and a 19S regulator sub-complex, responsible for the regulation of protein degradation in eukaryotes (Livneh et al., 2016;

Tanaka et al., 2009; Reading et al., 2013). Specifically, the genes identified include four members of the proteasome 20S (core) subunit alpha (*PSMA2*, *PSMA6*, *PSMA8*) and beta (*PSMB5*), eight of the 19S (regulator) ATPase (*PSMC1*, *PSMC2*, *PSMC4*, *PSMC6*) and non-ATPase subunits (*PSMD6*, *PSMD12*, *PSMD13*, *PSMD14*), and one of the gamma subunit of the 11S alternate regulator (*PSME3*). Of the top canonical pathways identified and shared between the Growth and Market Size comparisons, this 26s proteasome cluster is associated to the following, all of which have established roles in protein degradation via the ubiquitin system: BAG2 Signaling, FAT10 Signaling, Protein Ubiquitination, Inhibition of ARE-Mediated mRNA Degradation, and Huntington's Disease Signaling (**Figures 3.9(B)** and **3.11(A,C)**). Other genes identified as associated to these pathways from the Growth and Market Size comparison include cytochrome c (*CYCS*) that has been linked to the initiation of apoptosis, cell death, and decay, and heat shock protein 90 alpha family class A member 1 (*HSP90AA1*), which has been linked to cellular repair mechanisms by its promotion of cell motility via signaling through the low density lipoprotein receptor-related protein 1 (LRP-1) receptor (Fossati et al., 2010; Jayaprakash et al., 2015).

The ubiquitin-proteasome system (UPS) is a major protein degradation system critical for skeletal muscle homeostasis, specifically through muscle atrophy, and has been linked to defects in muscle regeneration and muscle diseases (Kitajima et al., 2020). For example, the mediation of protein degradation via proteolytic pathways (e.g., proteasome) has been linked to the reduced diameter of myofibers measured from human patients with the wasting syndrome cachexia (Kitajima et al., 2020). Similarly, suppression of the 26S proteasome in mice has been found to facilitate myofiber protein activity that supports myofiber size (i.e., muscle hypertrophy) (Wang et al., 2019). Although not as well characterized in fishes as their terrestrial counterparts, a

negative correlation between 20S proteasome activity in the white muscle tissue of spotted wolffish (*Anarhichas minor*) and specific growth rate has been identified such that degradation via the UPS is thought to underlie lower protein efficiency and growth rate observed when fish were at lower temperatures (Lamarre et al., 2010).

The overlapping elements shown in the network diagrams of predicted upstream regulators for Inferior SB (**Figure 3.10**), Under Size SB (**Figure 3.12**), and Market Size SB (**Figure 3.13**) similarly point to an activated immune response in SB of inferior growth performance. Specifically, upstream regulators involved in T lymphocyte (T-cell) development and proliferation are shared between Inferior SB and Market Size SB, although opposite in activation/inhibition (**Figures 3.10** and **3.13**). T-cells are generally associated with degeneration of skeletal muscle as part of an immune response, although some recent evidence in mice points to their role in skeletal muscle repair and regeneration after severe damage (Deyhle and Hyldahl, 2018). The CD28 molecule shown as activated in Inferior SB is essential for T-cell proliferation and survival and is predicted to activate LCK proto-oncogene (LCK) and interleukin 2 (IL-2) in these fish (**Figures 3.10**). The same CD48 and LCK molecules are shown as inhibited in Market Size SB, whereby an indirect inhibitory action from CD28 to interleukin 4 (IL-4) is also shown (**Figure 3.13**). The LCK protein has a critical signaling role in the selection and maturation of developing T-cells and interleukins, particularly IL-2 and IL-4, which shares a gamma chain with IL-2 and IL-7 in the heterotrimeric receptor of the interleukin cytokine protein complex, have been identified as important in eliciting the pro-inflammation response required for the repair of damaged muscle tissue, as well as the proliferation of T (and B) lymphocytes. (Chapman et al., 2013; Deyhle and Hyldahl, 2018; McKenzie et al., 1999; Rosa Neto et al., 2011; Zhou et al., 2018).

It may be that the inferior growth phenotype is in part due to insufficient immune response mechanisms that perpetuate limited muscle growth. This phenomenon may be best explained by the identified activation of the hypoxia-inducible factor α (HIF-1 α) Signaling pathway in Market Size SB, as well as Dam B offspring, that was inhibited in Inferior and Under Size SB (**Figures 3.9, 3.11, and 3.14**). The HIF-1 α Signaling pathway represents the response to hypoxia, including after transient hypoxic events following acute exercise or tissue damage events, and plays an important role in the physiology and homeostasis of skeletal muscle (Valle-Tenney et al., 2020). The HIF-1 α Signaling pathway is known to regulate proteins involved in angiogenesis and growth, as well as coordinate the metabolic demand such that supply and production of ATP are aligned and a bioenergetic collapse avoided (Wheaton and Chandel, 2011). The inhibition of HIF-1 α Signaling in SB exhibiting inferior growth is suggestive of dysfunction in balancing metabolic processes. Pyruvate kinase M1/2 (PKM) was associated to the HIF-1 α Signaling pathway as up-regulated in Dam B offspring and Market Size SB and down-regulated in Under Size SB. PKM2 catalyzes the last step in glycolysis whereby pyruvate and adenosine triphosphate (ATP) are formed from the transfer of a phosphate from phosphoenolpyruvate (PEP) to adenosine diphosphate (ADP) and PKM2 has been identified as a coactivator in stimulating HIF-1 α to initiate the switch from oxidative to glycolytic metabolism (Luo et al., 2012). Glycolysis I was identified as the top pathway in Superior SB and was also enriched in Market Size SB via up-regulation of glucose-6-phosphate isomerase (GPI) and triosephosphate isomerase 1 (TPI1) in addition to PKM (**Figure 3.9 and 3.11**). The trends elucidated through the pathway analysis may suggest that the SB of superior growth traits are able to employ a hypometabolic strategy to mitigate stress, whether muscle repair, crowding (i.e., hypoxia), or other, and subsequently preserve accumulated muscle mass to some extent possible,

whereas the SB of inferior growth resort to protein degradation and muscle atrophy.

Hypertrophic muscle growth via metabolic reprogramming mechanisms akin to those characterized for rapidly growing cancer cells (e.g., Warburg effect) and whereby glycolysis and oxidative phosphorylation enable a growth advantage of myotubes have been identified in mice and may be what is occurring in these SB (Stadhouders et al., 2020). Although speculative, it is possible the greater number of muscle fibers that are smaller in diameter in Inferior SB compared to Superior SB are not indicative of a hyperplastic growth, but rather a muscle wasting syndrome or similar catabolism of protein (Nguyen et al., 2021).

The determination of the extent to which the potential to elicit a more advantageous stress mitigation strategy is heritable is outside the scope of the study described herein. However, certain trends between expressed genes in the dataset and predicted pathways and regulators thereof provide some insight into the differences between responses in SB that grow well and those that do not. Specifically, enriched pathways of SB produced by Large sires included some associated to cellular stress, injury, and/or inflammation response (CLEAR Signaling Pathway, ID1 Signaling Pathway, Ceramide Signaling, Pyroptosis Signaling Pathway) and several associated with cellular growth, proliferation, and development processes (STAT3 Pathway, Acute Phase Response Signaling, Regulation of the Epithelial Mesenchymal Transition by Growth Factors Pathway, Retinoate Biosynthesis II, FGF Signaling, CDK5 Signaling). A number of these pathways are aligned with the immune response that would be anticipated as a result of muscle differentiation, high activity (i.e., swimming), and/or stress mitigation, and would subsequently suggest a connection to the observable phenotypic differences in offspring produced from Large sires compared to Small sires (**Table 3.5, Figure 3.17(A)**).

For example, the inhibitor of DNA binding 1 (ID1) Signaling Pathway was identified as enriched for offspring of Dam B and of Large sires (**Figures 3.14 and 3.17**). The ID1 proteins

have been implicated in an adaptive survival strategy to HIF-1 α inhibition, whereby up-regulation of ID1 is positively correlated to the regulation of HIF-1 α (Geng et al., 2021). In doing so, ID1 proteins support tumor growth and compensate for the loss of HIF-1 α by up-regulating glutaminase 2 (GLS2) and glutamine metabolism such that HIF-1 α -inhibited cells transfer metabolic dependency from glucose to glutamine (Geng et al., 2021). Further, the CLEAR (Coordinated Lysosomal Expression and Regulation) Signaling Pathway is the most significant for the SB produced from Large Sires and has known primary functions in cellular stress and injury, including being linked to immune response and protein degradation (**Figure 3.17(A)**) (Palmieri et al., 2011). The Signal Transducer and Activator of Transcription 3 (STAT3) Pathway is the second most significant canonical pathway identified for the Large Sire offspring group and is well-characterized as a regulator of skeletal muscle mass and repair (Guadagnin et al., 2018). These pathways share three associated genes from the Sire Size dataset, specifically: fibroblast growth factor receptor 3 (*FGFR3*), mitogen-activated protein kinase 14 (*MAPK14*), and nerve growth factor receptor (*NGFR*). Abou Sawan et al. (2020) recently summarized the potential roles of lysosomes (i.e., the CLEAR motif) in anabolic processes in skeletal muscle, such as in muscle mass maintenance with supporting evidence from animal studies that found hypertrophy may be positively correlated with autophagy markers. Additionally, STAT3 signaling has a role in the regulation of several skeletal muscle cell types, including muscle stem cells, myofibers, and macrophages as well as stimulating cellular glycolysis (Douros et al., 2018; Sala and Sacco, 2016; Seif et al., 2017). The STAT3 Pathway was not identified among the top pathways for Superior SB as it was for Superior Other SB from the Market Size comparison (**Figures 3.9 and 3.11**), however, Glycolysis I was the top canonical pathway identified for this group as mentioned above. Increased glycolysis in skeletal muscle has

been linked to accelerated glucose disposal and high metabolic rate in mice and other organisms (Xiang et al., 2021). Therefore, the STAT3 pathway, glycolysis, and related functions are likely a major component of the physiological processes that differentiate SB exhibiting superior growth from those that do not (Xiang et al., 2021).

Acknowledgements

We thank Dr. Andrew S. McGinty, Michael S. Hopper, and Robert W. Clark for their efforts in producing these fish; Dr. Sarah Rajab, Connor Neagle, and Fara Marin for assistance in rearing and sampling; Dr. David A. Baltzegar for assistance in data acquisition; Valerie Williams for coordinating microsatellite genotyping; and Fatma Kahn for assisting with gene annotation. This work was supported by funding provided from the following sources: the Foundation for Food and Agriculture Research (FFAR) *New Innovator Award*, the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), the National Oceanic and Atmospheric Administration (NOAA) and National Sea Grant (E/2019-AQUA-02, a project to establish a Striped Bass Aquaculture Hub, the *StriperHub*), and North Carolina Sea Grant. Striped bass is a priority species for the USDA National Research Support Project 8 (NRSP-8; *National Animal Genome Research Project*) and funding to support the National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry was provided by the NRSP-8 and the USDA NIFA (Hatch Multistate Project), the USDA Agricultural Research Service (ARS, Harry K. Dupree Stuttgart National Aquaculture Research Center), the North Carolina State University College of Agriculture and Life Sciences and College of Sciences, the North Carolina Agricultural Foundation William White Endowment, and various industry stakeholders including Premex, Clay Chappell, Locals Seafood, and several others who wish to remain anonymous. L.K. Andersen also received support Center for Environmental Farming Systems Graduate Fellowship program, the North Carolina State University, Raleigh, NC, Biotechnology Program (BIT), and the Coastal Conservation Association of North Carolina David and Ann Speaks Coastal Conservation Association Scholarship. This will be of the publications from to the North Carolina State University Pamlico Aquaculture Field Laboratory.

References

- About Sawan, S., Mazzulla, M., Moore, D.R. and Hodson, N. (2020). More than just a garbage can: emerging roles of the lysosome as an anabolic organelle in skeletal muscle. *American Journal of Physiology-Cell Physiology*, 319(3), pp.C561-C568. DOI: 10.1152/ajpcell.00241.2020.
- Andersen, L.K., Abernathy, J., Berlinsky, D.L., Bolton, G., Booker, M.M., Borski, R.J., Brown, T., Cerino, D., Ciaramella, M., Clark, R.W., Frinsko, M.O., Fuller, S.A., Gabel, S., Green, B.W., Herbst, E., Hodson, R.G., Hopper, M.S., Kenter, L.W., Lopez, F., McGinty, A.S., Nash, B., Parker, M., Pigg, S., Rawles, S., Riley, K., Turano, M.J., Webster, C.D., Weirich, C.R., Won, E., Woods III, L.C., and Reading, B.J. 2021. The status of striped bass, *Morone saxatilis*, as a commercially ready species for US marine aquaculture. *Journal of the World Aquaculture Society*, 52(3), pp.710-730. DOI: 10.1111/jwas.12812.
- Bayless, J. D., 1972. Artificial Propagation and Hybridization of Striped Bass, *Morone Saxatilis* (Walbaum). South Carolina Wildlife Resources Department.
- Brown, K.M., Baltazar, G.A., and Hamilton, M.B. 2005. Reconciling nuclear microsatellite and mitochondrial marker estimates of population structure: breeding population structure of Chesapeake Bay striped bass (*Morone saxatilis*). *Heredity*, 94(6):606-615. DOI: 10.1038/sj.hdy.6800668.
- Chapman, N.M., Connolly, S.F., Reinl, E.L. and Houtman, J.C., 2013. Focal adhesion kinase negatively regulates Lck function downstream of the T cell antigen receptor. *The Journal of Immunology*, 191(12), pp.6208-6221. DOI: 10.4049/jimmunol.1301587.

- Conte, C., Riant, E., Toutain, C., Pujol, F., Arnal, J., Lenfant, F., & Prats, A. (2008). FGF2 Translationally Induced by Hypoxia Is Involved in Negative and Positive Feedback Loops with HIF-1 α . *Plos ONE*, 3(8), e3078. DOI: 10.1371/journal.pone.0003078.
- Couch, C.R., Garber, A.F., Rexroad III, C.E., Abrams, J.M., Stannard, J.A., Westerman, M.E., and Sullivan, C.V. 2006. Isolation and characterization of 149 novel microsatellite DNA markers for striped bass, *Morone saxatilis*, and cross-species amplification in white bass, *Morone chrysops*, and their hybrid. *Molecular Ecology Notes*, 6(3):667-669. DOI: 10.1111/j.1471-8286.2006.01292.x.
- De Domenico, I., McVey Ward, D., & Kaplan, J. (2008). Regulation of iron acquisition and storage: consequences for iron-linked disorders. *Nature Reviews Molecular Cell Biology*, 9(1), 72-81. DOI: 10.1038/nrm2295.
- Deyhle, M., & Hyldahl, R. (2018). The Role of T Lymphocytes in Skeletal Muscle Repair From Traumatic and Contraction-Induced Injury. *Frontiers In Physiology*, 9. DOI: 10.3389/fphys.2018.00768.
- Douros, J., Baltzegar, D., Reading, B., Seale, A., Lerner, D., Grau, E., & Borski, R. (2018). Leptin Stimulates Cellular Glycolysis Through a STAT3 Dependent Mechanism in Tilapia. *Frontiers In Endocrinology*, 9. DOI: 10.3389/fendo.2018.00465.
- Fossati, S., Cam, J., Meyerson, J., Mezhericher, E., Romero, I.A., Couraud, P.O., Weksler, B.B., Ghiso, J., & Rostagno, A. (2009). Differential activation of mitochondrial apoptotic pathways by vasculotropic amyloid- β variants in cells composing the cerebral vessel walls. *The FASEB Journal*, 24(1), 229-241. DOI: 10.1096/fj.09-139584.
- Fournier, B., Mahlaoui, N., Moshous, D. and de Villartay, J.P., 2022. Inborn errors of immunity caused by defects in the DNA damage response pathways: Importance of minimizing

- treatment-related genotoxicity. *Pediatric Allergy and Immunology*, 33(6), p.e13820. DOI: 10.1111/pai.13820.
- Ganz, T. (2008). Iron Homeostasis: Fitting the Puzzle Pieces Together. *Cell Metabolism*, 7(4), 288-290. DOI: 10.1016/j.cmet.2008.03.008.
- Geng, H., Ko, H., Pittsenbarger, J., Harvey, C., Xue, C., & Liu, Q. et al. (2021). HIF1 and ID1 Interplay Confers Adaptive Survival to HIF1 α -Inhibition. *Frontiers In Cell And Developmental Biology*, 9. DOI: 10.3389/fcell.2021.724059.
- Guadagnin, E., Mázala, D., & Chen, Y. (2018). STAT3 in Skeletal Muscle Function and Disorders. *International Journal Of Molecular Sciences*, 19(8), 2265. DOI: 10.3390/ijms19082265.
- Han, K., Li, L., Leclerc, G.M., Hays, A.M., and Ely, B. 2000. Isolation and characterization of microsatellite loci for striped bass (*Morone saxatilis*). *Marine Biotechnology*, 2(5):405-408. DOI: 10.1007/s101260000014.
- Hodson, R.G. and Sullivan, C.V. 1993. Induced maturation and spawning of domestic and wild striped bass (*Morone saxatilis*) broodstock with implanted GnRH analogue and injected hCG. *Journal of Aquaculture and Fisheries Management*, 24:271-280. DOI: 10.1111/j.1365-2109.1993.tb00562.x.
- Huang, H., Nguyen, T., & Pickett, C. (2000). Regulation of the antioxidant response element by protein kinase C-mediated phosphorylation of NF-E2-related factor 2. *Proceedings Of The National Academy Of Sciences*, 97(23), 12475-12480. DOI: 10.1073/pnas.220418997.
- Jayaprakash, P., Dong, H., Zou, M., Bhatia, A., O'Brien, K., & Chen, M. et al. (2015). Hsp90 α and Hsp90 β Co-Operate a Stress-Response Mechanism to Cope With Hypoxia and

- Nutrient Paucity during Wound Healing. *Journal Of Cell Science*. DOI: 10.1242/jcs.166363
- Krämer, A., Green, J., Pollard, Jr., J., and Tugendreich, S. 2014. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*. 30(4):523–30.
- Kitajima, Y., Yoshioka, K., & Suzuki, N. (2020). The ubiquitin–proteasome system in regulation of the skeletal muscle homeostasis and atrophy: from basic science to disorders. *The Journal Of Physiological Sciences*, 70(1). DOI: 10.1186/s12576-020-00768-9.
- Lamarre, S. G., Blier, P. U., Driedzic, W. R., & Le Francois, N. R. (2010). White muscle 20S proteasome activity is negatively correlated to growth rate at low temperature in the spotted wolffish *Anarhichas minor*. *Journal of Fish Biology*, 76(7), 1565-1575. DOI: 10.1111/j.1095-8649.2010.02581.x.
- Li, C., Beck, B.H., Fuller, S.A., and Peatman, E. 2014. Transcriptome annotation and marker discovery in white bass (*Morone chrysops*) and striped bass (*Morone saxatilis*). *Animal Genetics*, 45(6):885-887. DOI: 10.1111/age.12211.
- Livneh, I., Cohen-Kaplan, V., Cohen-Rosenzweig, C., Avni, N., & Ciechanover, A. (2016). The life cycle of the 26S proteasome: from birth, through regulation and function, and onto its death. *Cell Research*, 26(8), 869-885. DOI: 10.1038/cr.2016.86.
- Luo, W., Hu, H., Chang, R., Zhong, J., Knabel, M., & O'Meally, R. et al. (2011). Pyruvate Kinase M2 Is a PHD3-Stimulated Coactivator for Hypoxia-Inducible Factor 1. *Cell*, 145(5), 732-744. DOI: 10.1016/j.cell.2011.03.054.
- Marshall, T.C., Slate, J.B.K.E., Kruuk, L.E.B., Pemberton, J.M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, 7(5), 639-655. <https://doi.org/10.1046/j.1365-294x.1998.00374.x>.

- McKenzie, G.J., Fallon, P.G., Emson, C.L., Grecis, R.K. and McKenzie, A.N., 1999. Simultaneous disruption of interleukin (IL)-4 and IL-13 defines individual roles in T helper cell type 2-mediated responses. *The Journal of experimental medicine*, 189(10), pp.1565-1572.
- Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., & Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621-628. DOI: 10.1038/nmeth.1226.
- Motulsky, H.J. and Brown, R.E. 2006. Detecting outliers when fitting data with nonlinear regression—a new method based on robust nonlinear regression and the false discovery rate. *BMC bioinformatics*, 7(1), pp.1-20. DOI: 10.1186/1471-2105-7-123.
- Mullis, A.W., and J.M. Smith. 1990. Artificial spawning and fry production of Striped Bass and in hybrids. Pages 7-16 in R.M. Harrell, J.H. Kerby, R.V. Minton, editors. *Culture and propagation of Striped Bass and its hybrids*. American Fisheries Society, Bethesda, Maryland.
- National Research Council (NRC), 1996. *Guide for the care and use of laboratory animals*. The National Academies Press. Washington, DC. DOI: 10.17226/5140.
- Nguyen, T., Conotte, S., Belayew, A., Declèves, A., Legrand, A., & Tassin, A. (2021). Hypoxia and Hypoxia-Inducible Factor Signaling in Muscular Dystrophies: Cause and Consequences. *International Journal Of Molecular Sciences*, 22(13), 7220. DOI: 10.3390/ijms22137220.
- Palmieri, M., Impey, S., Kang, H., di Ronza, A., Pelz, C., Sardiello, M., & Ballabio, A. (2011). Characterization of the CLEAR network reveals an integrated control of cellular

- clearance pathways. *Human Molecular Genetics*, 20(19), 3852-3866. DOI: 10.1093/hmg/ddr306.
- Pircher, T., Wackerhage, H., Aszodi, A., Kammerlander, C., Böcker, W., & Saller, M. (2021). Hypoxic Signaling in Skeletal Muscle Maintenance and Regeneration: A Systematic Review. *Frontiers In Physiology*, 12. DOI: 10.3389/fphys.2021.684899.
- Reading, B.J., Chapman, R.W., Schaff, J.E., Scholl, E.H., Opperman, C.H., and C.V. Sullivan. 2012. An ovary transcriptome for all maturational stages of the striped bass (*Morone saxatilis*), a highly advanced perciform fish. *BMC Research Notes*, 5:111. DOI: 10.1186/1756-0500-5-111.
- Reading, B.J., Williams, V.N., Chapman, R.W., Williams, T.I. and Sullivan, C.V., 2013. Dynamics of the striped bass (*Morone saxatilis*) ovary proteome reveal a complex network of the translasome. *Journal of proteome research*, 12(4), pp.1691-1699. DOI: 10.1021/pr3010293.
- Rees, R.A., Harrell, R.M., 1990. Artificial spawning and fry production of striped bass and in hybrids, in: Harrell, R.M., Kerby, J.H., Minton, R.V. (Eds.) *Culture and Propagation of Striped Bass and Its Hybrids*. American Fisheries Society, Bethesda, Maryland, pp. 43-72.
- Rexroad, C., Vallejo, R., Coulibaly, I., Couch, C., Garber, A., Westerman, M., and Sullivan, C. 2006. Identification and characterization of microsatellites for striped bass from repeat-enriched libraries. *Conservation Genetics*, 7(6):971-982. DOI: 10.1007/s10592-006-9122-0.

- Rosa Neto, J., Lira, F., Zanchi, N., Oyama, L., Pimentel, G., & Santos, R. et al. (2011). Acute exhaustive exercise regulates IL-2, IL-4 and MyoD in skeletal muscle but not adipose tissue in rats. *Lipids In Health And Disease*, 10(1), 97. DOI: 10.1186/1476-511x-10-97.
- Sala, D. & Sacco, A. (2016). Signal transducer and activator of transcription 3 signaling as a potential target to treat muscle wasting diseases. *Current Opinion In Clinical Nutrition And Metabolic Care*, 1. DOI: 10.1097/mco.0000000000000273.
- Seif, F., Khoshmirisafa, M., Aazami, H., Mohsenzadegan, M., Sedighi, G., & Bahar, M. (2017). The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. *Cell Communication And Signaling*, 15(1). DOI: 10.1186/s12964-017-0177-y.
- Skalski, G.T., Couch, C.R., Garber, A.F., Weir, B.S., and Sullivan, C.V. 2006. Evaluation of DNA pooling for the estimation of microsatellite allele frequencies: a case study using striped bass (*Morone saxatilis*). *Genetics*, 173(2):863-875. DOI: 10.1534/genetics.105.053702.
- Smith, J.M. and Whitehurst, D.K. 1990. Tank spawning methodology for the production of striped bass. Pages 73-77 in R.M. Harrell, J.H. Kerby and R.V. Minton (Eds.) *Culture and propagation of striped bass and its hybrids*. Striped Bass Committee, Southern Division, American Fisheries Society, Bethesda, Maryland, USA.
- Stadhouders, L.E., Verbrugge, S.A., Smith, J.A., Gabriel, B.M., Hammersen, T.D., Kolijn, D., Vogel, I.S., Mohamed, A.D., de Wit, G.M., Offringa, C., Hoogaars, W.M., Gehlert, S., and Jaspers, R.T. 2020. Myotube hypertrophy is associated with cancer-like metabolic reprogramming and limited by PHGDH. DOI: 10.1101/2020.12.01.403949.
- Sullivan, C.V., Hiramatsu, N., Kennedy, A.M., Clark, R.W., Weber, G.M., Matsubara, T., Hara, A., 2003. Induced maturation and spawning: opportunities and applications for research

- on oogenesis. *Fish Physiology and Biochemistry*, 28(1-4), 481-486. DOI: 10.1023/B:FISH.0000030635.92568.0a.
- Sullivan, C. V., Chapman, R. W., Reading, B. J., & Anderson, P. E. (2015). Transcriptomics of mRNA and egg quality in farmed fish: some recent developments and future directions. *General and comparative endocrinology*, 221, 23-30. DOI: 10.1016/j.ygcen.2015.02.012.
- Tanaka, K. (2009). The proteasome: Overview of structure and functions. *Proceedings of The Japan Academy, Series B*, 85(1), 12-36. DOI: 10.2183/pjab.85.12.
- Tierbach, A., Groh, K.J., Schönenberger, R., Schirmer, K., Suter, M.J. (2018) Glutathione S-transferase protein expression in different life stages of zebrafish (*Danio rerio*). *Toxicological sciences : an official journal of the Society of Toxicology*. 162(2):702-712. DOI: 10.1093/toxsci/kfx293.
- Valle-Tenney, R., Rebolledo, D., Acuña, M., & Brandan, E. (2020). HIF-hypoxia signaling in skeletal muscle physiology and fibrosis. *Journal Of Cell Communication And Signaling*, 14(2), 147-158. DOI: 10.1007/s12079-020-00553-8.
- Van de Vyver, M., Engelbrecht, L., Smith, C. and Myburgh, K.H. (2016). Neutrophil and monocyte responses to downhill running: Intracellular contents of MPO, IL-6, IL-10, pstat3, and SOCS 3. *Scandinavian journal of medicine & science in sports*, 26(6), pp.638-647. DOI: 10.1111/sms.12497.
- Wang, C., Zhang, B., Ratliff, A., Arlington, J., Chen, J., & Xiong, Y. et al. (2019). Methyltransferase-like 21e inhibits 26S proteasome activity to facilitate hypertrophy of type IIb myofibers. *The FASEB Journal*, 33(8), 9672-9684. DOI: 10.1096/fj.201900582r.

Wheaton, W.W. and Chandel, N.S., 2011. Hypoxia. 2. Hypoxia regulates cellular metabolism.

American Journal of Physiology-Cell Physiology, 300(3), pp.C385-C393. DOI:

10.1152/ajpcell.00485.2010.

Woods III L.C., Li Y., Ding Y., Liu J., Reading B.J., Fuller S.A., and Song J. 2018 DNA

methylation profiles correlated to striped bass sperm fertility. BMC Genomics 19:244.

DOI: 10.1186/s12864-018-4548-6.

Xiang, C., Zhang, Y., Chen, Q., Sun, A., Peng, Y., Zhang, G., Zhou, D., Xie, Y., Hou, X., Zheng,

F. and Wang, F. (2021). Increased glycolysis in skeletal muscle coordinates with adipose

tissue in systemic metabolic homeostasis. Journal of Cellular and Molecular Medicine,

25(16), pp.7840-7854. DOI: 10.1111/jcmm.16698.

Zhou, J., Zhang, Q., Henriquez, J., Crawford, R., & Kaminski, N. (2018). Lymphocyte-Specific

Protein Tyrosine Kinase (LCK) is Involved in the Aryl Hydrocarbon Receptor-Mediated

Impairment of Immunoglobulin Secretion in Human Primary B Cells. Toxicological

Sciences, 165(2), 322-334. DOI: 10.1093/toxsci/kfy133.

Table 3.1. Descriptions of comparisons made between striped bass (SB, n=72) offspring. Each comparison is generally referred to as the name listed in the left-most column. The sub-population size “(n=#)” of each group for a given comparison, referred to as “classes” in machine learning (ML) analyses, are also provided along with a brief description of each group.

Comparison	Groups/Classes	Description
Growth *	Superior (n=36) Inferior (n=36)	larger SB of the 72 sampled smaller SB of the 72 sampled
Market Size *	Market Size (n=12) Superior Other (n=24) Inferior Other (n=24) Under Size (n=12)	reached/exceeded 1.36 kg at final sample Superior Growth SB that did not reach market size Inferior Growth SB that were not “Under Size” smallest SB at final sample
Dam	Dam A (n=55) Dam B (n=17)	offspring born to Dam A offspring born to Dam B
Sire	1-L (n=7) 2-L (n=5) 3-L (n=9) 4-S (n=10) 5-S (n=2) 6-S (n=22) 7-L (n=1) 8-L (n=3) 9-L (n=2) 10-S (n=6) 11-S (n=3) 12-S (n=2)	offspring born to Sire 1-L offspring born to Sire 2-L offspring born to Sire 3-L offspring born to Sire 4-S offspring born to Sire 5-S offspring born to Sire 6-S offspring born to Sire 7-L offspring born to Sire 8-L offspring born to Sire 9-L offspring born to Sire 10-S offspring born to Sire 11-S offspring born to Sire 12-S
Sire Size **	Large (n=27) Small (n=45)	offspring born to “Large” sires (i.e., sires with “-L” following ID number as listed above) offspring born to “Small” sires (i.e., sires with “-S” following ID number as listed above)

*The sample population (n=72 of 173) fell along a bimodal distribution by weight and the larger SB (n=36, mean \pm standard deviation, SD, weight was 1.32 ± 0.18 kg) significantly differed from the smaller SB (n=36, 0.85 ± 0.11 kg; Student’s *t*-Test, $p < 0.0001$), thus these fish were designated into groups of Superior and Inferior growers, respectively. Twelve of the Superior SB had reached or exceeded the market size for these fish (1.36 kg, or 3.0 lbs) at the time of final sampling and additional comparisons between these largest, “Market Size” fish, the smallest, “Under Size” fish, and the remaining intermediate fish in the Superior and Inferior Other groups were made for deeper insight into underlying processes of the extremes of the sample population.

** The females Dam A and Dam B were both crossed with six males, three from each size group “Large” and “Small.” The mean weight and total length (TL) of the Large sires was 2.76 ± 0.15 kg and 570.67 ± 7.84 mm, respectively, and these values significantly differed from those of the Small sires ($p < 0.0001$, mean weight was 1.75 ± 0.15 kg and TL was 490.00 ± 21.14 mm).

Table 3.2. Outcomes of negative control machine learning (ML) analyses of striped bass (SB) gene expression data examined through five comparisons of growth performance (Growth, Market Size) and parentage (Dam, Sire, Sire Size). Each negative control was conducted by processing the optimal dataset* with randomized labels (group identifiers for a given comparison) associated with expression data ten separate times with each of the eight cross-validated algorithms (four algorithms x two cross-validated strategies). True learning by algorithms is said to occur if the mean percent correct classification of each negative control run is approximately what can be yielded through random chance. The number of attributes included in the optimal dataset and classes (groups) for each comparison are indicated followed by the percent correct classification that can be predicted under the assumption of random probability based on the number of classes (possible outcomes). The grand mean \pm standard deviation (SD) percent correct classification of all cross-validated ML algorithms on randomized optimal datasets for a given comparison is reported in the right-most column.

Comparison	Optimal Attributes	Classes	Correct Classification Assuming Random Probability (%)	Actual Correct Classification (Mean \pm SD %)
Growth	300	2	50.00	49.17 \pm 7.84
Market Size	284	4	25.00	29.53 \pm 7.01
Dam	150	2	50.00	67.01 \pm 8.83
Sire	35	12	8.00	18.82 \pm 6.15
Sire Size	75	2	50.00	56.22 \pm 6.69

*To establish the optimal dataset for a given comparison, values based on the information gained by the algorithm from the consideration of each attribute (i.e., gene transcript) are calculated and attributes are subsequently assigned a rank from low entropy (high information) to high entropy (little or no information). The lower-ranking of the ranked attributes (i.e., those assigned an information gain value greater than 0.00) are then removed (approximately 50 at a time) from the dataset used for training and testing by all ML algorithms. Algorithm performance measured as percent correct classification is then plotted against the number of input top-ranked attributes to determine points of model performance improvement or deterioration with the iterative exclusion of attributes. The number of top-ranked attributes included in input and yielding optimal or close-to optimal classification performance across multiple cross-validated algorithms is designated as the optimal dataset.

Table 3.3. Summary of striped bass (SB) growth and feeding data collected over the course of one year starting at five months of age. SB were raised in triplicate indoor recirculating aquaculture system (RAS) tanks (908 L/tank from four to eight months of age and 2006 L/tank thereafter). The rightmost column indicates the age of SB at each sampling point followed by the total population size (N) and mean \pm standard deviation (SD) weight (g) and total length (TL, mm). The amount of feed (g) offered (sum offered to all tanks over a given duration) for a given period between sampling and the mean \pm SD feed conversion ratio (FCR) are also provided. These values specific to each replicate tank are provided in italics. Differentiating letters assigned to weight and/or TL values for a given sampling point indicate statistical differences between replicate tanks were identified via one-way ANOVA and Tukey's HSD ($\alpha=0.05$). The absence of letters indicates no significant differences were identified for a given measurement and timepoint.

Age					
<i>Replicate</i>	N	Weight (g)	TL (mm)	Feed (g)	FCR
5 months	194	53.67 \pm 1.61	166.95 \pm 10.63	-	-
<i>1</i>	<i>64</i>	<i>55.41 \pm 13.045</i>	<i>167.73 \pm 12.068</i>	-	-
<i>2</i>	<i>66</i>	<i>52.23 \pm 10.34</i>	<i>165.67 \pm 9.80</i>	-	-
<i>3</i>	<i>64</i>	<i>53.38 \pm 10.046</i>	<i>167.50 \pm 9.94</i>	-	-
8 months	190	181.03 \pm 8.49	241.80 \pm 16.14	30211	1.26 \pm 0.017
<i>1</i>	<i>63</i>	<i>175.40 \pm 47.41</i>	<i>240.095 \pm 17.065^A</i>	<i>9314</i>	<i>1.24</i>
<i>2</i>	<i>65</i>	<i>176.92 \pm 37.58</i>	<i>238.59 \pm 12.95^A</i>	<i>10169</i>	<i>1.26</i>
<i>3</i>	<i>62</i>	<i>190.81 \pm 43.77</i>	<i>246.89 \pm 17.17^B</i>	<i>10728</i>	<i>1.28</i>
11 months	177	385.00 \pm 31.24	294.19 \pm 19.023	48103	1.44 \pm 0.037
<i>1</i>	<i>58</i>	<i>373.28 \pm 92.69^A</i>	<i>290.81 \pm 19.85^A</i>	<i>15645</i>	<i>1.48</i>
<i>2</i>	<i>64</i>	<i>361.25 \pm 75.81^A</i>	<i>289.77 \pm 15.80^A</i>	<i>16302</i>	<i>1.40</i>
<i>3</i>	<i>55</i>	<i>420.36 \pm 89.77^B</i>	<i>302.89 \pm 18.97^B</i>	<i>16156</i>	<i>1.43</i>
14 months	177	762.37 \pm 47.78	380.83 \pm 23.19	105207	1.58 \pm 0.038
<i>1</i>	<i>58</i>	<i>765.52 \pm 166.55^{AB}</i>	<i>383.069 \pm 24.83</i>	<i>34959</i>	<i>1.54</i>
<i>2</i>	<i>64</i>	<i>713.13 \pm 131.85^A</i>	<i>375.19 \pm 20.58</i>	<i>36191</i>	<i>1.61</i>
<i>3</i>	<i>55</i>	<i>808.55 \pm 159.56^B</i>	<i>385.018 \pm 23.39</i>	<i>34057</i>	<i>1.60</i>
18 months	173	1109.00 \pm 33.42	428.27 \pm 25.37	121767	2.14 \pm 0.087
<i>1</i>	<i>57</i>	<i>1081.40 \pm 249.49</i>	<i>426.31 \pm 26.86</i>	<i>37378</i>	<i>2.17</i>
<i>2</i>	<i>62</i>	<i>1100.32 \pm 202.92</i>	<i>426.73 \pm 22.67</i>	<i>46032</i>	<i>2.04</i>
<i>3</i>	<i>54</i>	<i>1145.93 \pm 244.98</i>	<i>432.11 \pm 26.71</i>	<i>38357</i>	<i>2.20</i>

Table 3.4. The mean \pm standard deviation (SD) weight (kg) and total length (TL, mm) of the sub-population of striped bass (SB) sacrificed for analysis (n=72 of 173). An equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (1.36 kg, or 3.0 lbs; n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) in subsequent analyses and the average weight and TL of those sub-groups are provided here. A Student’s *t*-Test was used to compare weight and length between Superior and Inferior groups for the “Growth” comparison described in the text and the one-way ANOVA and Tukey’s HSD post hoc test were used to compare the four groups in the “Market Size” comparison described in the text. The p-value for all comparisons of weight and TL was $p < 0.0001$ as indicated in the table below. Differentiating letters are used to indicate differences of weight and/or total length identified between groups in the Growth comparison and groups in the Market Size comparison.

Group (n)	Weight (kg)	TL (mm)	p-value
Superior (36)	1.32 \pm 0.18 ^A	449.00 \pm 16.54 ^A	<0.0001
Inferior (36)	0.85 \pm 0.12 ^B	399.58 \pm 15.96 ^B	
Market Size (12)	1.52 \pm 0.13 ^A	465.67 \pm 13.27 ^A	<0.0001
Superior Other (24)	1.21 \pm 0.090 ^B	440.67 \pm 10.72 ^B	
Inferior Other (24)	0.91 \pm 0.076 ^C	408.083 \pm 11.27 ^C	
Under Size (12)	0.72 \pm 0.060 ^D	382.58 \pm 8.50 ^D	

Table 3.5. Weight (kg) and total length (TL, mm) of the striped bass (SB) offspring sacrificed for analysis (n=72) and grouped according to parentage: SB Dam, Sire Size, and Dam x Sire Size. Briefly, two female SB (dams) were crossed with twelve male SB (sires) that were either “Large” or “Small” in size (“Sire Size” comparison), whereby six males total (three from either size group) were crossed with a single female to produce twelve half-sibling families. The number of SB produced from either Dam, Large or Small sires, or specific Dam x Sire Size is listed in the table below (“Offspring”) followed by the mean \pm standard deviation (SD) weight and TL of SB offspring belonging to a given size group. Weight and TL metrics were compared between groups via Student’s *t*-Test (Dam, Sire Size) or one-way ANOVA and Tukey’s HSD post hoc test (Dam x Sire Size), p-values are listed below for each comparison and metric. Differentiating letters are used to indicate differences between weight of offspring in the Dam x Sire Size comparison.

Comparison	Offspring	Offspring Weight (kg)	Offspring Total Length (mm)
<i>Dam</i>			
Dam A	55	1.06 \pm 0.28	421.95 \pm 30.32
Dam B	17	1.16 \pm 0.27	431.88 \pm 26.84
		<i>p</i> =0.1689	<i>p</i> =0.2298
<i>Sire Size*</i>			
Large Sires	27	1.18 \pm 0.31	433.63 \pm 33.55
Small Sires	43	1.02 \pm 0.25	418.69 \pm 25.86
		<i>p</i> =0.0212	<i>p</i> =0.0376
<i>Dam x Sire Size**</i>			
Dam A x Large Sires	21	1.19 \pm 0.34 ^A	434.81 \pm 35.70
Dam A x Small Sires	34	0.97 \pm 0.20 ^B	414.00 \pm 23.70
Dam B x Large Sires	6	1.13 \pm 0.23 ^{AB}	429.50 \pm 27.00
Dam B x Small Sires	11	1.18 \pm 0.30 ^{AB}	433.18 \pm 27.99
		<i>p</i> =0.0177	<i>p</i> \geq 0.0511

*Large SB sires (2.76 \pm 0.15 kg weight, 570.67 \pm 7.84 mm TL) significantly differed from the Small SB sires (1.75 \pm 0.15 kg weight, 490.00 \pm 21.14 mm TL) by weight and TL (*p*<0.0001).

** The Large and Small sires crossed with Dam A or Dam B did not significantly differ within size groups, however, Large sires crossed with Dam A (Dam A: 5.68 kg, 722 mm x Large Sires: 2.73 \pm 0.20 kg weight, 567.67 \pm 6.66 mm TL) and Dam B (Dam B: 6.15 kg, 703 mm x Large Sires: 2.79 \pm 0.11 kg weight, 573.67 \pm 9.074 mm TL) did significantly differ in weight (*p* \leq 0.0003) and TL (*p* \leq 0.0032) from Small sires crossed with Dam A (Dam A x Small Sires: 1.69 \pm 0.20 kg weight, 491.33 \pm 17.62 mm TL) and Dam B (Dam B x Small Sires: 1.81 \pm 0.070 kg weight, 488.67 \pm 28.31 mm TL).

Table 3.6. The number of gene transcripts of the 32,018 measured that were identified as important to the comparisons of striped bass (SB) listed below based upon information gain (“Entropy”) determined through the machine learning (ML) approach or significant difference determined through traditional statistical approaches (false discovery rate p-value, “ $p \leq 0.05$ ”). The gene transcripts measured map back to 22,746 unique genes; the number of unique genes each set of transcripts map back to is provided in parentheses immediately following each value. The numbers provided in the “Optimal” row for each comparison are those gene transcripts and unique genes that were considered the optimal dataset* for each comparison. The number of transcripts and unique genes that were identified as important based upon information gain and significantly differed are listed in the “Shared” column, the values in this column associated with the “Optimal” row is this number of transcripts and genes that are also within the optimal dataset*. The final column, “Total” provides the total number of unique transcripts and genes identified as important to a given comparison based upon information gain and statistical tests.

Comparison	Entropy	FDR $p \leq 0.05$	Shared	Total
Growth	965 (946)	401 (360)	91 (89)	1275 (1211)
<i>Optimal</i>	300 (298)		32 (32)	
Market Size	284 (282)	621 (549)	43 (43)	862 (785)
<i>Optimal</i>	284 (282)		43 (43)	
Dam	257 (253)	102 (96)	31 (31)	328 (317)
<i>Optimal</i>	150 (147)		21 (21)	
Sire	35 (35)	1177 (976)	22 (22)	1190 (987)
<i>Optimal</i>	35 (35)		22 (22)	
Sire Size	378 (367)	150 (131)	28 (26)	500 (471)
<i>Optimal</i>	75 (72)		16 (15)	

*To establish the optimal dataset for a given comparison, values based on the information gained by the algorithm from the consideration of each attribute (i.e., gene transcript) are calculated and attributes are subsequently assigned a rank from low entropy (high information) to high entropy (little or no information). The lower-ranking of the ranked attributes (i.e., those assigned an information gain value greater than 0.00) are then removed (approximately 50 at a time) from the dataset used for training and testing by all ML algorithms. Algorithm performance measured as percent correct classification is then plotted against the number of input top-ranked attributes to determine points of model performance improvement or deterioration with the iterative exclusion of attributes. The number of top-ranked attributes included in input and yielding optimal or close-to optimal classification performance across multiple cross-validated algorithms is designated as the optimal dataset.

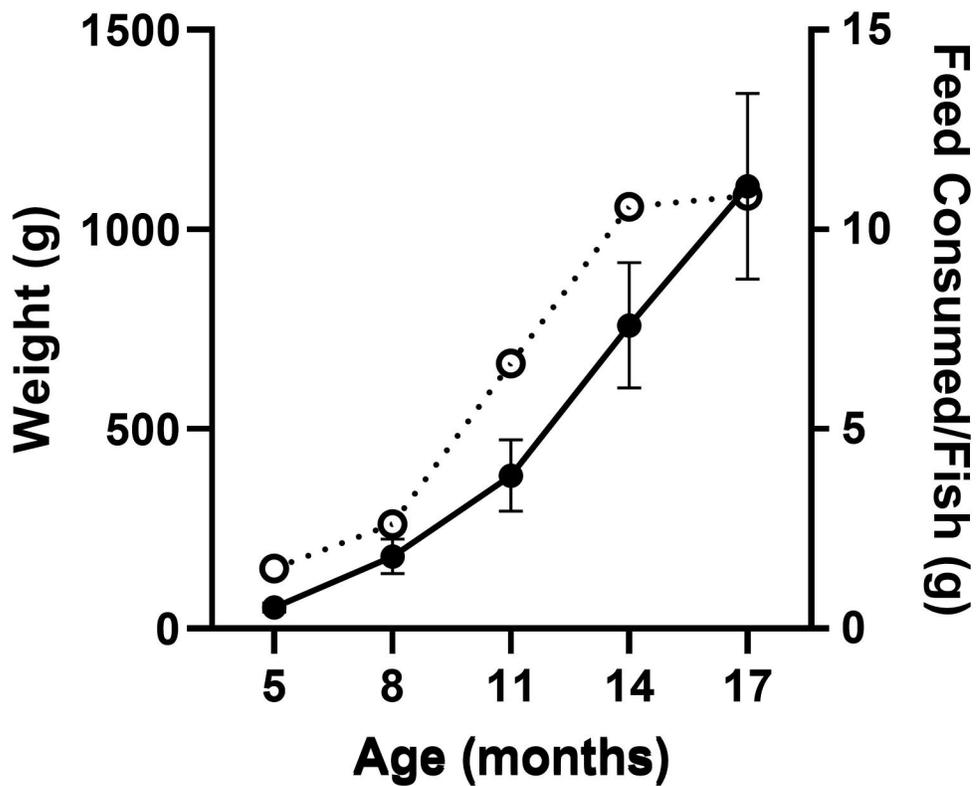


Figure 3.1. Growth and feed consumption of a population of striped bass (SB; Initial N=194, Final N=173) raised in triplicate recirculating aquaculture system (RAS) tanks at Grinnell's Animal Health Laboratory (North Carolina State University, Raleigh, NC, USA) for a year from five to eighteen months of age. Fish were weighed every three months and feed was administered *ad libitum* for the entirety of the study to monitor consumption and enable the calculation of amount of feed consumed per number of fish over time and while accounting for mortality. The average weight in grams (g) of the population (left y-axis) is represented by the black solid line and bars represent the standard deviation. The average amount of feed consumed (g) per fish is represented by the dotted line and open circles.

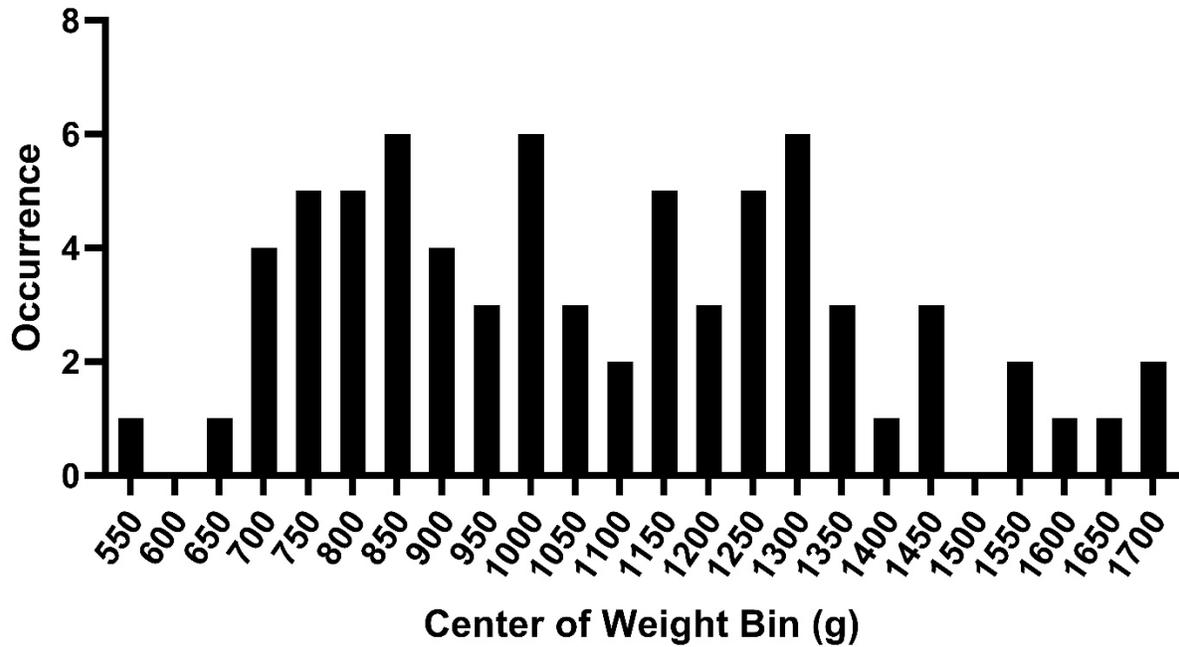


Figure 3.2. Frequency distribution of the weights (g) of striped bass (SB) sacrificed at eighteen months of age (n=72 of 173) for muscle histology and gene expression analysis. The x-axis indicates the center value of each bin (e.g., the four fish in the 700 g bin weighed between 675 and 725 g). Fish were reared in triplicate recirculating aquaculture system (RAS) tanks at Grinnell's Animal Health Laboratory (North Carolina State University, Raleigh, NC, USA) for a year from five to eighteen months of age.

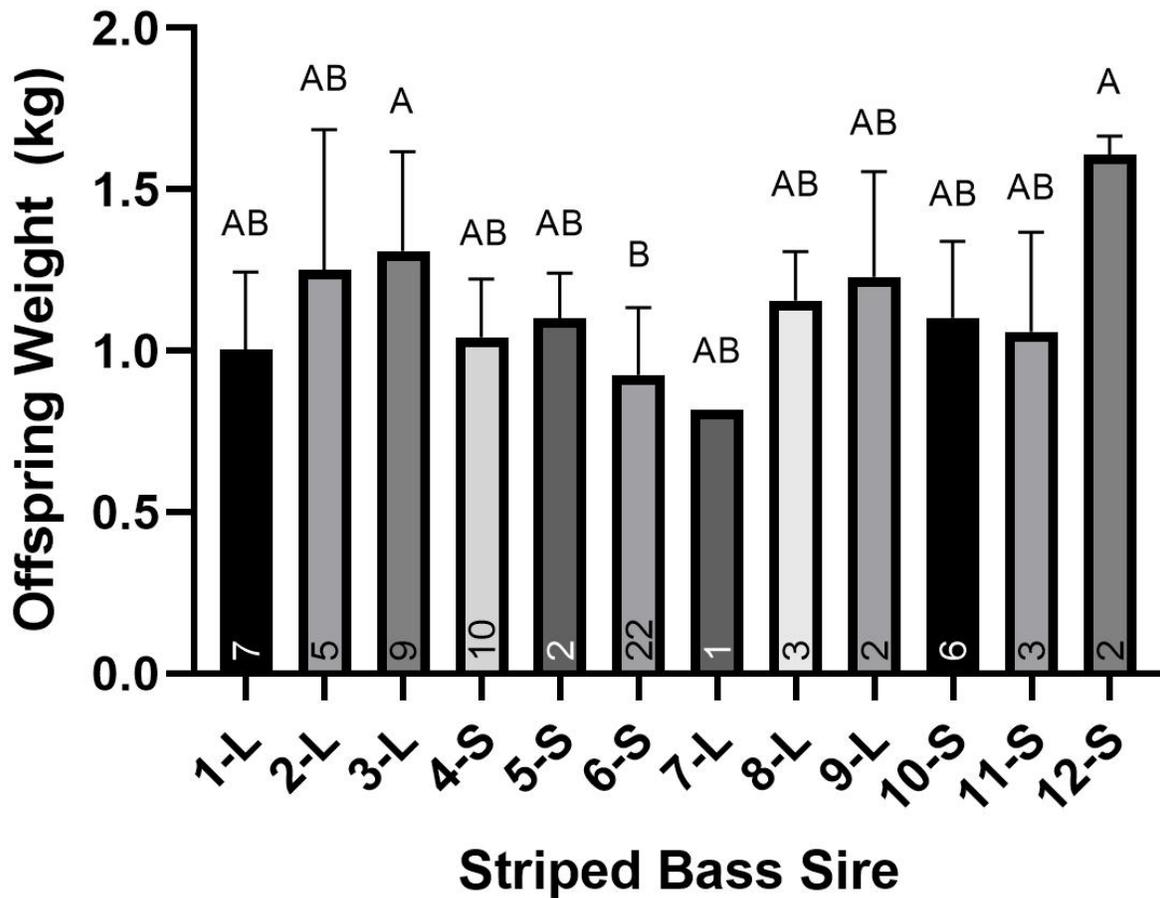


Figure 3.3. Mean weight (kg) of striped bass (SB) offspring (y-axis) produced from one of twelve SB sires (x-axis). Sires numbered 1–6 (“#-L/S”) were crossed with the same SB female (“Dam A”) and are therefore of half sibling families. Similarly, sires numbered 7–12 were crossed with a different SB female (“Dam B”). Sires belonged to two different groups, Large, indicated by “-L”, or Small, indicated by “-S”, that significantly differed in weight and total length, TL (Large: 2.76 ± 0.15 kg weight, 570.67 ± 7.84 mm TL; Small: 1.75 ± 0.15 kg weight, 490.00 ± 21.14 mm TL; $p < 0.0001$). Vertical numbers indicate the number of offspring produced from each sire of the total sacrificed for this study ($n=72$). Standard deviation bars are present if more than one offspring was produced from a given sire. Differentiating letters indicate significant differences between mean weight of offspring born to individual sires as determined by one-way ANOVA and Tukey’s HSD post-hoc test ($p \leq 0.0201$).

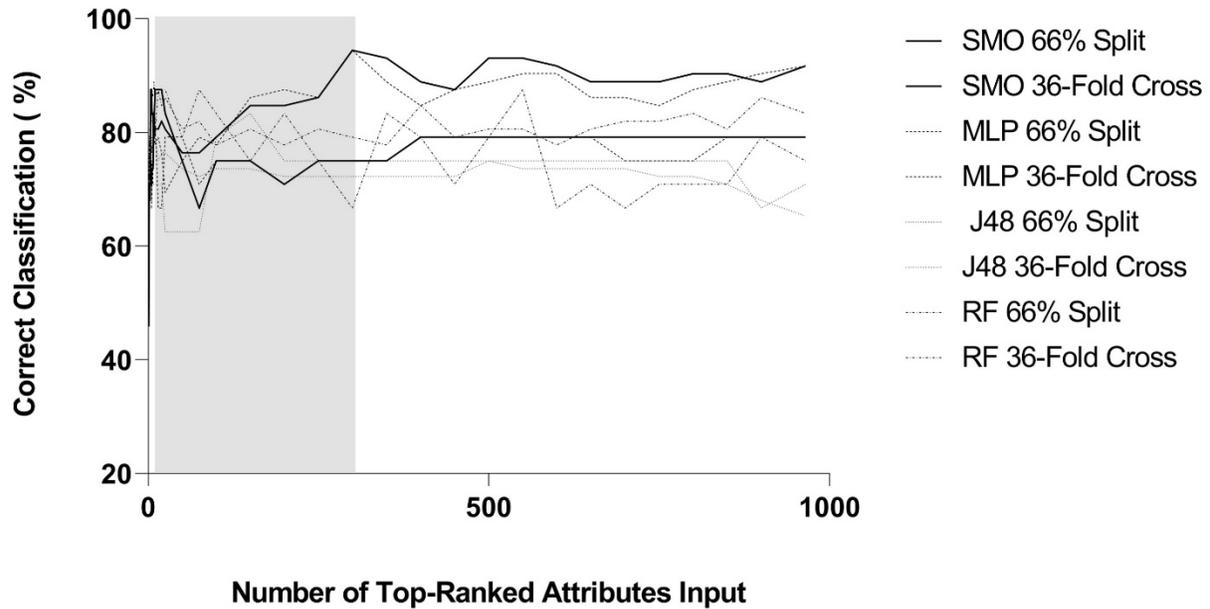


Figure 3.4. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) striped bass (SB) into classes (groups) for the Growth Comparison (*see*: **Table 3.1**) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (K = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (965 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison, which were determined to be 10 and 300 attributes, respectively.

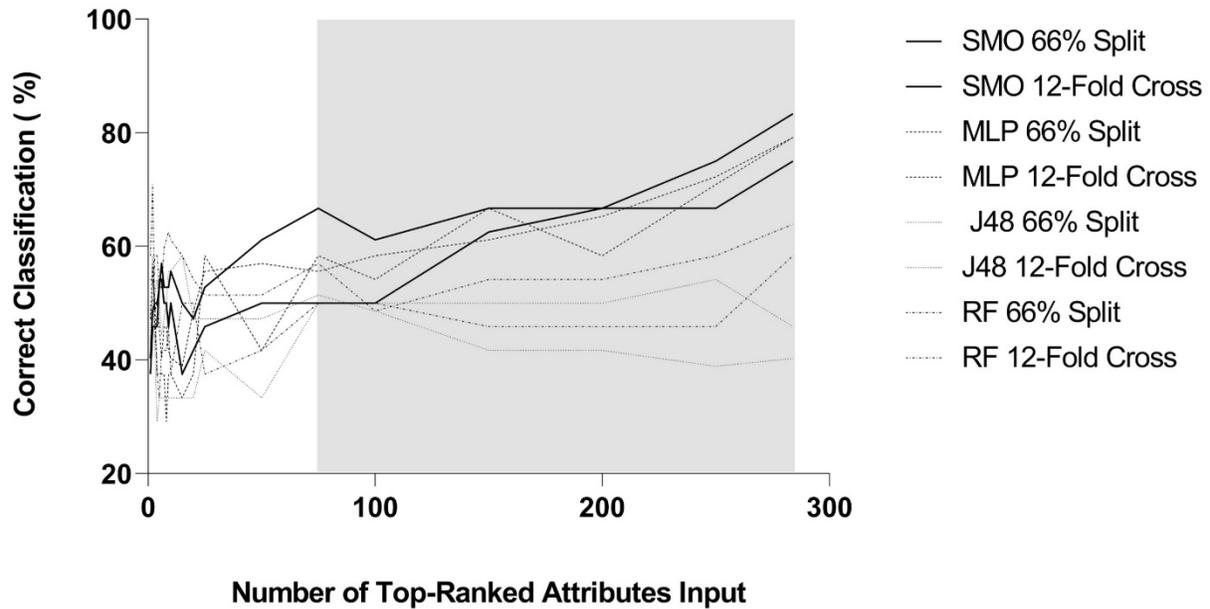


Figure 3.5. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (%) striped bass (SB) into classes (groups) for the Market Size Comparison (*see*: **Table 3.1**) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (K = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (284 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison, which were determined to be 75 and all 284 attributes, respectively.

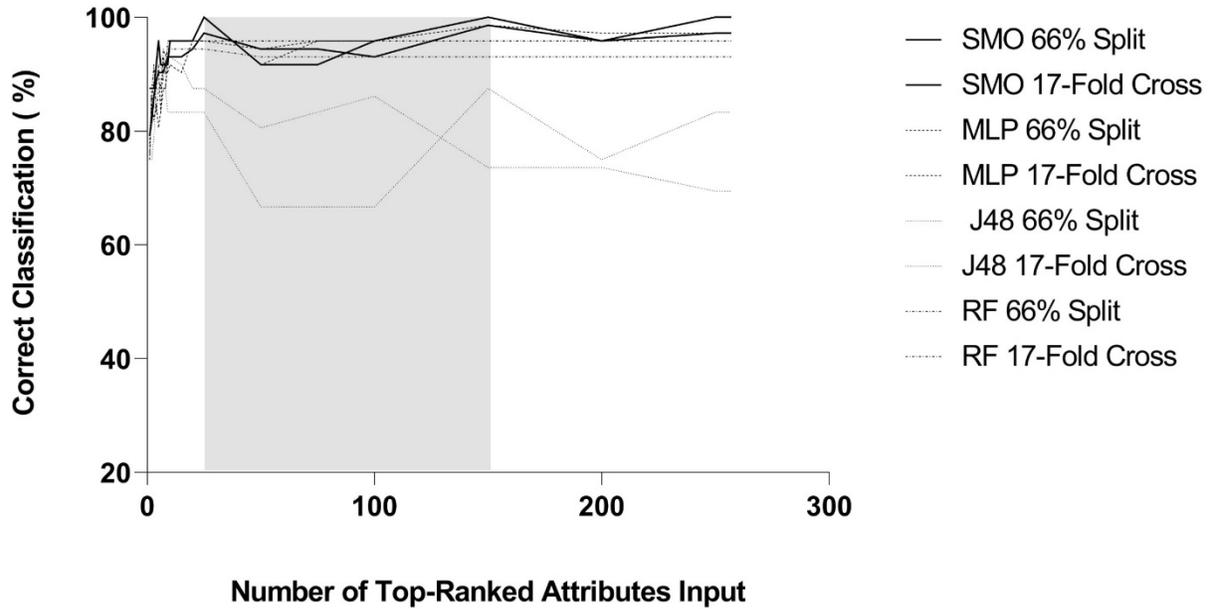


Figure 3.6. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (%) striped bass (SB) into classes (groups) for the Dam Comparison (see: **Table 3.1**) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (K = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (257 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison, which were determined to be 25 and 150 attributes, respectively.

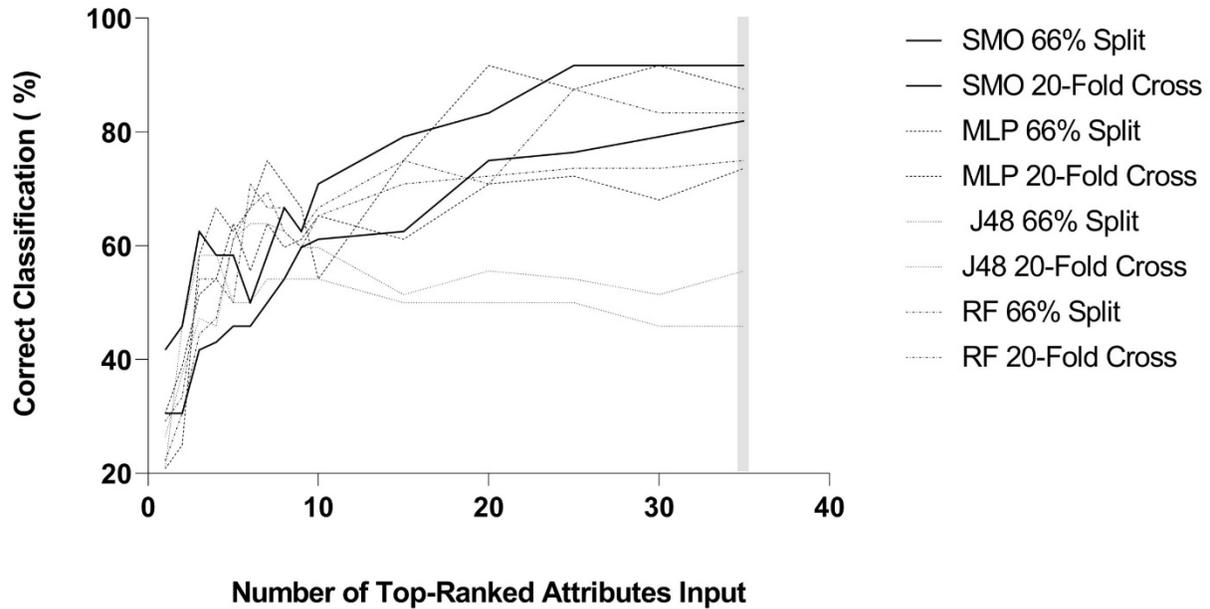


Figure 3.7. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (% , y-axis) striped bass (SB) into classes (groups) for the Sire Comparison (see: **Table 3.1**) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (K = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (35) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison was determined to be including all 35 ranked attributes.

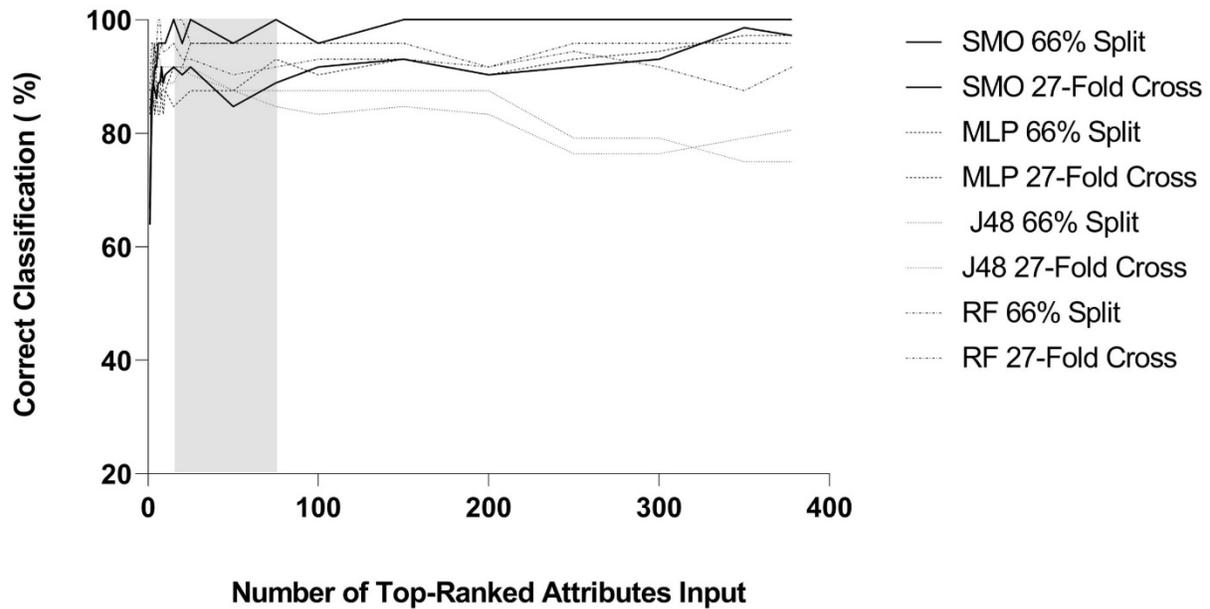
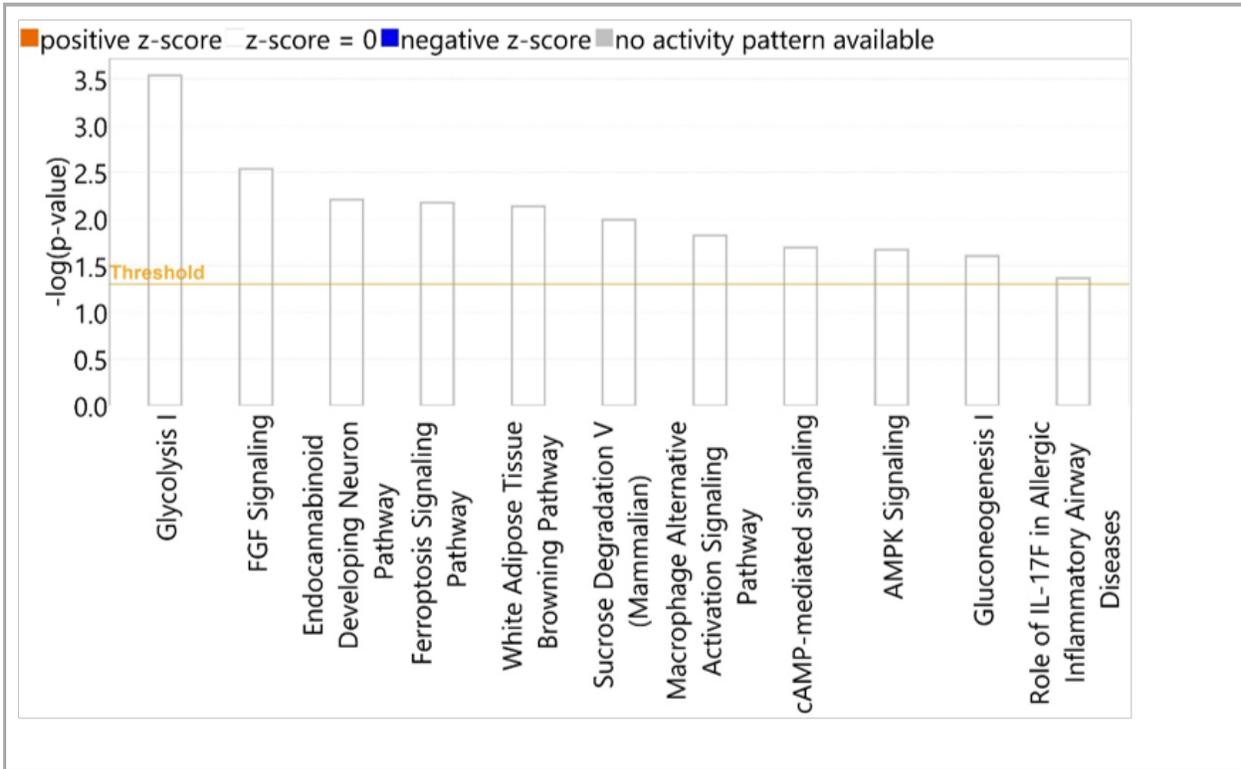


Figure 3.8. The performance of four cross-validated machine learning (ML) algorithms in correctly classifying (%) striped bass (SB) into classes (groups) for the Sire Size Comparison (*see*: **Table 3.1**) based upon expression patterns of gene transcripts (attributes) identified as having an information gain value above 0.0 and thus providing information for learning. Specifically, any information gain value above 0.0 indicates that information is gained by including a given attribute in learning and were considered “top-ranked” among all 32,018 transcripts measured for the given comparison. Greater information gain values indicate more information is gained from a given attribute than those assigned lesser information gain values. The ML algorithms used were sequential minimal optimization (SMO), a support vector machine; multilayer perceptron (MLP) an artificial neural network (ANN); J48, a decision tree; and Random Forest (RF), an ensemble (i.e., combination of models) decision tree. The holdout method (66.0 % split) and the stratified K -fold cross-validation (K = minimum number of samples in any class) were the cross-validation strategies applied with each run of ML algorithm. All top-ranked attributes (378 total) and subsets thereof including fewer and fewer of the ranked attributes until only the highest ranking remained were run through each cross-validated algorithm. The grey box indicates the threshold of underfitting (i.e., too few highly-ranked attributes included thus negatively impacting algorithm performance) and overfitting (i.e., too many highly-ranked attributes included) for this comparison, which were determined to be 15 and 75 attributes, respectively.

Figure 3.9. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) designated as Superior (A) or Inferior (B) based upon growth performance. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those in the Superior group were significantly greater in weight (1.32 ± 0.18 kg/fish) and total length (TL, 449.00 ± 16.54 mm/fish) than those in the Inferior group (0.85 ± 0.12 kg/fish and 399.58 ± 15.96 mm/fish; Student's *t*-Test, $p < 0.0001$). Genes included in analysis (294 unique, analysis-ready molecules from a list of 300 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher's Exact Test ($p = 0.05$ indicated by "Threshold" line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made.

A. Superior Growth SB



B. Inferior Growth SB

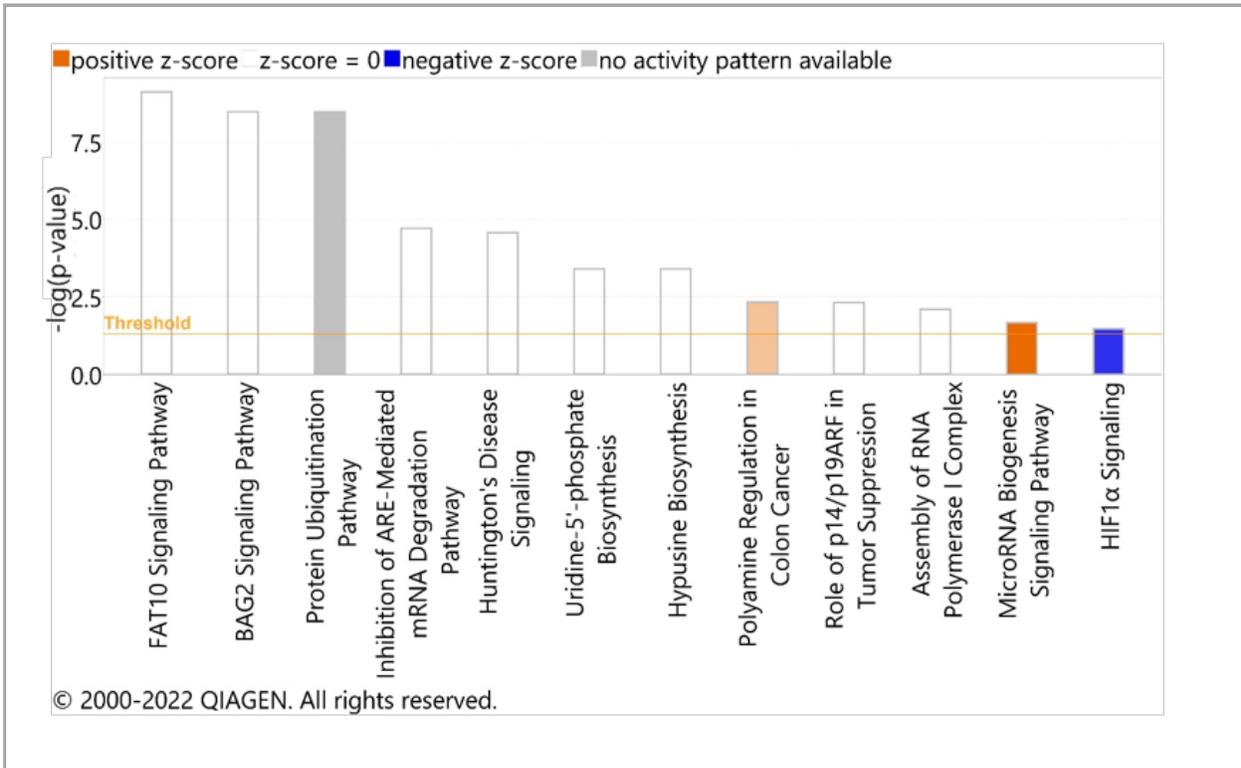
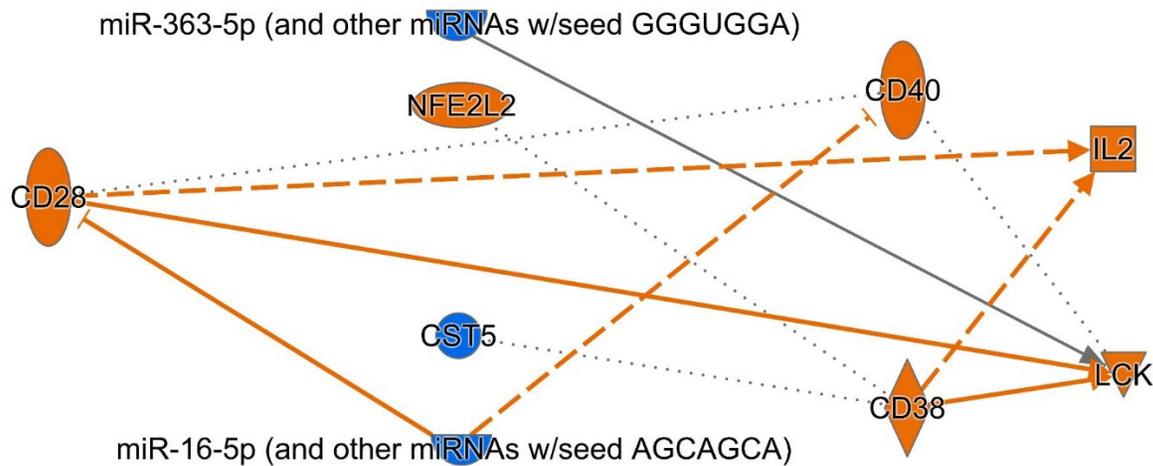


Figure 3.9.



©-2022 QIAGEN. All rights reserved.

Figure 3.10. Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among striped bass (SB) designated as Inferior based upon growth performance. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36) and these groups significantly differed in weight and total length (Student's *t*-Test, $p < 0.0001$). Genes included in analysis (294 unique, analysis-ready molecules from a list of 300 gene transcripts total, 272 of which were up-regulated in Inferior SB) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance through the application of a machine learning workflow. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., HGF), transcription regulators are displayed as horizontal ovals (e.g., MYC), mature microRNAs are displayed as half circles (e.g., miR-16-5p), functions are displayed as octagons (e.g., Cell Survival), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted.

Figure 3.11. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) designated as Under Size (A), Inferior Other (B), Market Size (C), or Superior Other (D) based upon growth to market size (1.36 kg, or 3.0 lbs) of these fish. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) and the remaining SB in each group were referred to as Superior Other and Inferior Other, respectively. SB in all groups significantly differed in weight and total length (one-way ANOVA and Tukey’s HSD post hoc test, $p < 0.0001$). Genes included in analysis (277 unique, analysis-ready molecules from a list of 284 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance to market size through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p = 0.05$ indicated by “Threshold” line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made.

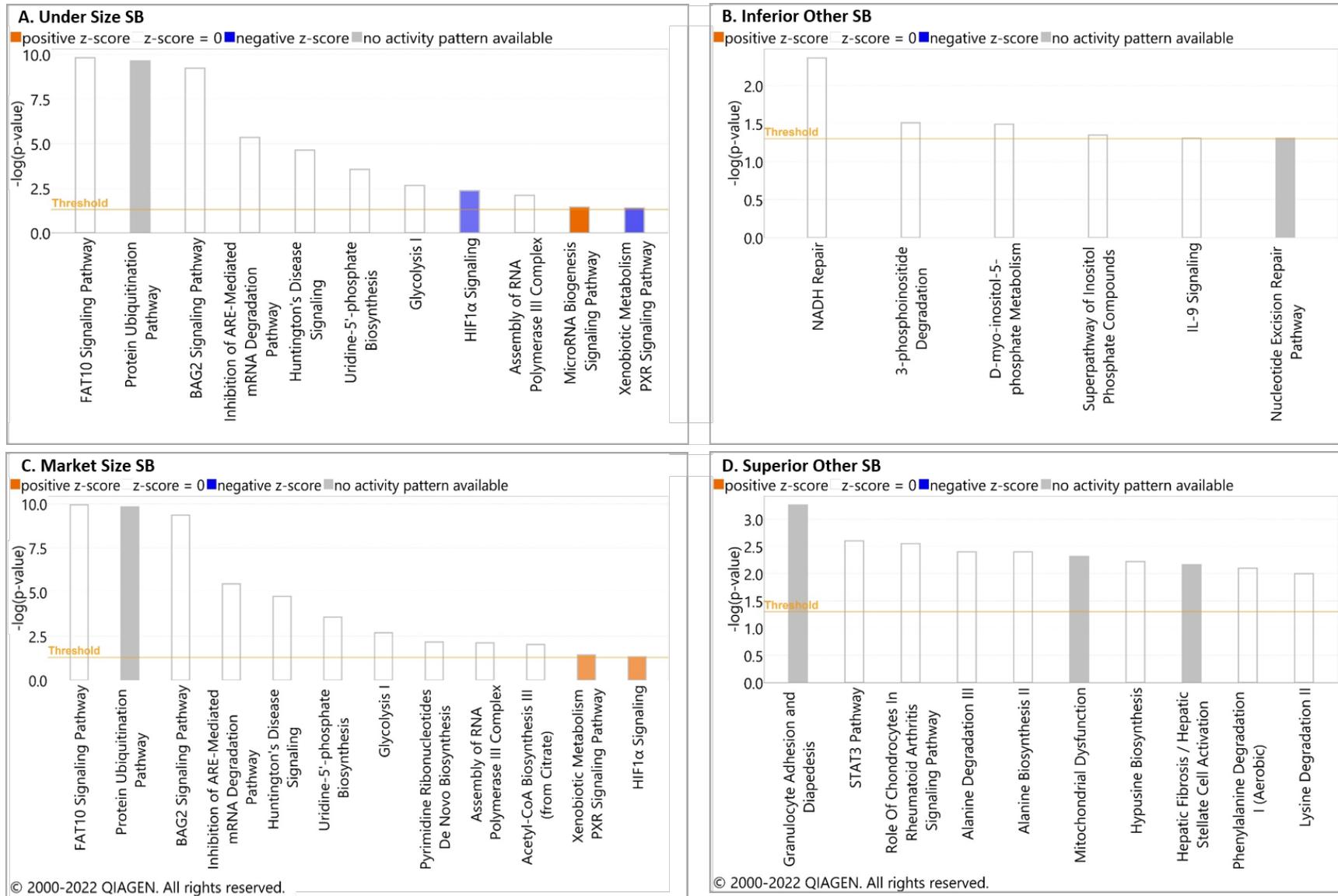
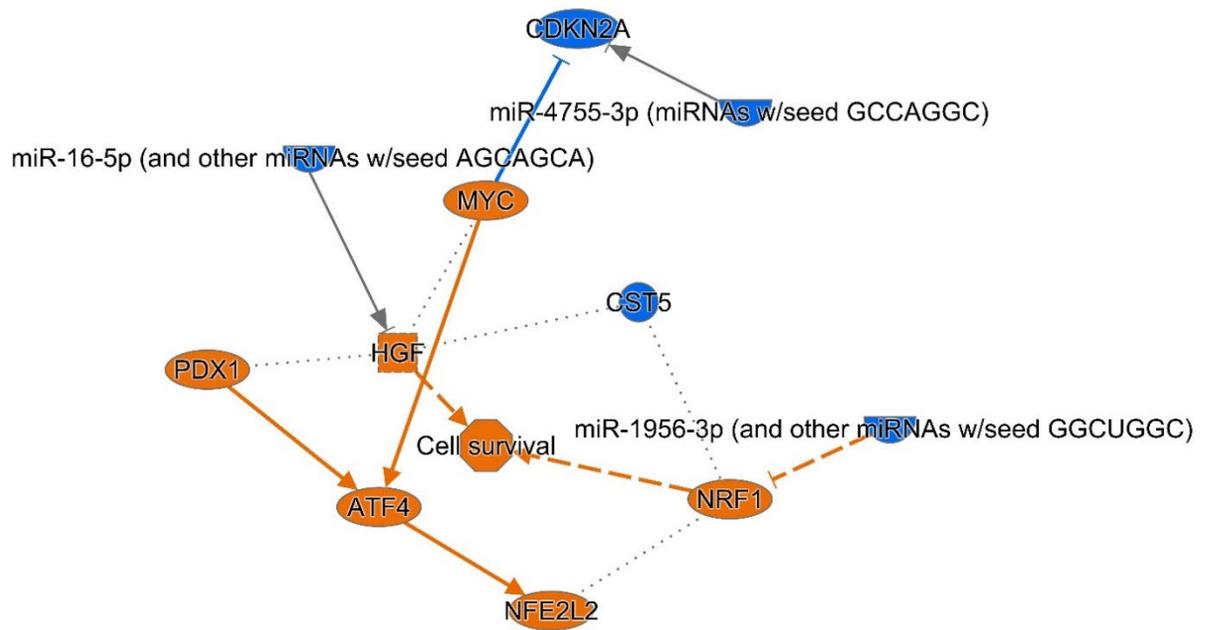


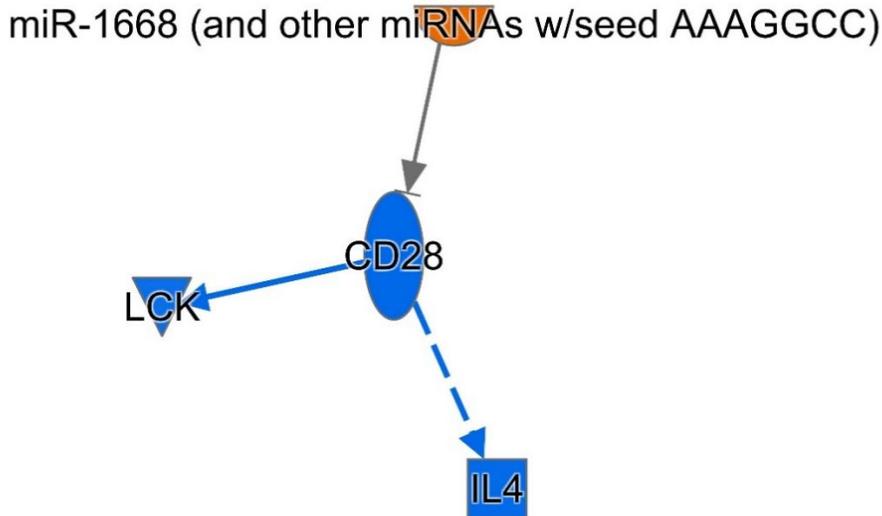
Figure 3.11.

Figure 3.12. Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among striped bass (SB) designated as Under Size. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (1.36 kg, or 3.0 lbs; n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) and the remaining SB in each group were referred to as Superior Other and Inferior Other, respectively. SB in all groups significantly differed in weight and total length (one-way ANOVA and Tukey’s HSD post hoc test, $p < 0.0001$). Genes included in analysis (277 unique, analysis-ready molecules from a list of 284 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance to market size through the application of a machine learning workflow. A total of 194 of the 277 unique genes were identified as up-regulated in Under Size SB. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., HGF), transcription regulators are displayed as horizontal ovals (e.g., MYC), mature microRNAs are displayed as half circles (e.g., miR-16-5p), functions are displayed as octagons (e.g., Cell Survival), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted.



© 2000-2022 QIAGEN. All rights reserved.

Figure 3.12.

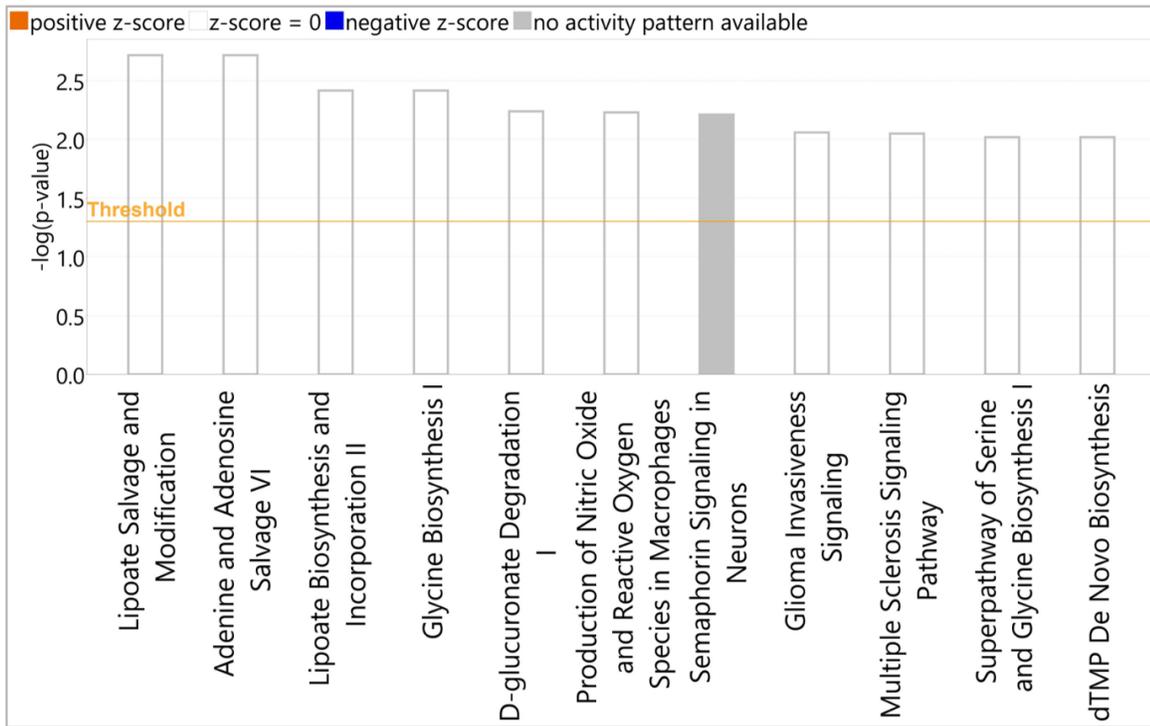


© 2000-2022 QIAGEN. All rights reserved.

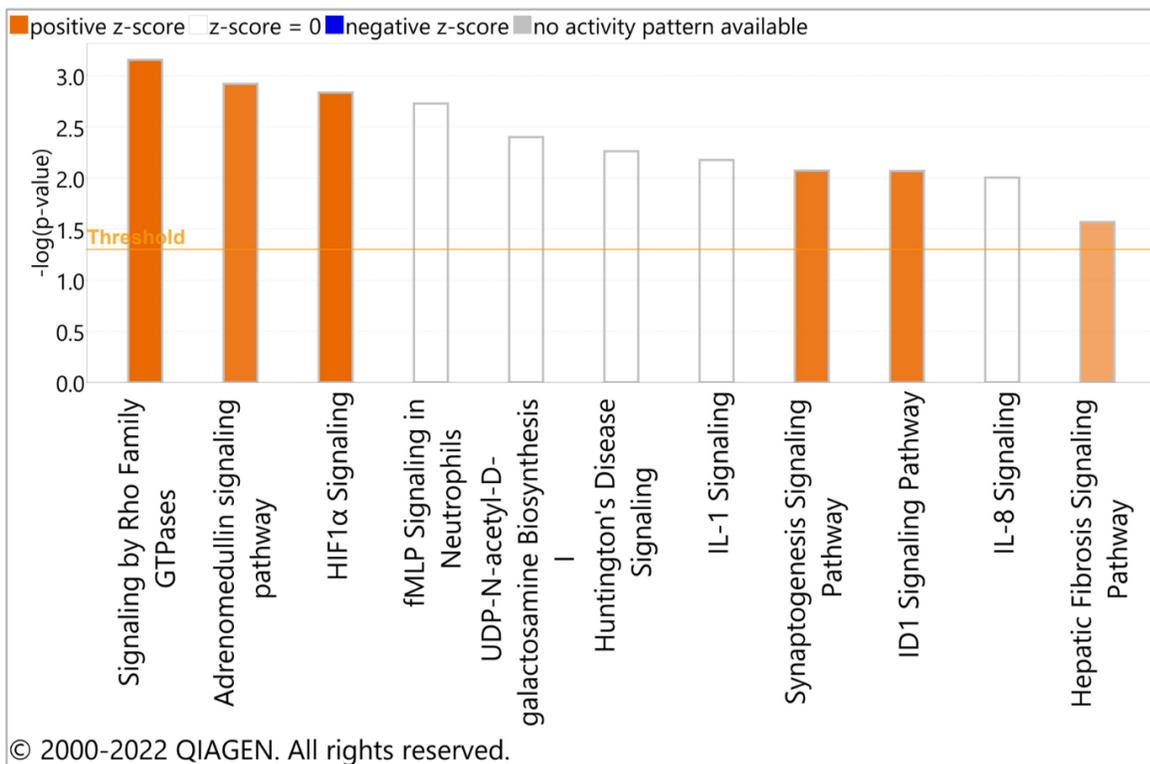
Figure 3.13. Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression among striped bass (SB) designated as Market Size. Briefly, an equal number of SB had been randomly selected for sacrifice from triplicate tanks of an indoor recirculating aquaculture system (RAS) at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth (n=36) and those that exhibited relatively poorer, or Inferior growth (n=36). Those Superior growing fish that reached or exceeded market size (1.36 kg, or 3.0 lbs; n=12 of 36 Superior, “Market Size”) were further compared to the smallest fish of the Inferior group (n=12 of 36 Inferior, “Under Size”) and the remaining SB in each group were referred to as Superior Other and Inferior Other, respectively. SB in all groups significantly differed in weight and total length (one-way ANOVA and Tukey’s HSD post hoc test, $p < 0.0001$). Genes included in analysis (277 unique, analysis-ready molecules from a list of 284 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB growth performance to market size through the application of a machine learning workflow. A total of 49 of the 277 unique genes were identified as up-regulated in Under Size SB. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., HGF), transcription regulators are displayed as horizontal ovals (e.g., MYC), mature microRNAs are displayed as half circles (e.g., miR-16-5p), functions are displayed as octagons (e.g., Cell Survival), and elements classified as “other” are displayed as circles (e.g., DAP3). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, and grey indicates that an effect is not predicted.

Figure 3.14. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) produced from one of two female SB, Dam A (A) or Dam B (B). Briefly, SB had been produced by crossing two female SB with six different male SB (sires) each, whereby three of the sires each dam was crossed with were designated as “Large” or “Small” as they significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Offspring were reared in triplicate tanks of an indoor recirculating aquaculture system (RAS) and sampled at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth ($n=36$) and those that exhibited relatively poorer, or Inferior growth ($n=36$). This sampled population was split into two equal groups based on size: those that exhibited Superior growth ($n=36$) and those that exhibited relatively poorer, or Inferior growth ($n=36$), and these fish significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Genes included in analysis (139 unique, analysis-ready molecules from a list of 150 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p=0.05$ indicated by “Threshold” line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made.

A. Dam A Offspring

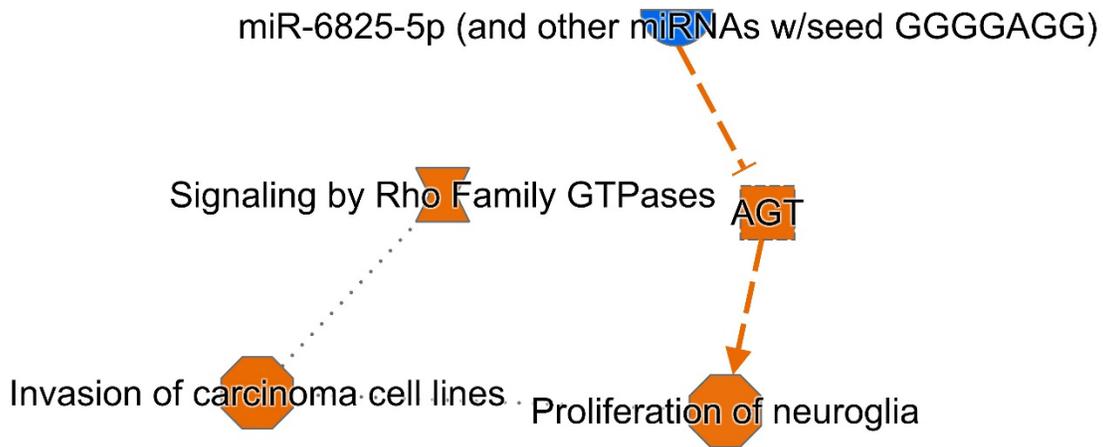


B. Dam B Offspring



© 2000-2022 QIAGEN. All rights reserved.

Figure 3.14.



© 2000-2022 QIAGEN. All rights reserved.

Figure 3.15. Network of upstream regulatory molecules and pathways predicted to underlie observed patterns in gene expression in striped bass (SB) offspring produced by a single female, Dam B, and crossed with six males, three belonging to a “Large” group and three belonging to a “Small” group, which significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Genes included in analysis (139 unique, analysis-ready molecules from a list of 150 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. A total of 94 of these 139 genes were up-regulated in offspring produced from Dam B. Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) software was used to make predictions of upstream regulators based the number of known (Ingenuity® Knowledge Base) targets of a given regulator among the molecules in the input dataset. Node type and relationships between nodes are indicated by shape and line style, respectively, as follows: cytokines are displayed as squares (e.g., AGT), mature microRNAs are displayed as half circles (e.g., miR-6825-5p), canonical pathways are displayed as hourglass hexagons (e.g., Signaling by Rho Family GTPases), and functions are displayed as octagons (e.g., Proliferation of neuroglia). Solid lines indicate a known direct interaction between two elements, dashed lines indicate an indirect interaction, and dotted lines represented a relationship that has been inferred based upon ML-approaches applied in IPA. Solid arrows represent directional activation, causation, or expression, perpendicular intersecting lines indicate inhibition, or ubiquitination. Lines without endpoint markers indicate interactions (e.g., chemical-chemical, protein-protein) and/or correlation. Predicted activity is indicated by orange for activation, blue for inhibition, yellow for findings that are inconsistent with the up- or down-regulation of a downstream molecule, and grey indicates that an effect is not predicted.

SB Sire Comparison

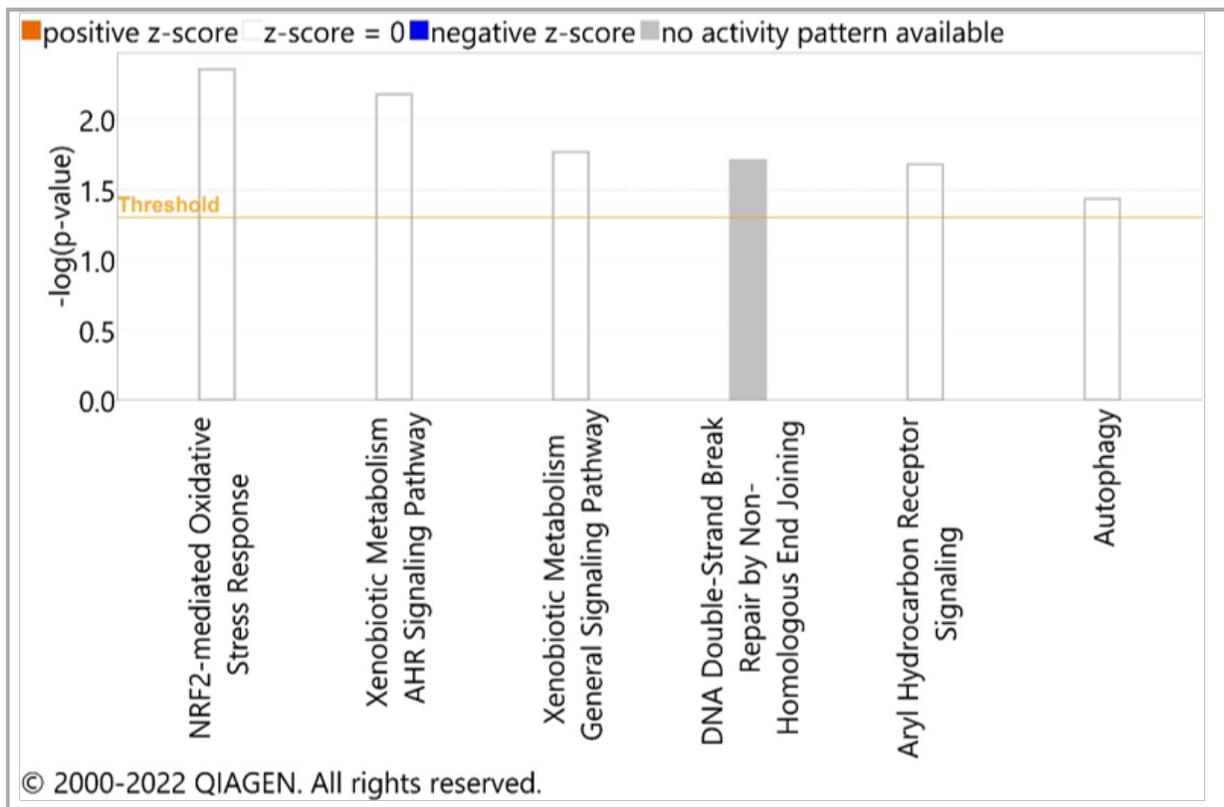


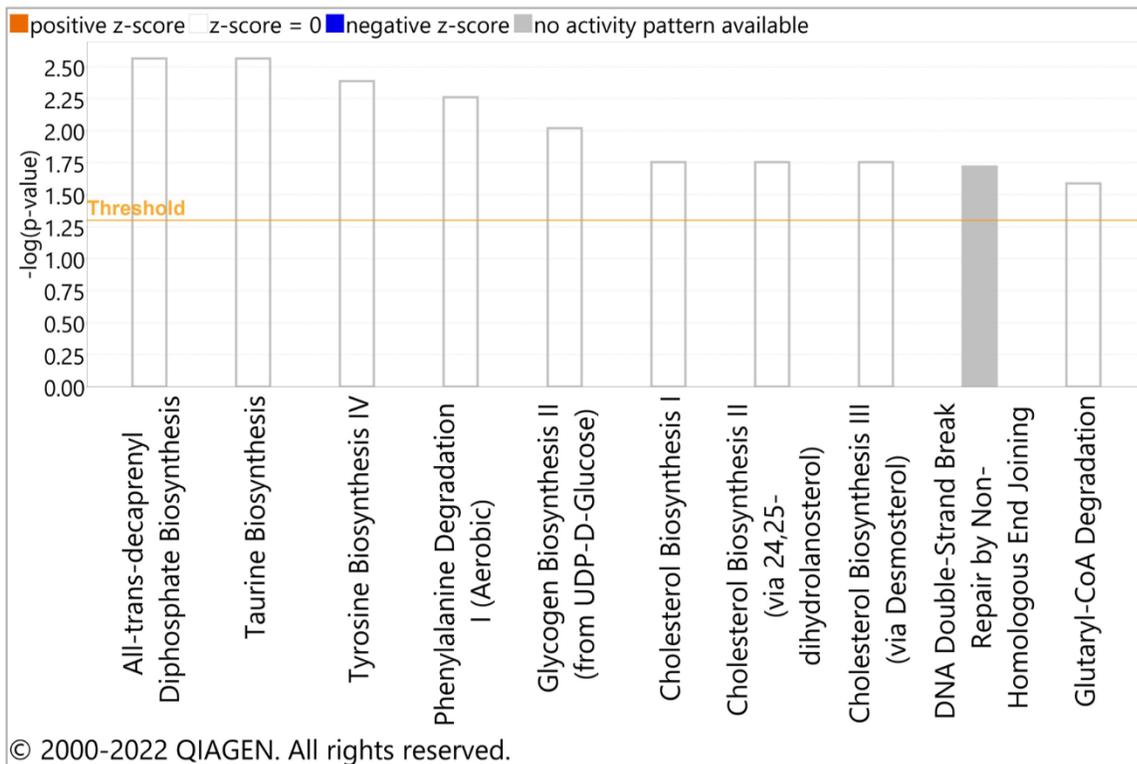
Figure 3.16. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) produced from one of twelve male SB Sires 1–12. Briefly, SB had been produced by crossing two female SB (Dam A and Dam B) with six different male SB (sires) each, whereby three of the sires each dam was crossed with were designated as “Large” or “Small” as they significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Offspring were reared in triplicate tanks of an indoor recirculating aquaculture system (RAS) and sampled at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth ($n=36$) and those that exhibited relatively poorer, or Inferior growth ($n=36$), and these fish significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Genes included in analysis (33 unique, analysis-ready molecules from a list of 35 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p=0.05$ indicated by “Threshold” line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made.

Figure 3.17. Enriched canonical pathways identified through Qiagen Ingenuity Pathway Analysis (IPA, Germantown, MD, USA) from the genes up-regulated in striped bass (SB) belonging to the “Large” (A) or “Small” (B) size group. Briefly, SB had been produced by crossing two female SB (Dam A and Dam B) with six different male SB (sires) each, whereby three of the sires each dam was crossed with were designated as “Large” or “Small” as they significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Offspring were reared in triplicate tanks of an indoor recirculating aquaculture system (RAS) and sampled at eighteen months of age. This sampled population was split into two equal groups based on size: those that exhibited Superior growth ($n=36$) and those that exhibited relatively poorer, or Inferior growth ($n=36$), and these fish significantly differed in weight and total length (Student’s *t*-Test, $p < 0.0001$). Genes included in analysis (71 unique, analysis-ready molecules from a list of 74 gene transcripts total) were quantitated in skeletal muscle tissue (i.e., fillet) and identified as those yielding optimal classification performance between groups (classes) of SB based upon dam parentage through the application of a machine learning workflow. Canonical pathways are designated as enriched based upon the number of molecules in the dataset associated to a given pathway and the calculated significance based upon the Fisher’s Exact Test ($p=0.05$ indicated by “Threshold” line parallel to x-axis). The color of each bar indicates the predicted activity as follows: blue represents inhibited pathways, orange represents activated pathways, grey represents pathways for which a prediction of activity cannot be made based upon the specific pathway construction and associated molecules, and white represents pathways for which the calculated z-score was 0.0 indicating that the evidence for activation and inhibition are equal preventing a prediction from being made.

A. Large Sire Offspring



B. Small Sire Offspring



© 2000-2022 QIAGEN. All rights reserved.

Figure 3.17.

CHAPTER 4. MACHINE LEARNING WORKFLOWS FOR BIOLOGICAL DATA

Linnea K. Andersen and Benjamin J. Reading

Abstract

Recent technological advancements have revolutionized research capabilities across the biological sciences by enabling the collection of large data that provides a broader picture of systems from the cellular to ecosystem level at a more refined resolution. The rapid rate of generating these data has exacerbated bottlenecks in study design and data analysis approaches, especially as conventional methods that incorporate traditional statistical tests and assumptions are not suitable or sufficient for highly-dimensional data (i.e., more than 1,000 variables). The application of machine learning techniques in large data analysis is one promising solution that is increasingly popular. However, limitations in expertise such that the results from machine learning models can be interpreted to gain meaningful biological insight pose a great challenge. To address this challenge, we provide here (1) a general overview of data analysis and machine learning approaches and considerations thereof and (2) a user-friendly machine learning workflow that can be applied to a wide variety of data types to reduce these large datasets to those variables (attributes) most determinant of experimental and/or observed conditions. The workflow presented here has been beta-tested with great success and is recommended to be incorporated into analysis pipelines of large data as a standardized approach of reducing data dimensionality. Moreover, the workflow is flexible and the underlying concepts and steps can be modified to best suit user needs, objectives, and study parameters.

Introduction

The exponential increase in the amount, type, and frequency with which we generate data is an earmark of the present Fourth Industrial Revolution (4IR or Industry 4.0) that has impacted all industries, research fields, and sectors of our current society. The biological sciences are no exception and the convergence of the digital and physical realms with the biological realm is another fundamental characteristic of the 4IR. The available and developing tools and methods to generate biological data allow for the collection of information that is broader in scope and of greater resolution than ever before. The continuously improving and emerging technologies to measure the cellular and molecular components that fall under the umbrella of ‘omics’ (e.g., genomics, epigenomics, transcriptomics, proteomics, metabolomics, metagenomics, ionomics) are paramount of this impact (Zampieri et al., 2019). Superior capacity, accuracy, and, perhaps most importantly, accessibility (i.e., usability and affordability) of such technologies has rapidly facilitated the incorporation of omics data into studies across the biological sciences and in adjacent fields (Noor et al., 2019; Schrider and Kern, 2018; Silva et al., 2019). Further, the ability to integrate multiple types of omics data (‘multi-’, ‘poly-’, or ‘integrative-’ omics) has empowered researchers to take a true systems biology approach whereby biological entities can be modeled as a whole consisting of multiple interacting elements sharing organizing principles rather than merely as a sum of its parts (Gilpin et al., 2020; Tavassoly et al., 2018; von Bertalanffy, 1968). This holistic approach to understanding biological systems has already led to groundbreaking innovations and novel research insights in areas such as personalized medicine, precision agriculture, nanotechnology, materials science, and so on (Graw et al., 2021; Karthikeyan and Priyakumar, 2022; Sarker, 2021). The burgeoning availability of high-throughput omics data has not come without challenges, however. Although these biological data

may not consistently meet the criteria to be considered ‘big data’, the 3Vs and/or the 5Vs (i.e., volume, velocity, variety, veracity, and value), they are typically ‘large’, or highly-dimensional (i.e., more than 1,000 variables/parameters measured), something that biologists are typically not accustomed to, and thus present similar bottlenecks in handling, analysis, and interpretation (Misra et al., 2019; Younas, 2019).

Such challenges of analyzing large (and big) data can be met with another characteristic element of the 4IR: the use of artificial intelligence (AI), or computer systems designed to perform decision-based tasks as an intelligent agent (i.e., human or other animal) without direct command from a human, to transform highly-dimensional data into information (i.e., data mining). The algorithms and technologies that enable pattern recognition and computational learning theory required for a computer system to make such intelligent decisions based on input data collectively fall under the umbrella of machine learning (ML). These ML algorithms allow AI systems to induce new knowledge from prior experiences (i.e., “learning) and numerous, versatile strategies to derive meaningful biological insight through the application of ML algorithms have been demonstrated with great success (Mishra et al., 2018; Villoutreix, 2021). Despite promise, limitations of computing power, algorithm scalability, and expertise in applying algorithms and interpreting models pose hurdles for integrating ML-based approaches across the biological sciences.

The hurdle of limited expertise in applying ML approaches to biological research is multifaceted and includes obvious elements, such as requiring a certain level of computational knowledge and resources and/or an understanding of the resources that may be necessary to perform ML analyses (e.g., high-performance computing systems), but are not necessarily commonplace among biologists and the institutions where they conduct research. A more

complex element of the limited ML expertise hurdle, however, is the stray from traditional methodologies and thought processes of experimental design and interpretation that have been the foundation of research in the biological sciences for over a century. Specifically, the adoption of ML approaches across the biological sciences requires a shift towards what Breiman (2001) describes as the “algorithmic modeling culture” in which the mechanisms of the system that causes the output y to result from the input x , also referred to as the “black box,” is understood to be unknown. Regarding the black box as unknown and therefore as having the potential to be far different or more complex than what is presently understood facilitates discovery-based data processing through which previously unconsidered factors or new phenomena are revealed from data in an exploratory, “bottom-up” fashion (**Figure 4.1**) (Schrider and Kern, 2017).

Traditional Approaches to Data Analysis

The complex, unknown black box is a far fall from the “data modeling culture” in which the black box is thought to be a stochastic model that accurately and appropriately represents the mechanism(s) that causes x to yield y (Breiman, 2001). This modeling culture is predicated on there being previously known information that allows for data to fit a given model with the goal of testing a hypothesis to draw conclusions in a “top-down” fashion that has become convention across the biological sciences (**Figure 4.1**). Indeed, planning and conducting research based on defined, justified (by observation) hypotheses is the backbone of the scientific method and doing so has enabled the reductionist methodology of breaking a system or phenomenon down into simpler, individual parts (i.e., units of testable hypotheses) to become standard practice (Fang and Casadevall, 2011). The concepts and models employed by the data modeling culture are familiar and so well-established in some areas that there are formally specified requirements for

conducting surveys and experiments to ensure data fit (Smith, 2018), such as providing a power analysis as a justification for experimental sample size alongside funding applications.

Although the reductionist methodology has led to great successes in biological research, the assumptions imposed by the data modeling culture and reductionist approach are restrictive and can easily lead scientists to sacrifice accuracy for interpretability (Breiman, 2001; Fang and Casadevall, 2011). At the core of the data modeling culture and reductionist, top-down approach are the assumptions that (1) a natural system or phenomenon can be accurately explained by a model we are able to establish, including methods implemented to account for any compounding and/or unknown factors, and (2) the information subsequently garnered from an isolated component of a whole, whether an organism or a single molecule, can provide enough explanation to understand how that component operates within the entire system itself (Breiman, 2001; Regenmortel, 2004). Although these assumptions may hold true in some cases, the insurgence of technologies to measure variables at a previously unthinkable resolution, or even those not yet discovered, raises the question of whether such assumptions are simply fallacies. The requirements of the mathematical models commonly used to test hypotheses under these approaches further restrict analyses as they often require strict adherence of the population (broader and/or the subpopulation sampled) to additional assumptions such as normal distribution, common (homogeneous) variance, random and independently derived observations, and additive factor effects. Although there are alternatives that are not bound by the same confining assumptions of population distribution (e.g., nonparametric tests), employing these methods is often at the cost of statistical power, precision, accuracy, and/or measure of effect.

The assumptions and restrictions imposed by these methods only narrow as we move from philosophy to actual practice. The statistical models utilized by the data modeling culture

typically test hypotheses by estimating and distinguishing a result from other possibilities (i.e., inferences based on known properties and sample distributions), perhaps most commonly by asking: “Does a given parameter differ between groups and can this difference be attributed to random chance?” (Anderson et al., 2000; Jones and Matloff, 1986). If observations can be attributed to random chance there is *not* a statistically significant relationship, difference, or other trend that connects the independent variables to the dependent variables and thus the “null” hypothesis is regarded as true. If observations cannot be attributed to random chance, the “alternative” hypothesis is considered true whereby there *is* some statistically significant relationship, difference, or other trend influencing the measured outcomes.

Introduced by Ronald Fisher in the 1920s, the p-value is a familiar and recognizable, if not *the* most familiar and recognizable, index for determining a statistically significant relationship, difference, or trend (Nuzzo, 2014). The p-value measures the probability that the observed results are consistent with random chance and deeming observations statistically significant is contingent on the p-value being less than or equal to a set alpha level indicative of the probability of rejecting the null hypothesis when it is true (i.e., a Type I error, false positive) (Nuzzo, 2014). The alpha level broadly accepted as standard is 0.05, meaning researchers accept up to a 5% chance that the null hypothesis (i.e., no relationship, difference, or trend identified among data analyzed) will be rejected in favor of the alternative when the null is actually true. Thus, the interpretation of statistical analyses that utilize the p-value is married to the rejection or acceptance of a null hypothesis. If the parameters of the study and actual measured values are well understood, this is not necessarily an issue. However, the high adaptation of the p-value and subsequent extrapolation of its meaning when presenting the inference gained about a given population is an issue (Nuzzo, 2014). Specifically, the p-value is limited by the risk of Type I

Error proclivity and the narrowing of scope caused by restricting analyses through the test assumptions, power analyses, known unknowns, unknown unknowns, etc. In other words, even if Type I (or Type II) Errors are not committed, the research question the statistical test is set to answer becomes so specific by the time researchers arrive at accepting or rejecting the null hypothesis that any interpretation of the findings beyond exactly the scope of analysis could easily lead to supposition or generalization (Nuzzo, 2014). Further still, the p-value fails to indicate the magnitude and range of a given effect, nor is statistical significance a proxy for measuring interactions between elements of a whole (Nuzzo, 2014; Halsey, 2019). In fact, considering p-value alone when interpreting the results and implications of a study is not considered a sufficient measure of a model or hypothesis by the American Statistical Association (Wasserstein & Lazar, 2016).

There are alternatives to the p-value that are still within the realm of traditional data modeling approaches, such as Bayesian methods (Bayes factor), which aim to indicate whether the data are better predicted by the null *or* alternative (i.e., not exclusively the null), and effect size along with confidence intervals, which aim to indicate how strong the effect is and how accurate the effect value is relative to the greater population effect (Halsey, 2019). However, many tests employed as alternatives to those that utilize the p-value are also restricted by assumptions such as the reasonable representation of the data by the chosen likelihood function (Bayesian models), or normality and homogeneity of variances (effect sizes, confidence intervals) and in some cases rely more heavily on graphical (visual) interpretation (Halsey, 2019). Moreover, and despite exceptions such as the principal component analysis (PCA) or partial least squares (PLS) regression, many traditional statistical tests are not equipped to account for datasets that have more variables than observations or the correlation of several

independent variables in a dataset (multicollinearity), operate based upon comparisons of means (averages), and are unable to appropriately address individual items (outliers) in concert with the full dataset, all common features of omics and other large biological datasets (i.e., Curse of Dimensionality). In all, exclusive use of the data modeling culture and reductionist approach are generally limited in application to biological research if the intent is to derive meaningful results and these limitations are exacerbated in areas such as multi-omics and quantum biology where the understanding that there are many unknown interactions and underlying mechanisms is inherent at the theoretical level and in practice (Edwards and Thiele, 2013; McFadden and Al-Khalili, 2018; Outeiral et al., 2020; Regenmortel, 2004).

The address limitations to the standard approaches and techniques employed across the biological sciences is critical as the amount of and speed with which data are generated is increasing exponentially and as we near projected timeframes for experiencing more severe impacts of global issues such as food security, resource shortages, and emerging infectious diseases, and so on (Bashura et al., 2021; Meshram et al., 2021; Selvarajoo, 2021). Thus, we are at a crossroads akin to the biomathematics revolution of the early last century (“Statistics”) that requires knowledge transfer, standardization, and benchmarking of ML approaches that can be utilized to derive insights that are both novel and leverageable in developing practical applications of research across the biological sciences (Quinn, 2012; Stone, 1961).

Below we provide an overview of ML and a workflow for applying ML techniques to determine the variables, referred to herein as attributes, most underlying the mechanism(s) that lead x input to yield y output. Specifically, we summarize the types (i.e., categories and subcategories) of ML and processes of learning, including descriptions of commonly used cross-validation (CV) methods, evaluation metrics, and considerations for utilizing ML. The ML

workflow incorporates multiple ML techniques and algorithms that are of distinct architecture from each other to facilitate the reduction of dimensionality of large data while accounting for limitations of each element and without imposing many restrictions or requirements of the data, system, or otherwise. The reduced dataset resulting from application of the ML workflow are the attributes that are most determinant of the black box to the extent it can be represented by the input data. Our goal is to present a standardized template for incorporating ML into analysis of large biological data and in doing so, further bridge the gap between the data modeling and algorithmic culture towards one of scientific advancement that does not shy away from discovery-based approaches.

An Overview of Machine Learning

ML approaches are often broadly categorized as belonging to one of four areas based on input (e.g., known or unknown groupings), objective (e.g., to cluster), and algorithm architecture (e.g., to perform a regression vs neural network): (1) Traditional Learning, (2) Reinforcement Learning, (3) Neural Nets and Deep Learning, and (4) Ensemble Learning (Karthikeyan and Priyakumar, 2022) (**Figure 4.2**).

Categorizing ML approaches based on input is dependent on whether or not distinct label(s) describing groupings of instances are known, or provided to the given algorithm, or not. If labels are known and therefore included in the training data to carry out a *task-driven* process that answers the question, “Is there a difference between classes?” (e.g., a difference between treatment groups such as Experimental and Control), the ML approach is considered “Supervised.” If labels are not known, the ML approach is considered “Unsupervised” and learning is considered a *data-driven* process that answers the question, “Are there meaningful

patterns, structures, and/or features?”. The Supervised Learning approach is more commonly used in biological research as it includes already well-known tests such as linear regression and the question posed by the *task-driven* approach is more familiar in the context of conventional hypothesis testing (i.e., experimental design). Subsequent categorizations of ML approaches based on objective and/or algorithm architecture are generally directed further once users determine whether the analysis will be supervised or unsupervised as described below. The terminology used to describe subtypes and techniques of ML algorithms herein are defined in **Table 4.1** alongside examples of comparable terms used to describe equivalent or similar concepts in traditional statistics and/or biological research. **Table 4.2** lists some of the popular algorithms belonging to each ML category and corresponding subtypes where applicable.

(1) Traditional Learning

Traditional Learning approaches are often categorized based upon whether they have historically been applied in a Supervised, Unsupervised, or Semi-Supervised manner and are further grouped as such here (**Table 4.2, Figure 4.2**). However, it should be noted that some algorithms of any type in the category of Traditional Learning have been developed further to accept supervised or unsupervised input. Moreover, many approaches that are not considered Traditional Learning can also be characterized as supervised, unsupervised, or semi-supervised.

Supervised ML approaches in the Traditional Learning category can be further divided into two types: Classification and Regression. Classification techniques predict the class of each instance based upon values associated with the attributes. Regression techniques predict a continuous value assigned to each instance along a numeric axis based upon values associated to

attributes and then subsequently determine the relationship (i.e., correlation). Supervised approaches that utilize past and present data to predict future outcomes are sometimes referred to as “forecasting.” As mentioned above, supervised learning approaches are likely to be the approach of choice for many biologists as existing familiarity with algorithms such as linear regression (Table 4.2) and concepts of conventional experimental study design lend themselves to an easy transition from statistically determining differences between user-defined groups to identifying factors underlying differences between user-defined groups (i.e., classification and regression).

Unsupervised ML techniques in the Traditional Learning category can be further divided into two types: Clustering and Association. Clustering techniques divide instances based on unknown features and the ML algorithm determines the best groupings based upon values associated to attributes and without the inclusion of predefined classes. Association techniques identify patterns of attribute arrangements based on their measurement or the sequence in which they occur. Dimensionality Reduction is sometimes regarded as a third type of Unsupervised Traditional Learning, whereby higher-level attributes are assembled from specific attributes into a smaller dimensional space based upon the associated values. Dimensionality Reduction techniques are often used to visualize data or as a pre-processing step prior to Supervised Learning. Unsupervised Learning approaches are less labor intensive than Supervised approaches as no information regarding labels needs to be known. Unsupervised Learning approaches lend themselves to discovery-based research whereby patterns amongst data can be uncovered without constraints imposed by labeled classes or in the absence of experimental design.

Semi-Supervised Learning is a hybridized technique of Supervised and Unsupervised Learning, whereby algorithm input includes both labeled and unlabeled data. Semi-Supervised Learning is useful in instances where some number of instances do not have known labels and instances that do have labels are able to aid in the ability of the algorithm to accurately predict outcomes of those unlabeled instances. As such, Semi-Supervised Learning is advantageous in areas of biological research such as prediction of drug-protein interactions whereby researchers have certain information known (i.e., highly characterized interactions) and must account for what are presumably numerous unknown interactions (Xia et al., 2010).

(2) Reinforcement Learning

Reinforcement Learning is an *environment-driven* process that does not require external supervision for training whereby algorithms learn via experiences resulting in positive and negative feedback (i.e., decisions made in an interactive environment). Reinforcement Learning approaches can be further grouped as Value-Based, whereby the value (a real number) of a given state is learned; Policy-Based, whereby the value of a given strategy of action (policy) is learned; or Model-Based, whereby the environment model is learned and strategy is built based upon this (Table 4.2, Figure 4.2). In all Reinforcement Learning approaches outcomes are defined as either reward or penalty (positive or negative) and algorithms are trained to make decisions based on what will increase the occurrence of reward and decrease the occurrence of penalty. Reinforcement Learning techniques have vast applications to the biological sciences, particularly with the rise of autonomous systems such as prosthetics or those programmed for unmanned ocean exploration (Gunnarson et al., 2021; Tang et al., 2022, Wen et al., 2020).

(3) Neural Nets and Deep Learning

Neural Nets and Deep Learning are a category of ML that can be supervised or unsupervised and used in lieu of Traditional Learning techniques for tasks such as Classification. Neural Nets, also referred to as Artificial Neural Networks (ANNs), are a system based on the neural network of the animal brain whereby nodes, also referred to as artificial neurons, are connected and transmit a signal to other nodes (**Table 4.2, Figure 4.2**). The network begins with an input layer where data is recognized and initial signals are processed before being transmitted to subsequently connected nodes organized into hidden, intermediate layers. The signal at a connection, or “edge,” is a real number computed by a nonlinear function of the sum of node inputs and then used to assign a corresponding weight. Nodes with heavier weights will have a greater effect on the next layer than those assigned lighter weights until the output layer, representative of the ultimate predictions, is reached. The direction of data input and processing can be used to describe ANNs. For example, a Feed-Forward network (e.g., Multilayer Perceptron, MLP) processes data from input and through the hidden layer(s) to output in one direction, whereas Recurrent networks (i.e., back propagation) feed the output of processing nodes back into the model for learning. Deep Learning is used to describe the application of ANNs that have several hidden layers and are therefore particularly well-suited to analyze large, complex data. Deep learning approaches have wide applications across the sciences ranging from particle discovery to quantum chemistry (Baldi et al., 2014; Ching et al. 2018; Goh et al., 2017). At present, one of the most prominent applications of deep learning to the biological sciences is in image and video analysis for the successful completion of tasks such as cancer screening to

high-throughput phenotyping of plant-crops (Esteva et al., 2017; Feng et al., 2019; LeCun et al., 2015).

(4) Ensemble Learning

The Ensemble Learning approach is to combine models to reduce noise, variance, and bias for the goal of improved accuracy in learning and decision making (Witten et al., 2011). The simplest Ensemble method is to run multiple ML algorithms on a given dataset and determine the ultimate prediction based upon the mean or mode of algorithm predictions (Witten et al., 2011). Three popular strategies among the more advanced Ensemble methods are Bagging, Stacking, and Boosting (Witten et al., 2011) (**Table 4.2, Figure 4.2**). Bagging, or “bootstrap aggregation”, is the fitting of several decision trees on different subsets of the same training data (with replacement) and averaging the prediction. Stacking, or “stacked generalization”, is the fitting of several algorithms, referred to as “Level-0” learners, on the same data and then applying a different, meta-algorithm, referred to as the “Level-1” model, to determine the best way to combine predictions. Boosting is the fitting of an initial algorithm to an entire dataset and then fitting subsequent algorithms on subsets that include residuals (i.e., inaccurately predicted attributes) from the previous boosting algorithms in a sequential manner. The workflow described below is an Ensemble Learning technique as it incorporates multiple algorithms and in doing so (and cross-referencing the results) allows for researchers to validate predictions with reduced dispersion compared to that of a single algorithm. Moreover, the use of multiple algorithms in this way that are distinct enough from each other allows for the nuance of one to address and/or account for the limitations of another. Ensemble Learning is particularly useful in

the biological sciences as researchers implementing these techniques advance in their own literacy and expertise with them and increased model robustness and confidence in outcomes are paramount.

How Learning Works and Cross-Validation Techniques

Learning occurs across all categories of ML via the processing of a training set and test set by a given algorithm. The training set is a representative subset of the data used to train the algorithm to recognize features and patterns of the data (i.e., build the model). The test set is a subset of the data independent from the training set (i.e., no replacement) that is used to evaluate the performance of the model in determining outcomes. In some cases, a single experimental dataset can be subdivided into training and test sets and in other cases the training and test sets can be replicate experiments.

Cross-validation (CV) techniques allow for users to define the partitioning of data into training and test sets for the purpose of evaluating how well a given model performs. Specifically, CV techniques facilitate the examination and comparison of model performance on a given dataset by providing benchmarks of model performance on multiple permutations of the complete dataset. Some of the most widely-used CV techniques include:

1. Holdout Method (Percentage Split)
2. K-Fold Cross-Validation
3. Stratified K-Fold Cross-Validation
4. Leave-P-Out Cross-Validation
5. Leave-One-Out Cross-Validation

The holdout method, or percentage split, randomly divides a single dataset into the training and test sets based on a user-defined percentage. For example, a 66% percent split would stipulate that 66% of data is included in the training set and 34% of data in the test set (Schilling et al., 2014, 2015). Ideally the CV technique will mitigate bias by incorporating representative samples for the training and testing sets such that an equal or near-equal number of instances from each category are included in both sets, however, holdout method does not guarantee this as much (Witten et al., 2011). To account for the possible lack of representation of the whole dataset, users may verify sample representation after the run of an algorithm and/or choose to repeat the training and testing process several times with different random selections of instances incorporated into training and testing sets. The latter method of error rate estimation is referred to as a repeated holdout, whereby users average the error rate from each run to calculate an overall error rate (Witten et al., 2011).

The K -fold cross-validation method randomly partitions the data into K subsets, or folds, and the model is trained on $K - 1$ number of folds and tested on the K th fold. This process is repeated until each of the folds has served as the test set and the algorithm is evaluated based upon the average of all performance metrics calculated from each K number of times the algorithm was applied. A tenfold cross (10) has become a general standard based on considerable testing of numerous datasets (Witten et al., 2011). Reassignment of folds to either the training or testing set reduces bias observed with other CV techniques that do not ensure that all instances are included in both the training and test set for the overall building of the model. However, the K -fold CV method lends itself to building an imbalanced model as the data are randomly divided into folds. The stratified K -fold CV method accounts for this by rearranging data to ensure each fold is representative of the whole dataset, a process referred to as stratification. Although

comprehensive, it is important to note that the requisite computational power to perform either K -fold CV method increases as the number of folds increases.

The Leave-P-Out and Leave-One-Out CV methods are considered exhaustive CV techniques, whereby “exhaustive” is defined as the testing of all possible partitions of the entire dataset into the training and test sets. The Leave-P-Out CV method will remove a number of data points, p , from the total number of data points, n . The model is trained on $(n - p)$ data points, tested on p data points, and this process is repeated until all possible combinations of p are removed from the training set and used as the test set. Similar to the K -fold CV techniques, the model is evaluated based upon the average performance of each iteration of this CV and as p increases the requisite computational power increases. The Leave-One-Out CV technique is similar albeit less exhaustive than the Leave-P-Out technique as p is defined simply as 1.

It is important to note that the above CV strategies may not be appropriate for the analysis of time series or other ordered data. In these instances creating subsets (folds) in a forward-chain manner, such that the first in order in the time series is used for training a model tested on the second in order, which are used next both used to train a model tested on third and so on and so forth.

Evaluating Learning: Measures of Model Performance

Learning is generally evaluated through measures that describe the accuracy of a model in making predictions about instances that agree with the actual value (i.e., the success rate).

Some examples of learning evaluation measures are provided below:

Loss Function - In the simplest case where predictions are either correct or incorrect, model performance can be evaluated through a $0 - 1$ loss function where loss is 0 if correct and 1

if incorrect (Witten et al., 2011). A given learning approach may also assign a prediction probability, or a value of probability that a prediction agrees with the actual value, and users may take this into account (e.g., apply the quadratic loss function to evaluate predictive probability) (Witten et al., 2011).

Confusion Matrix: True and False Predictions and Kappa Statistic - In other scenarios, users may want to also evaluate the *cost* that incorrect predictions may have by calculating an overall success rate based upon the rate of correct predictions (true positives, TP, and true negatives, TN) and incorrect predictions (false positives, FP, and false negatives, FN) (Witten et al., 2011). Algorithm performance summaries may display this information in a confusion matrix of TP, TN, FP, and FN values for each predicted outcome. Many other performance metrics can be calculated from these values. For example, the rate of correct predictions (i.e., true positive rate, or sensitivity) is calculated as all true positives divided by all positives [$= (TP) / (TP + FN)$] and the rate of incorrect predictions (false positive rate) is calculated as all false positives divided by all negatives [$= (FP) / (FP + TN)$] (Witten et al., 2011). The overall success rate (i.e., accuracy) is then calculated as the total correct predictions divided by all predictions [$= (TP + TN) / (TP + TN + FP + FN)$] (Witten et al., 2011).

To account for the number of correct predictions that may occur as a result of random chance, a metric referred to as the Kappa statistic can be taken into account (Witten et al., 2011). The Kappa statistic can be calculated as $[K = (p_o - p_e) / (1 - p_e)]$, where p_o is the overall accuracy of the model and p_e is the measure of agreement between the predictions and actual outcomes if occurring by chance, or $[p_e = ((TP + FN) * (TP + FP) + (FP + TN) * (FN + TN)) / (TP + TN + FP + FN)]$. Kappa values range from 0.0 to 1.0, where a value of 0.0 indicates prediction is expected based on random chance and a value of 1.0 indicates prediction performance is based

on true-learning (i.e., more information than random chance) (Witten et al., 2011). The range of Kappa statistic values and corresponding prediction strength of ML algorithms is provided in **Table 4.3**.

Area Under the Receiver Operating Characteristics (AUROC) Curve - The term “Receiver Operating Characteristics” (ROC) specifically is used to characterize the compromise between actual hits (TP and TN) and false hits (FP and FN) over a noisy channel (Witten et al., 2011). The ROC curve is a way to visualize algorithm performance without accounting for distribution among classes or the cost of incorrect predictions with the true positive rate on the y-axis and the true negative rate on the x-axis (Witten et al., 2011). In this way, the y-axis of the ROC curve is similar to a lift chart, a graph depicting the improvement of a model compared to random assignment where lift is the measured change between values, except values are expressed as a percentage here (Witten et al., 2011). An optimal classifier will chart in the upper left corner of a ROC curve graph, thus the area under the ROC, or the AUROC, can be used as a measure of performance for a given algorithm where a greater area is indicative of better performance. The range of AUROC values and corresponding prediction strength of ML algorithms is provided in **Table 4.3**.

Recall-Precision Curves - The Recall-Precision curve is a graphical representation of the recall (x-axis) and precision (y-axis) of a given algorithm, where recall is calculated as the true positive rate (*see above*) and precision is calculated as $[= (TP) / (TP + FP)]$. A higher recall-precision area under the curve (AUC) is preferred as with the AUROC value, however, here algorithms with high recall and high precision will chart in the upper right (Witten et al., 2011).

Evaluating Numeric Predictions - In classification or similar ML tasks, errors are ultimately present or absent, whereas with numeric predictions errors will vary in size (Witten et

al., 2011). The most commonly used strategy to evaluate numerical predictions is to calculate the mean-squared error (Witten et al., 2011). Depending on how well-suited certain test parameters are for the evaluation of a given algorithm, users may opt to evaluate performance based upon the root mean-squared error, mean-absolute error, relative-squared error, root relative-squared error, relative-absolute error, and/or Correlation coefficient (Witten et al., 2011).

Machine Learning Workflow

Specific Workflow Elements

Here we present a supervised ML workflow for the reduction of highly-dimensional biological data to only the attributes (variables) that most underlie the distinction between instances (samples, observations) into user-defined classes (groups) representative of some trait(s) and/or feature(s). Definitions of ML terminology used to describe this approach are provided in **Table 4.1**.

The workflow enables a conservative reduction of data dimensionality by employing multiple empirical ML techniques in an Ensemble Learning fashion, whereby four orthogonal algorithmic approaches, each applied with at least two CV strategies are utilized. A diagram of the workflow is provided in **Figure 4.3(A)**, complemented by a description of this workflow using an example dataset on the right panel **(B)**. In this example, 10,000 genes were identified among twenty biological replicates (i.e., individuals), or instances that are split equally into two classes (groups), “Good” and “Bad” based upon some observable quality. Additional text referencing the example is included at various stages of the workflow description below in *italic* typeface.

Data is first reduced based upon measures of information gain (“Shannon’s Entropy”), by which attributes that are not informative to the algorithm in building a well-trained model are removed and the remaining attributes are assigned a rank relative to the amount of information gained. These attributes are included in subsequent analysis as the “*ranked attribute set*.” The four CV algorithms are then applied to “*subsets*” of the *ranked-attribute set* that include fewer and fewer bottom-ranked attributes until only those highest ranking attributes (i.e., greatest information gain) are included in learning. Model performance (success) values from each application of the Ensemble method are compared to determine the “*optimal dataset*”, or the fewest number of highly-ranked attributes required (i.e., included as input) to yield well-trained models from all or the majority of cross-validated algorithms. The attributes in the *optimal dataset* can be considered those of all attributes measured that are driving the observed differences between classes and that account for model overfitting (Type II Error, false negatives) and underfitting (Type I Error, false positives).

The algorithms applied in this workflow fall under three of the four broad categories of ML: Traditional (Classical) Learning, Neural Nets and Deep Learning, and Ensemble Learning. Specifically, two algorithms used fall under the Traditional (Classical) Learning category and Classification subtype: the Sequential Minimal Optimization (SMO, a support vector machine or SVM), and J48, a decision tree. The third algorithm, Multilayer Perceptron (MLP), is an ANN and thus categorized under Neural Nets and Deep Learning. The fourth algorithm, Random Forest, is a Boosting, or Bootstrap Aggregation, algorithm that in and of itself is categorized as Ensemble Learning. Each application of the overall Ensemble approach to the full dataset or a given subset thereof consists of eight “*runs*”, whereby a *run* is the use of one algorithm cross-validated with one of the two strategies outlined in this workflow, the Holdout Method

(Percentage Split) and the Stratified K-Fold Cross-Validation. Predictor performance will be evaluated based upon Percent Correct Classification, Kappa statistic, and AUROC values returned after each *run*.

Employing the use of algorithms belonging to multiple categories of ML (i.e., fundamentally different at the mathematical and conceptual levels) and two different CV strategies allows users to account for disadvantages or limitations of each and that may be encountered as a result of the study design, data type, and similar factors. Further, agreement between such orthogonal algorithms indicates robustness of the relationship reported by the model. By accounting for such, this workflow provides a sound, reliable method of reducing data dimensionality at a high-resolution to identify the most important attributes that underlie a predictive relationship of samples (biological and technical replicates) to a given user-designated classification of treatment, trait, or outcome of interest.

Software - This workflow was developed using open source software, Weka 3 (Waikato Environment for Knowledge Analysis), produced by researchers at the University of Waikato, New Zealand (Eibe et al., 2016, <https://www.cs.waikato.ac.nz/ml/weka/>). Weka tasks can be executed in the graphical user interface (GUI) or using Java packages (i.e., command-line Java code or Weka application program interface, API) and there are several Weka-associated resources of value, such as the ARFF (Attribute-Relation File Format) file format that is of particular use for input datasets containing more than 16,383 attributes, as this amount will not be compatible with Microsoft Excel for CSV formatting (i.e., limit to number of columns when transposing data). However, the workflow instructions herein are generalized to the extent that users are able to adapt this workflow to an environment or software of their choice; the

algorithms and CV methods or close equivalents are widely available for use in software packages and/or command-line interface (CLI).

Workflow Instructions

The ML workflow is broadly divided into three stages: (I) Data Collection and Initial Exploration; (II) Model Building or Pattern Identification; and (III) Cross-Validation or Verification of Prediction with New Data. The first stage of data collection and initial exploration includes quality control, normalization, filtering, and formatting of data. The second stage of model building or pattern identification typically involves fitting the complex yet "generic", or more "mainstream", models to the data that are not related to any reasoning or theoretical understanding of underlying causal processes. The third stage of CV or verification of prediction with new data serves to demonstrate accurate predictions or classifications using the dataset used in CV. The workflow concludes with the (IV) Interpretation of Results and (V) Potential Next Steps.

(I) Data Collection and Initial Exploration

Once an experimental design has been established, data should be collected and treated as you would normally based upon the study design and any other conventions specific to the given data type(s). This includes data cleaning, transformation, address of outliers, etc. A metadata file should be created and include the assigned class labels that will be used to perform the supervised ML workflow. Curating a clear, interpretable file of metadata is essential for interpreting analysis results in the greater context of the experiment and for facilitating future

usability of the data. In biological datasets it is likely, and therefore expected, that instances (i.e., samples/replicates) can be assigned or categorized into multiple groupings per the experimental design (e.g., Control and Treatment, Before Treatment and After Treatment, tissue sample types, geographical location of origin) and outcomes measured during and/or after the conclusion of the study (e.g., phenotype, survival, success of a measured action). As such, users may decide to apply this workflow multiple times to a given dataset whereby labels associated with different groupings of instances are included with each application and thus allowing for comparisons to be made between reduced datasets later. Once data and metadata have been curated users will likely have to format separate files for input into Weka or another program. **Figure 4.4** shows what the input file for the example described here would look like if formatted for Weka analysis.

Example: A metadata key for the example study would include unique sample IDs (Sample 1, Sample 2...Sample 20) for each individual, or instance, any morphometric, phenotypic, or other measured data collected, and labels for each categorization of instances. In this example there are twenty instances ($N = 20$ individuals) and therefore in each category, such as quality, there will be twenty labels. If Samples 1-10 are labeled as “Good” and Samples 11-20 are labeled as “Bad”, ten instances belong in the class “Good” and ten instances in the class “Bad”. If the labels entered were unique (e.g., “Good1, Good2, ... Bad10”) Weka would treat each as a distinct class and build a model based on the classification of instances into twenty classes (groups) rather than two.

(II) Model Building

Establish Cross-Validation Strategies - This workflow instructs users to use the holdout method (percent split) and Stratified K-Fold CV as CV strategies for each algorithm. Prior to analysis users should be sure to consider the suitability of the standard parameters for these strategies, described in greater detail below, and any adjustments that should be made (e.g., selection of different strategies).

We generally recommend the 66% split for the holdout method (default in Weka), which stipulates that 66% of data is included in the training set and 34% of data in the test set. If using Weka, an equal number of instances from each class will be included in the training and test sets and a random number generator is employed to select instances, whereby an average performance for sampling possibilities is calculated and therefore users only need to run one model to obtain average performance (Reading, 2018; Witten et al., 2011).

Example: In the ideal division of instances into the training and testing sets, the 66% split cross-validation technique will incorporate seven instances from the Good category and seven instances from the Bad category into the training set and the test set will consist of the remaining three instances from each category for cross-validation.

The stratified K-fold cross-validation partitions a set of n instances into K sets ('folds') of size n/K and the training set includes the full dataset minus a random fold, which is then used as the test set. This is repeated K times and the accuracy estimate is an average for each of the K folds. The Weka default is a tenfold cross (10) as this has become a general standard based on considerable testing of numerous datasets (Witten et al., 2011). However, users cannot stipulate

more folds than instances in any given class. For example, if the smallest number of instances in any class is 8, an 8-fold cross would be performed.

Example: As there are 10 instances in each class, Good and Bad, a 10-fold cross is appropriate. In this case, eighteen instances (9 from each class, $n - 1$ for each) would be randomly selected and incorporated into the training set and the test set will consist of the remaining two instances (1 from each class, $n/K = 20/10$) for cross-validation.

Analyze the Complete Dataset - Apply the following four algorithms to the complete dataset (i.e., with all attributes), each algorithm should be CV with both strategies (holdout method and stratified K-fold CV, or other as described above):

- SMO, sequential minimal optimization, a support-vector machine (SVM);
- MLP, multilayer perceptron, an artificial neural network (ANN);
- J48, a statistical classifier also referred to as C4.5, a decision tree; and
- Random Forest, a multitude of trees, or forest, used typically in bagging.

Record the outcomes of each run of a CV algorithm. We recommend maintaining two output logs: (1) a text document where the summary output can be copied and pasted, and (2) a spreadsheet where specific algorithm performance metrics (Percent Correct Classification, Kappa Statistic, and AUROC) from summary output will be extracted and used to determine points of overfitting and underfitting.

Feature Selection using Information Gain - Information gain, related to the concept of "entropy", is a measure of information gained from any one random variable (i.e., attribute) about possible outcomes (i.e., assignment to a class). Here, information gain values will be calculated to rank attributes based on how much information they provide to the algorithm for decision making towards the classification of instances. An information gain value of 0.0

indicates no information is gained and a value of 1.0 indicates maximum information is gained. Attributes that provide little to no information can be excluded from subsequent analysis to reduce data dimensionality.

Using the complete dataset with all attributes, run the SMO algorithm again with the cross-validation strategy that yielded the superior percent correct classification. The output will be an ordered list of all attributes from lowest to highest entropy (i.e., most to least informative). Assign all attributes a rank from 1 to n , where 1 is assigned to the attribute with the lowest entropy and n is the total number of attributes in the complete dataset assigned to the attribute with the highest entropy. Associate (i.e., link) the attribute rank values to the data for each instance and ensure values correctly correspond.

Select Features for Further Analysis - The extent to which information gain values can be used to reduce data dimensionality will vary dataset to dataset. In some cases, many attributes will be assigned a value of 0.0 and thus it will be clear that these attributes can be excluded from subsequent analyses. The attributes that are *not* excluded based on information gain values and/or other user decisions are referred to here as “*ranked attributes*.”

In other cases, few or no attributes will be assigned an information gain value of 0.0 and the user will have to determine a cutoff. Users may choose to set an arbitrary cutoff to remove attributes with an information gain value equal to or less than 0.05, or may decide based on factors such as the distribution of information gain values across all attributes. Creating a histogram of information gain values binned to the nearest tenths or hundredths decimal place may be useful to this end. Alternatively, users may opt to use p-values (or similar) to reduce data dimensionality. Specifically, users may opt to apply the ML procedure described above to only attributes that have been determined to meet a cutoff of significant difference (e.g., $p \leq 0.05$)

between groups based upon traditional statistical tests. In this case, the initial input data would be all *significant* attributes rather than all attributes for which measurements were collected.

Similarly, users may also opt to rank attributes by p-value or similar metric in lieu of information gain or consider p-value (or similar) alongside information gain rankings. The inclusion and/or reference to results derived from a more traditional statistical analysis approach alongside those from the ML workflow described here can serve as an orthogonal confirmation, a benchmark of sorts for evaluating methodological approaches, and so on. Incorporating outcomes from traditional statistical tests into the ML analysis described above is entirely at the discretion of the user and users are responsible for clearly describing such decision making (i.e., cutoffs, justification. interpretation) in the reporting of methodology, results, and findings.

Example: If 1,500 of the 10,000 attributes are assigned information gain values of 0.0, the user may choose to omit these from the subsequent analysis and create the “ranked attributes” dataset of the 8,500 attributes assigned information gain values greater than 0.0.

Model Building using Reduced Dataset - Run all twice-CV algorithms as described above using the reduced dataset and record the summary information after each run as above. We recommend ordering the attributes such that the first attribute listed (i.e., first column containing attribute name and corresponding data) is the highest ranked attribute (i.e., information gain value closest to 1.0) for ease of subsequent steps.

(III) Cross-Validation or Verification of Prediction with New Data

Determine Points of Model Overfitting and Underfitting - As described above, model overfitting is excessive complexity, whereby the output describes more random error and noise

rather than the underlying relationship which can lead to a high rate of false negatives. Conversely, model underfitting is oversimplification whereby the model is not representative of the population leading to a high rate of false positives (Reading, 2018). Thus, the reduction of data dimensionality generally serves to address overfitting, however, this does not address underfitting as the solution would be to include more attributes. To address this, users will remove bottom-ranked attributes (of the *ranked attributes* set) in an iterative fashion and then plot model performance to identify the thresholds of overfitting as the improvement of performance with the removal of attributes (e.g., improvement when 150 attributes are included compared to 200) and underfitting as the improvement of performance with the inclusion of attributes (e.g., improvement when 50 are included compared to 10). We will consider the “*optimal dataset*” to be that which includes the minimum number of attributes demonstrated to be important to include in learning for the correct classification of instances into classes.

First, users must create *subsets* of the ranked attributes dataset, whereby “subset” refers to a dataset that includes fewer attributes than the *ranked attribute* set such that each subset includes fewer of the highest ranking attributes by information gain value (or alternative, such as p-value). In other words, a greater number of bottom-ranking attributes from the *ranked attributes* set will be omitted from each subset. The number of bottom ranked attributes that can be omitted from each subset should be reasonable relative to the total number of attributes included in the ranked attributes set. For example, if a reduced dataset includes ~30,000 attributes a user may decide it more reasonable to create subsets of 25,000, 20,000, 15,000, 10,000 ... 1,000, 500, 400, 300, 200, 100, 50, 10 top-ranked attributes than 29,000, 28,000, 27,000, and so on and so forth.

Example: With a reduced dataset of 8,500 attributes, subsets may be determined as including 8,000, 7,500, 7,000, 6,500 ... 500, 400, 300, 200, 100, 50, 10, 5, 3, 2, and 1 top-ranked attribute(s). In this case, the subset with 8,000 ranked attributes would be all ranked attributes minus the 500 that had the lowest assigned information gain values among all 8,500, the subset with 10 ranked attributes would be the ten ranked attributes with the highest assigned information gain values, and the subset with 1 ranked attribute would be only the attribute with the highest assigned information gain value.

Run all twice-CV algorithms as described above using the reduced dataset and record the summary information after each run as above. Once complete, plot the percent correct classification (i.e., performance) for each of the four CV algorithms (y-axis) against the number of attributes included (x-axis). Add a trend line (polynomial or log function) to model the data. The threshold of overfitting will be where algorithm performance distinctly improves to some peak or maximum value with the removal of bottom-ranked attributes. The threshold of underfitting will be where algorithm performance distinctly improves to some peak or maximum value with the inclusion of top-ranked attributes. Refer to **Table 4.3** as needed to interpret the Kappa statistic and AUROC values.

Example: In keeping with the subsets described above, the performance of each cross-validated algorithm is plotted against the number of attributes included as input: all 8,500 ranked attributes and subsets of 8,000, 7,500, 7,000, 6,500 ... 500, 400, 300, 200, 100, 50, 10, 5, 3, 2, and 1 top-ranked attribute(s). We find that performance distinctly improves when the 8,000 bottom-ranked attributes are excluded (when input is only the top 500) and that algorithm performance reaches a peak when the top 50 attributes are included but decreases as more of the top-ranked attributes are excluded (when input is top 10, 5, 3, 2, and 1). The thresholds of

*overfitting and underfitting would be 500 and 50, respectively. See: **Figure 4.3(B)** for an example of this graph with the grey shaded region indicating both threshold values.*

It is important to note that the size of the dataset will influence the marked improvement or deterioration of algorithm performance. In smaller datasets (i.e., hundreds of attributes) there may not be as clear of a deterioration in performance due to overfitting, or algorithm performance may be generally poor across all subsets due to too few attributes having been measured for algorithms to clearly identify patterns in classification. If the algorithm has perfect performance (i.e., 100% correct classification, Kappa statistic of 1.0, and AUROC of 1.0) with all ranked attributes, reduce data dimensionality by identifying the minimum set of variables required to maintain optimal classification and increase the data dimensionality to identify the maximum set of variables that maintain the optimal classification. If algorithms have perfect performance with only the very top-ranked attributes (i.e., ≤ 5) included and therefore a point of underfitting is too difficult to ascertain, users can validate the influence on algorithm performance by performing all runs on subsets that include all or some other attributes. If users encounter the latter, they should take note that these top-ranking attributes are extremely familiar.

Compare Algorithm Performance - In order to determine the “*optimal dataset*” users must compare algorithm performance across all CV algorithms and identify the instance(s) of greatest agreement between algorithms at which a minimum number of attributes yielded peak (i.e., maximum or best) or near-peak performance. If agreement between CV algorithms is not clear (e.g., the inclusion of 500 top-ranked attributes was the overfitting threshold for three of four algorithms and including 400 top-ranked attributes was the threshold for the fourth), continue analyses using the least amount of attributes (i.e., 400). If performance seems to vary

greatly between algorithms, it may be useful to conduct a statistical analysis (e.g., ANOVA and post-hoc test, as appropriate) comparing the eight percent correct classification values for each subset (or at least those subsets with the greatest apparent variance between algorithms) (Witten et al., 2011). The results of this test may help users identify any outliers and subsequently justify why a given algorithm may be excluded from consideration should users decide to do so. For example, if one algorithm has significantly worse performance than the other algorithms it may not be worth a user attempting to find consensus among this and the other algorithms. In this case, users should look further into the limitations of that algorithm to gain insight into why the performance might be significantly different than the others.

Determine which cross-validated algorithm is optimal for the given dataset by identifying which algorithm has the best performance the greatest number of times (i.e., for each subset). Save the model learned from this algorithm as your “*optimal model*.” If the performance of the CV algorithms that are not considered the *optimal model* are drastically different or yield unexpected results, users may consider what the features and processes of the *optimal model* enable the superior performance as this may provide insight into the “black box” of what underlies the given observations amongst the data.

Establish the Negative Control for Learning - To demonstrate that algorithm performance is due to true, pattern-based learning users will establish a negative control for learning whereby the labels associated to instances in the optimal dataset will be randomized and run through each algorithm. Create ten copies of the optimal dataset with randomized labels (e.g., with the =*RAND()* function in Microsoft Excel) for each instance and verify that labels are randomized differently between each copy of the optimal datasets.

Run all twice-CV algorithms as described above on each of the ten randomized *optimal datasets* and record the summary information after each run as above. Calculate the mean and standard deviations of the percent correct classification, Kappa statistic, and AUROC for the set of runs performed on each randomized optimal dataset. The mean values of performance metrics should be approximate to what would be predicted from random assignment (i.e., Law of Probability) based on the number of classes and as according to **Table 4.4** (Reading, 2018). The average percent correct classification of algorithms run on negative control data being approximately what would be predicted from random assignment indicates that true learning has occurred in your previous predictions (i.e., when using input that does not have randomized labels). The results of this negative control should be reported in all papers using the method.

Example: In the example case, there are two classes—Good and Bad—therefore the correct classification based on randomized data should be around 50% (1:1, Law of Probability).

(IV) Interpretation of Results

After applying this workflow, users will have:

- A list of ranked attributes for which rank is indicative of the amount of information gained (Shannon's entropy) when a given attribute and values associated to each instance are included in learning and therefore could be considered of greater importance for the differentiation of instances into user-defined classes relative to attributes assigned a lesser rank (i.e., information gain of a relatively low value or 0.0).

- An optimal dataset of the minimum number of attributes required for optimal classification of these data into defined groups (i.e., classes); these attributes are the most important for differentiating between your samples into those defined groups and could be considered akin to being “significant” in terms of how they may be regarded in subsequent analysis.
- A performance comparison of several different ML algorithms that allows for deeper investigation such that changes in predictive performance may be indicative of interactions between attributes (i.e., how the influence of an attribute may change with the exclusion or inclusion of another).
- An optimal model for analysis of the input dataset that can be applied to future datasets, if desired.
- A negative control demonstrating that the ML procedure is valid and also that the Law of Probability has been upheld during your modeling.

(V) Potential Next Steps and Directions

There are many potential next steps and directions after applying the workflow, many of which depend on the type of data and intended use. Two examples are provided here, followed by examples of workflow applications.

Make Predictions on New Data - Use the model developed to make predictions on new data. There are a variety of reasons users may opt to apply the optimal model to new data. Users may simply want to gain insight into how well the model performs if fed different attributes, such as those determined to be significantly different regardless of information gain. Users may also want to gain further insight into the impact of attributes associated with a certain sub-grouping or functionality (e.g., gene function) on the performance of the optimal model by including and/or excluding those attributes from the input. Alternatively, users may want to make predictions about a different population based upon the information learned from the optimal dataset. The application of the optimal model to make predictions on new data that is unsupervised is possible depending on the algorithm from which the optimal model was learned (i.e., of the four algorithms in the workflow, only MLP and RandomForest will be available with supervised and unsupervised data).

Pathway Analysis - If the data are of the molecular variety (e.g., gene expression), we suggest performing a pathway analysis using the optimal dataset further to elucidate the physiological processes underlying the observed differences between groups. A variety of tools are available to perform pathway analyses depending on omics data type(s). Huang et al. (2017), Misra et al. (2019), and Krassowski et al. (2020) provide comprehensive lists of tools, softwares, and/or databases that may be useful for the integrated analysis of omics data, several of which can be used for single-omics studies and enable or are associated with downstream pathway analysis functions/tools. Some pathway analysis tools may limit the number of attributes that can be included in input. If a reduced dataset still contains some number of attributes that exceeds this threshold, users may consider establishing a cutoff based upon the information gain-rank assigned to each attribute (e.g., Top 50 of ranked attributes from optimal dataset). This cutoff

point may be further informed by creating additional subsets with the removal of one or few attributes at a time and plotting performance as described above.

The findings from the pathway analysis may be used to help direct further investigations of the factors (i.e., attributes) that are the strongest drivers of the difference, such as biomarker identification or similar. As described in the example applications of this workflow below, the findings from studies have supported further hypotheses that have served as the basis of other cellular and molecular research (e.g., Douros et al., 2018).

Examples of Workflow Applications

This workflow has been used to analyze a wide variety of biological data in both published and unpublished research studies. These data include, but are not limited to, various omics datasets, climate data, fisheries modeling data, and so on. Consistent success in reducing the number of attributes and therefore data dimensionality has been achieved in each application of this workflow or modified iteration thereof, depending on project scope and data suitability. As such, researchers employing this workflow or other collaborative research efforts from a variety of fields have had success in deriving more directed, meaningful insights has demonstrated that this workflow lends itself to being a widely-utilized standard approach to reduction of large biological data.

Among the published and in preparation findings yielded from the application of this workflow (or variation thereof) include the following, which specifically highlight the use of this approach to gain meaningful insights from large omics data:

Reading et al. (2013) used the SMO SVM to validate an unsupervised learning technique (k-means clustering) applied to characterize stages of striped bass (*Morone saxatilis*) oocyte

growth based upon proteomic profiles. Schilling et al. (2014) used the SMO SVM to accurately characterize the cytosolic and membrane fractions of white perch (*Morone americana*) ovary tissues based upon the expression of 242 measured proteins, providing further information of ovarian protein profiles during oogenesis and early embryogenesis. Schilling et al. (2015) employed the SMO SVM to model the proteomic differences of sexually mature male and female white perch (*Morone americana*) before and after exposure (via injection) to an endocrine disrupting compound, 17 β -Estradiol (E2). The SVMs were able to classify samples between sex and treatment with 100.0 % accuracy when the 104 proteins quantified via Nanoscale liquid chromatography coupled to tandem mass spectrometry (nanoLC-MS/MS) (Schilling et al., 2015). Chapman et al. (2014) applied ANNs to develop a 233-gene transcriptomic fingerprint predictive of quality and survival of striped bass (*Morone saxatilis*) eggs. Notably, none of the genes identified by Chapman et al. (2014) were found to significantly differ ($\alpha = 0.05$) when analyzed using traditional statistical tests (ANOVA). Sullivan et al. (2015) reports conducting a study designed identically to Chapman et al. (2014) on wild and domesticated striped bass (*Morone saxatilis*) broodfish and found the ANNs predicted 78.0-91.0 % variation of egg quality. Sullivan et al. (2015) also calculated the information gain associated with each of the 233 most informative genes identified by Chapman et al. (2014) and found that the lowest entropy, or greatest information, was associated with genes characteristic of an “intermediate” state of egg quality, suggesting there may be a transition from poor to good egg quality, rather than either being an inherited, static occurrence. Douros et al. (2018) used the MLP ANN to model the effects of leptin on gene expression in the tilapia (*Oreochromis niloticus*) pituitary. Pathway analysis of the reduced transcriptome data (400 genes) led researchers to hypothesize that one function of leptin is to stimulate glycolysis to meet energy

needs when responding to stress (Douros et al., 2018). Phillips et al. (2020) used this workflow (SMO and MLP) to reduce a the 16,000-gene transcriptomes of broiler chickens to the 450 genes most determinant of wooden breast myopathy (WBM, “woody breast”), a condition that negatively impacts meat quality, and ultimately characterize this disease as systemic rather than solely occurring to cellular dysfunction in the muscle alone. Giacomini et al. (BMC Genomics preprint) utilized the information gain approach and SMO SVM to identify genes underlying a positive immune response observed in sunflower-fed bumble bees (*Bombus impatiens*) compared to wildflower-fed bees.

Machine Learning Considerations and Challenges

There are numerous challenges inherent to or that may arise when conducting a ML analysis of big (or large) data. Challenges of applying ML techniques are compounded by limitations of ML tool accessibility (i.e., ease of use), computing power and algorithm scalability, and limitations based on what resources are available. The range of human expertise in applying algorithms and interpreting models also poses a challenge. Selecting the appropriate algorithm(s) for the specific biological question, data structure (data type and sample type, e.g., gene expression data are different than time series or spatial data) is of high import and with few environment-dependent exceptions (e.g., working in a data science research group), requires at least *some* human input, and therefore conceptual understanding at this time. Troubleshooting is anticipated as ML methods are incorporated across fields of biological research and in new ways, and as technologies that enable the generation of large biological data continue to advance. As such, it is also expected that educational references and materials will become more and more available, concepts and jargon will be more commonplace across fields of study, and that

evaluating analyses on the basis of being sound and replicable will be more easily completed by a greater number of scientists. Factors that will contribute to this end include the standardization of baseline approaches (*see*: workflow presented here), more common incorporation of ML into the scientist's toolkit, continued adaptation and development for different areas of research, and the presentation and publication of information.

Here we review some of the challenges and considerations of ML analyses, some of which are likely to be familiar based on their similar importance in completing traditional statistical analyses:

Metadata - To maximize comparability and leveraging potential of data and ML analysis results (i.e., compare between studies, approaches, and/or allow for the utilization of data in future studies that have not yet even been conceived) emphasis must be placed on the collection, curation, and publication/availability of metadata. Moreover, high quality (i.e., detailed), well-curated metadata enhances the extent to which a model can be interpreted and reasonably adapted to analyze and/or predict new data (garbage in, garbage out).

Scalability - The computational power required to conduct ML analyses is greatly influenced by factors such as the size and format of input data; framework for performing analyses (i.e., interface, language and/or software program); processors (CPUs, GPUs, TPUs, ASICs); model training based upon CV strategy (e.g., Holdout Method versus Stratified K-Fold CV); algorithm procedure or decision-making (e.g., feed forward versus recurrent neural network); and so on. An overview of possible solutions and resources that can be utilized to perform ML analyses when a standard computer is insufficient is provided in Mirza et al. (2019).

Sample Size - A critical design consideration for any study whether using traditional statistics or ML is sample size, or more specifically, determining the size of the subpopulation to be sampled that can be deemed representative of the population for the purpose of analysis and interpretation of results. In addition to the traditional statistical power analysis, a number of algorithms have been developed to determine the required sample size to appropriately perform machine learning analyses. For example, Figueroa et al. (2012) developed an algorithm to determine annotation sample size for supervised ML analysis and Dobbin et al. (2008) developed an online calculator for determining sample size necessary for classification with highly dimensional data. Other groups such as Vabalas et al. (2019) expand upon biases and other study design considerations that arise when working with small sample sizes.

Degrees of Freedom - In traditional statistics, degrees of freedom (df) refers to the number of independent variables that impact a statistic, or the number of values in the final calculation that can change. The power to reject a fall null hypothesis and find a significant result increases as df increases. In machine learning df more often refers to the number of parameters in a given model that are estimated (i.e., how many parameters must be estimated from the training data), the specific parameters varying slightly based on the algorithm in consideration. For example, df may be the number of training instances in SVMs or the number of terminal nodes (leaves) in a decision tree. The importance of df in ML is largely dependent on the algorithm(s) used. For example, df may be used when reporting a regression model. Alternatively, df have been applied as model selection criteria in deep neural networks (Gao and Jojic, 2016).

Missing Data - Similar to how missing data is accounted for in traditional statistical analyses, it must first be established if the data is missing at random, across all samples, or

otherwise (Mirza et al., 2019; Stevens et al., 2020). Researchers must make informed decisions regarding the address of missing data based upon these and other factors in order to mitigate bias and account for possible limitations of findings based upon the imputation and/or omission of data, the criteria for which must be clearly reported in either case (Mirza et al., 2019; Stevens et al., 2020).

Rarity and Class Imbalance - Class imbalance occurs in cases where one of the classes (groups) can be considered “rare” relative to the whole population (i.e., where the dataset has more instances that do *not* belong to a certain class than instances that *do* belong) (Mirza et al., 2019). Many ML algorithms assume a balanced class distribution whereby the number of instances from each class is approximately equal (in training) leading to an overestimation of the majority class and possible disregard of the minority class (Mirza et al., 2019). Class imbalance learning (CIL) methods include balancing the classes prior to analysis (“data sampling”); applying algorithms that are sensitive to the cost of incorrect predictions (“algorithm modification”); and utilizing an ensemble learning approach to improve generalization (Mirza et al., 2019).

Data Dimensionality - The curse of data dimensionality (CoD) refers to the challenges that come with analyzing datasets that include numerous attributes and in particular when the number of instances the attributes are measured from remain relatively low (i.e., more is not always better) (Bzdok et al., 2017). Among these challenges are interconnected issues of overfitting, sparsity, multicollinearity, and multiple testing (Altman & Krzywinski, 2018).

As the number of attributes increases the higher the likelihood that one or more of these attributes contain multiple zeros as the measured value and are therefore sparse (“sparsity”). The challenge of sparsity increases the space and time of a given model and results in data being less

representative to the greater population, as a larger sample size is needed to capture all possible combinations of attributes (Altman & Krzywinski, 2018). With sparsity comes overfitting, whereby an increased number of attributes may lead to the model being excessively complex and the output describes random error and “noise” rather than the actual underlying relationship (i.e., specificity is high and generalization is low, high rate of false negatives) (Reading, 2018). When the number of attributes exceeds the number of instances, attributes and dimensions of the data are able to be expressed in terms of other attributes and/or dimensions, leading to multiple correlations and muddying the impact (significance) of each attribute independently (i.e., multicollinearity) (Altman & Krzywinski, 2018). As with traditional statistics, testing whether or not each variable attribute impacts the response (i.e., multiple testing) can lead to a high rate of false positives, often corrected by controls for false discovery rate (FDR), however, this in turn may lead to high false negative rates (Altman & Krzywinski, 2018). Conversely, when too few attributes are included in the dataset the model becomes too simplistic and the output cannot be considered representative of the population (i.e., specificity is low and generalization is high) (Reading, 2018). This occurrence is also referred to as underfitting and leads to models having a high rate of false positives.

Heterogeneous Data - The challenge of data heterogeneity is especially pertinent to applications of ML in multi-omics and systems biology studies where the scale, distribution, number of measured variables, modalities (types), etc. of data being analyzed in concert may vary greatly (Mirza et al., 2019; Xu & Jackson, 2019). Although the establishment of best practices remains, there are multiple approaches to addressing challenges that arise with heterogeneous data (Krassowski et al., 2020; Mirza et al., 2019; Wang, 2017; Xu & Jackson, 2019). One such approach is to use multiple kernel learning strategies, which calculate individual

matrices for each data type prior to merging them into a global model and therefore accounts for known and unknown heterogeneities among the data (Mirza et al., 2019). Other approaches for dealing with heterogeneous data are particularly well-demonstrated in multi-omics studies, as this challenge is inherent to the integration of multiple (three or more) omics data (e.g., genomics, transcriptomics, proteomics, metabolomics, epigenomics) likely generated via different methods and subsequently represented by different measurements of molecular signatures (Mirza et al., 2019; Xu & Jackson, 2019). As such, multiple tools, frameworks, and strategies facilitating integration of multi-omics data have been developed under the umbrella of “Machine Learning and System Genomics (MLSG) approaches, and can possibly serve as a template for other issues of heterogeneous data analysis (Krassowski et al., 2020; Lin & Lane, 2017; Misra et al., 2019).

Conclusions

The application of ML techniques and approaches across the biological sciences is a necessity for advancements in research to keep pace with the ever-developing technologies that facilitate it as well as the challenges society is currently and will be facing in the future. Great successes of applying ML methods to biological research have been demonstrated and are continuing in effort and energy, however, the gap between those doing so and those researchers that do not will widen in part due to limitations of computational literacy, computing power, and interpretability of findings. To address this gap we provided an overview of ML and a supervised ML workflow that allows for a comprehensive yet flexible reduction of large biological datasets. The flexibility of the workflow in particular allows it to be applied to multiple data types to identify only those attributes (variables) most important to the classification of instances

(samples, individuals) into user-defined groups. Therefore, this workflow can be considered a baseline approach for reducing data dimensionality across biological (and other) research studies that can become a standard in reducing data dimensionality to improve interpretation, reusability, and leveraging of data and findings from ML research.

Acknowledgements

In addition to the published examples of utilizing the workflow presented here, this approach has been used to analyze research data generated for thesis projects as a component of a graduate-level course titled “Machine Learning Approaches in Biological Sciences” (AEC 510) taught at North Carolina State University (Raleigh, NC, USA) by B.J.R. This work was supported by funding provided from the following sources: The Foundation for Food and Agriculture Research (FFAR) *New Innovator Award*, the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), the National Oceanic and Atmospheric Administration (NOAA) and National Sea Grant (E/2019-AQUA-02, a project to establish a Striped Bass Aquaculture Hub, the *StriperHub*), and North Carolina Sea Grant. Striped bass is a priority species for the USDA National Research Support Project 8 (NRSP-8; *National Animal Genome Research Project*) and funding to support the *National Program for Genetic Improvement and Selective Breeding for the Hybrid Striped Bass Industry* was provided by the NRSP-8 and the USDA NIFA (Hatch Multistate Project), the USDA Agricultural Research Service (ARS, Harry K. Dupree Stuttgart National Aquaculture Research Center), the North Carolina State University College of Agriculture and Life Sciences and College of Sciences, and the North Carolina Agricultural Foundation William White Endowment. L.K.A. received funding support from the Center for Environmental Farming Systems Graduate

Fellowship Program, North Carolina State University Biotechnology Program (BIT), and the Coastal Conservation Association of North Carolina David and Ann Speaks Coastal Conservation Association Scholarship.

References

- Altman, N. & Krzywinski, M. (2018). The curse(s) of dimensionality. *Nature Methods*, 15(6). DOI: 10.1038/s41592-018-0019-x.
- Anderson, D.R., Burnham, K.P., & Thompson, W.L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *The Journal of Wildlife Management*, pp.912-923. DOI: 10.2307/3803199.
- Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(1). DOI: 10.1038/ncomms5308.
- Bashura, J., Burke, M., Lapitan, R., Mikulec, J., Owens, T., Reading, B.J., Richt, J.A., Spencer, D., Valdivia-Granda, W., Weekes, J., Wittrock, M., & Wright, D. (2021). Threats to Food and Agriculture Resources. United States, Department of Homeland Security, Office of Intelligence and Analysis (I&A) Public-Private Analytic Exchange Program (AEP). Retrieved: https://www.dhs.gov/sites/default/files/publications/threats_to_food_and_agriculture_resources.pdf.
- Bergandi, D. & Blandin, P. (1998). Holism vs. Reductionism: Do Ecosystem Ecology and Landscape Ecology Clarify the Debate? *Acta Biotheoretica*, 46(3), pp.185-206. DOI: 10.1023/A:1001716624350.
- Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3). DOI: 10.1214/ss/1009213726.
- Bzdok, D., Krzywinski, M., & Altman, N. (2017). Machine learning: a primer. *Nature methods*, 14(12), p.1119. DOI: 10.1038/nmeth.4526.
- Chapman, R.W., Reading, B.J., & Sullivan, C.V. (2014). Ovary Transcriptome Profiling via Artificial Intelligence Reveals a Transcriptomic Fingerprint Predicting Egg Quality in

- Striped Bass, *Morone saxatilis*. PLoS One, 9(5), p.e96818. DOI: 10.1371/journal.pone.0096818.
- Ching, T., Himmelstein, D., Beaulieu-Jones, B., Kalinin, A., Do, B., & Way, G. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface, 15(141), 20170387. DOI: 10.1098/rsif.2017.0387.
- Dobbin, K., Zhao, Y., & Simon, R. (2008). How Large a Training Set is Needed to Develop a Classifier for Microarray Data?. Clinical Cancer Research, 14(1), 108-114. DOI: 10.1158/1078-0432.ccr-07-0443.
- Douros, J.D., Baltzegar, D.A., Reading, B.J., Seale, A.P., Lerner, D.T., Grau, E.G., & Borski, R.J. (2018). Leptin stimulates cellular glycolysis through a STAT3 dependent mechanism in Tilapia. Frontiers in Endocrinology, 9. DOI: 10.3389/fendo.2018.00465.
- Ducharme, E.E. (2019). Understanding Striped Bass (*Morone saxatilis*) and Sunshine Hybrid Striped Bass (FEMALE *M. chrysops* x MALE *M. saxatilis*) Growth Using Metabolomic Analysis of Liver Tissues. North Carolina State University. MS Thesis. <https://repository.lib.ncsu.edu/handle/1840.20/36994>.
- Edwards, L. & Thiele, I. (2013). Applying systems biology methods to the study of human physiology in extreme environments. Extreme Physiology and Medicine, 2(1). DOI: 10.1186/2046-7648-2-8.
- Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. DOI: 10.1038/nature21056.
- Fang, F. & Casadevall, A. (2011). Reductionistic and Holistic Science. Infection And Immunity, 79(4), 1401-1404. DOI: 10.1128/iai.01343-10.

- Feng, X., Zhan, Y., Wang, Q., Yang, X., Yu, C., Wang, H., Tang, Z., Jiang, D., Peng, C., & He, Y. (2020). Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. *The Plant Journal*, 101(6), pp.1448-1461. DOI: 10.1111/tpj.14597.
- Figuroa, R., Zeng-Treitler, Q., Kandula, S., & Ngo, L. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1). DOI: 10.1186/1472-6947-12-8.
- Gao, T. & Jovic, V. (2016). Degrees of freedom in deep neural networks. *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI'16*, AUAI Press, Arlington, Virginia, United States, pp. 232-241. DOI: 10.48550/arXiv.1603.09260.
- Goh, G., Hodas, N., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal Of Computational Chemistry*, 38(16), 1291-1307. DOI: 10.1002/jcc.24764.
- Giacomini, J.J., Adler, L.S., Reading, B.J. & Irwin, R.E., 2021. Differential bumble bee gene expression associated with pathogen infection and pollen diet. DOI: 10.21203/rs.3.rs-912647/v1.
- Gilpin, W., Huang, Y., & Forger, D. (2020). Learning dynamics from large biological data sets: Machine learning meets systems biology. *Current Opinion in Systems Biology*, 22, 1-7. DOI: 10.1016/j.coisb.2020.07.009.
- Graw, S., Chappell, K., Washam, C., Gies, A., Bird, J., Robeson, M., & Byrum, S. (2021). Multi-omics data integration considerations and study design for biological systems and disease. *Molecular Omics*, 17(2), 170-185. DOI: 10.1039/d0mo00041h.

- Gunnarson, P., Mandralis, I., Novati, G., Koumoutsakos, P., & Dabiri, J. (2021). Learning efficient navigation in vortical flow fields. *Nature Communications*, 12(1). DOI: 10.1038/s41467-021-27015-y.
- Halsey, L.G. (2019). The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Biology Letters*, 15(5). DOI: 10.1098/rsbl.2019.0174.
- Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers In Genetics*, 8. DOI: 10.3389/fgene.2017.00084.
- Johnson, R.R., & Kubly, P. (2004). *Just the Essentials of Elementary Statistics* (9th ed.). Thomson Brooks/Cole.
- Jones, D. & Matloff, N., 1986. Statistical hypothesis testing in biology: a contradiction in terms. *Journal of Economic Entomology*, 79(5), pp.1156-1160. DOI: 10.1093/jee/79.5.1156.
- Karthikeyan, A., & Priyakumar, U.D. (2022). Artificial intelligence: machine learning for chemical sciences. *Journal Of Chemical Sciences*, 134(1). DOI: 10.1007/s12039-021-01995-2.
- Krassowski, M., Das, V., Sahu, S.K., & Misra, B.B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Frontiers In Genetics*, 11. DOI: 10.3389/fgene.2020.610798.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. DOI: 10.1038/nature14539.
- Lin, E. & Lane, H.-Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomarker Research*, 5(1). DOI: 10.1186/s40364-017-0082-y.

- McFadden, J. & Al-Khalili, J. (2018). The origins of quantum biology. *Proceedings of the Royal Society A*, 474(2220), p.20180674. DOI: 10.1098/rspa.2018.0674.
- Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramkteke, S. (2021). Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in The Life Sciences*, 1, 100010. DOI: 10.1016/j.ailsci.2021.100010.
- Mirza, B., Wang, W., Choi, H., Chung, N.C., & Ping, P. (2019). Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes*, 10(2). DOI: 10.3390/genes10020087.
- Mishra, B., Kumar, N. & Mukhtar, M.S., 2019. Systems biology and machine learning in plant–pathogen interactions. *Molecular Plant-Microbe Interactions*, 32(1), pp.45-55. DOI: 10.1094/MPMI-08-18-0221-FI.
- Misra, B.B., Langefeld, C., Olivier, M., & Cox, L.A. (2019). Integrated omics: tools, advances and future approaches. *Journal Of Molecular Endocrinology*, 62(1). DOI: 10.1530/jme-18-0055.
- Noor, E., Cherkaoui, S., & Sauer, U. (2019). Biological insights through omics data integration. *Current Opinion in Systems Biology*, 15, 39-47. DOI: 10.1016/j.coisb.2019.03.007.
- Nuzzo, R. (2014). Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487), pp.150-153. DOI: 10.1038/506150a.
- Outeiral, C., Strahm, M., Shi, J., Morris, G.M., Benjamin, S.C., & Deane, C.M., 2021. The prospects of quantum computing in computational molecular biology. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11(1), p.e1481. DOI: 10.1002/wcms.1481.

- Phillips, C.A., Reading, B.J., Livingston, M., Livingston, K., & Ashwell, C.M. (2020). Evaluation via supervised machine learning of the broiler pectoralis major and liver transcriptome in association with the muscle myopathy wooden breast. *Frontiers in Physiology*, 11. DOI: 10.3389/fphys.2020.00101.
- Quinn, F. (2012). A revolution in mathematics? What really happened a century ago and why it matters today. *Notices of the AMS*, 59(1), pp.31-37. DOI: 10.1090/noti787.
- Rajab, S.A.S. (2020). An Integrated Metabolomic and Transcriptomic Approach for Understanding White Muscle Growth Regulation in Hybrid Striped Bass Aquaculture. North Carolina State University. Doctoral Dissertation.
<https://repository.lib.ncsu.edu/handle/1840.20/38272>.
- Reading, B. (2018). Machine Learning Approaches in Biological Sciences [lecture notes]. North Carolina State University, AEC 592-012.
- Reading, B.J., Williams, V.N., Chapman, R.W., Williams, T.I., & Sullivan, C.V. (2013). Dynamics of the Striped Bass (*Morone saxatilis*) Ovary Proteome Reveal a Complex Network of the Translasome. *Journal of Proteome Research*, 12(4). DOI: 10.1021/pr3010293.
- Regenmortel, M.H.V. (2004). Reductionism and complexity in molecular biology: Scientists now have the tools to unravel biological complexity and overcome the limitations of reductionism. *EMBO reports*, 5(11), pp.1016-1020. DOI: 10.1038/sj.embor.7400284.
- Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3). DOI: 10.1007/s42979-021-00592-x.
- Schilling, J., Nepomuceno, A.I., Planchart, A., Yoder, J.A., Kelly, R.M., Muddiman, D.C., Daniels, H.V., Hiramatsu, N., & Reading, B.J. (2015). Machine learning reveals sex-

- specific 17 β -estradiol-responsive expression patterns in white perch (*Morone americana*) plasma proteins. *Proteomics*, 15(5). DOI: 10.1002/pmic.201400606.
- Schilling, J., Nepomuceno, A.I., Schaff, J.E., Muddiman, D.C., Daniels, H.V., & Reading, B.J. (2014). Compartment proteomics analysis of white perch (*Morone americana*) ovary using support vector machines. *Journal of Proteome Research*, 13(3). DOI: 10.1021/pr401067g.
- Schrider, D. & Kern, A. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends In Genetics*, 34(4), 301-312. DOI: 10.1016/j.tig.2017.12.005.
- Selvarajoo, K. (2021). The need for integrated systems biology approaches for biotechnological applications. *Biotechnology Notes*, 2, 39-43. DOI: 10.1016/j.biotno.2021.08.002.
- Silva, J., Teixeira, R., Silva, F., Brommonschenkel, S., & Fontes, E. (2019). Machine learning approaches and their current application in plant molecular biology: A systematic review. *Plant Science*, 284, 37-47. DOI: 10.1016/j.plantsci.2019.03.020.
- Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).
- Smith, R.J. (2018). The continuing misuse of null hypothesis significance testing in biological anthropology. *American Journal of Physical Anthropology*, 166(1), pp.236-245. DOI: 10.1002/ajpa.23399.
- Stone, M. (1961). The revolution in mathematics. *The American Mathematical Monthly*, 68(8), pp.715-734. DOI: 10.2307/2311976.

- Stevens, L.M., Mortazavi, B.J., Deo, R.C., Curtis, L., & Kao, D.P. (2020). Recommendations for Reporting Machine Learning Analyses in Clinical Research. *Circulation: Cardiovascular Quality And Outcomes*, 13(10). DOI: 10.1161/CIRCOUTCOMES.120.006556.
- Sullivan, C.V., Chapman, R.W., Reading, B.J., & Anderson, P.E. (2015). Transcriptomics of mRNA and egg quality in farmed fish: Some recent developments and future directions. *General and Comparative Endocrinology*, 221. DOI: 10.1016/j.ygcen.2015.02.012.
- Tang, Y., Zhao, C., Wang, J., Zhang, C., Sun, Q., Zheng, W.X., Du, W., Qian, F. and Kurths, J. (2022). Perception and Navigation in Autonomous Systems in the Era of Learning: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*. DOI: 10.1109/TNNLS.2022.3167688.
- Tavassoly, I., Goldfarb, J., & Iyengar, R. (2018). Systems biology primer: the basic methods and approaches. *Essays in Biochemistry*, 62(4), pp.487-500. DOI: 10.1042/EBC20180003.
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. (2019). Machine learning algorithm validation with a limited sample size. *PLoS One*, 14(11), e0224365. DOI: 10.1371/journal.pone.0224365.
- Villoutreix, P. (2021). What machine learning can do for developmental biology. *Development*, 148(1), p.dev188474. DOI: 10.1242/dev.188474.
- von Bertalanffy, L. (1968). *General system theory: Foundations, development, applications*. G. Braziller.
- Wang, L. (2017). Heterogeneous data and big data analytics. *Automatic Control and Information Sciences*, 3(1). DOI: 10.12691/acis-3-1-3.
- Wasserstein, R.L. & Lazar, N.A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2). DOI: 10.1080/00031305.2016.1154108.

- Wen, Y., Li, M., Si, J., & Huang, H. (2020). Wearer-prosthesis interaction for symmetrical gait: a study enabled by reinforcement learning prosthesis control. *IEEE transactions on neural systems and rehabilitation engineering*, 28(4), pp.904-913. DOI: 10.1109/TNSRE.2020.2979033.
- Witten, I.H., Frank, E., & Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA.
- Xia, Z., Wu, L., Zhou, X., & Wong, S. (2010). Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology*, 4(S2). DOI: 10.1186/1752-0509-4-s2-s6.
- Xu, C. & Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biology*, 20(1). DOI: 10.1186/s13059-019-1689-0.
- Younas, M. (2019). Research challenges of big data. *Service Oriented Computing And Applications*, 13(2). DOI: 10.1007/s11761-019-00265-x.
- Zampieri, G., Vijayakumar, S., Yaneske, E., & Angione, C. (2019). Machine and deep learning meet genome-scale metabolic modeling. *PLoS Computational Biology*, 15(7), e1007084. DOI: 10.1371/journal.pcbi.1007084.

Table 4.1. Definitions of common machine learning (ML) terminology. Comparable terms used to describe similar concepts in traditional statistics and/or biological research are provided in italics, if applicable. Defined terminology used in another provided definition are underlined.

ML Term	Definition
Algorithm <i>classifier</i>	The ML method (procedure) implemented to perform a given task
Attribute <i>feature, variable</i>	The items of data measured
Class <i>category</i>	One of a set of enumerated target values for a <u>label</u>
Cross-Validation (CV)	Techniques to evaluate model performance by providing benchmarks on multiple permutations of the complete input dataset
Dimension	The number of entries in a list of values that represent an <u>instance</u> entered into a model (i.e., number of entries in a feature vector)
Entropy <i>Shannon's Entropy</i>	A value describing how unpredictable a probability distribution is, or how much information is associated with each <u>attribute</u>
Information Gain	A measure of information an <u>attribute</u> provides about a <u>class</u> based in part on <u>entropy</u> calculated for each attribute
Instance <i>sample, individual, replicate, observation</i>	One row of a dataset
Label <i>groups, treatments</i>	Identifiers of <u>instances</u> belonging to one of the designated <u>classes</u>
Model	The representation of what a ML system has learned from the training data as the result of an <u>algorithm</u>
Overfitting	A model that has poor predictive ability due to matching the training data too closely; a curse of dimensionality (i.e., too many <u>attributes</u>)
Test Set <i>testing data</i>	A subset of data independent from the training set used to evaluate model performance in determining outcomes
Training Set <i>training data</i>	A representative subset of the data used to train the algorithm to recognize features and patterns of the data (i.e., build the model)
Underfitting	A model that has poor predictive ability due to a lack of capturing the complexity of the training data (i.e., too few <u>attributes</u>)

Table 4.2. Examples of machine learning (ML) algorithms within the four broad ML categories: (1) Traditional (Classical) Learning, (2) Reinforcement Learning, (3) Neural Nets and Deep Learning, (4) Ensemble Learning, and subtype(s) thereof as applicable.

ML Category	Subtype(s)	Task (if applicable)	Algorithm Example(s)	
(1) Traditional Learning	Supervised	<i>Classification</i>	Naive Bayes	
			Decision tree	
				Logistic regression
				K-Nearest Neighbors (KNN or k-NN)
				Support-Vector Machine (SVM)
		<i>Regression</i>		Simple linear regression
				Multiple linear regression
			Polynomial regression	
			Least absolute shrinkage and selection operator (LASSO)	
Unsupervised	<i>Clustering</i>		K-means clustering	
			Meanshift	
			Density-based spatial clustering of applications with noise (DBSCAN)	
	<i>Association</i>		Apriori	
			FP-growth (frequent pattern tree)	
			Equivalence Class Clustering and bottom-up Lattice Traversal (ECLAT)	
	<i>Dimensionality Reduction</i>		Principal component analysis (PCA)	
			Singular value decomposition (SVD)	
			Latent dirichlet allocation (LDA)	
			Latent semantic analysis (LSA, pLSA, GLSA)	
			t-distributed Stochastic Neighbor Embedding (t-SNE)	
	Semi-Supervised		<i>user choice / approach-based</i>	
(2) Reinforcement Learning	Value-Based		Q-Learning	
			Deep Q network	
	Policy-Based		Proximal Policy Optimization (PPO)	
	Model-Based		Monte Carlo Tree Search (MCTS)	

Table 4.2. (continued).

(3) Neural Nets and Deep Learning Artificial Neural Nets (ANNs)	Multilayer Perceptron (MLP) Convolutional Neural Network (CNN, or ConvNet) Long Short-Term Memory Recurrent Networks (LSTM-RNN) GameGAN (generative adversarial network)
Other Deep Learning Architecture	Restricted Boltzmann machine (RBM) Deep belief network (DBN) Auto-encoders (AE)
(4) Ensemble Learning Bagging (Bootstrap Aggregation)	Random Forest (RF) Extra Trees
Stacking (Stacked Generalization)	Blending Super Ensemble
Boosting	AdaBoost Gradient Boosting XGBoost (Stochastic gradient boosting)

Table 4.3. Kappa statistic (left) and Area Under the Receiver Operating Characteristic Curve (AUROC) values (right) and corresponding prediction strength of machine learning (ML) algorithms. The Kappa statistic is a measure of precision, specifically “interrater reliability”, or agreement between raters. The AUROC measures algorithm performance by characterizing the relationship between correct outcomes (true positives and true negatives) and incorrect outcomes (false positives and false negatives). Prediction strength varies from “worse than random” to “Optimal”, where “worse than random” indicates accuracy of predictions made was worse than what could be expected based on random chance and “optimal” is a perfectly trained algorithm.

Kappa Statistic	Prediction Strength	AUROC Value	Prediction Strength
>0.00	worse than random	0.00 – 0.49	worse than random
0.00 – 0.20	poor	0.50	poor
0.21 – 0.40	fair	0.51 – 0.60	fair
0.41 – 0.60	moderate	0.61 – 0.70	moderate
0.61 – 0.80	good	0.71 – 0.80	good
0.81 – 0.99	very good	0.81 – 0.99	very good
1.00	optimal	1.00	optimal

Table 4.4. The percent correct classification of machine learning (ML) algorithms that can be predicted with the assumption of random probability (i.e., Law of Probability) based on the number of designated classes (groups) in an analyzed dataset.

Classes	Odds Against	Correct Classification Assuming Random Probability (%)
2	1:2	50.0 %
3	1:3	33.3 %
4	1:4	25.0 %
5	1:5	20.0 %
6	1:6	16.7 %
7	1:7	14.3 %
8	1:8	12.5 %
9	1:9	11.1 %
10	1:10	10.0 %

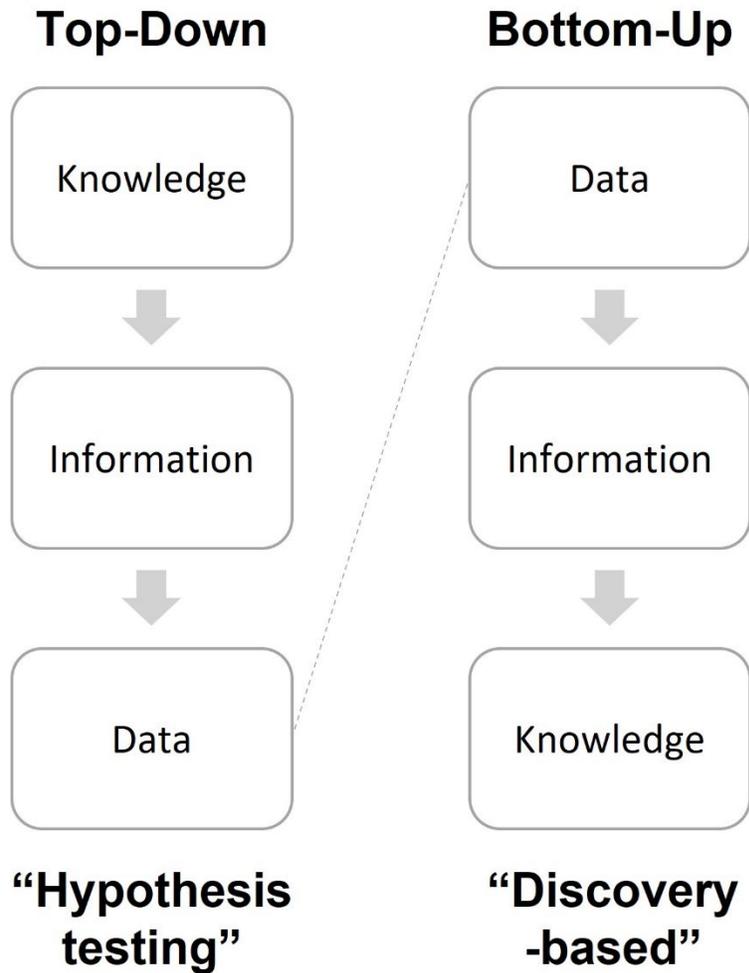


Figure 4.1. Top-down and bottom-up approaches to processing. The top-down approach is more commonly associated with principles of hypothesis testing and reductionism to draw specific conclusions from general knowledge. The bottom-up approach is more commonly associated with principles of discovery-based, holistic, and integrative methodologies that are data-driven whereby conclusions are based solely on what is represented by the data and little to no assumptions are made prior to. Top-down and bottom-up approaches to processing are not mutually exclusive; they are connected through the handling of data (represented by the dotted line) and are implemented in many procedures and processes beyond data analysis and research.

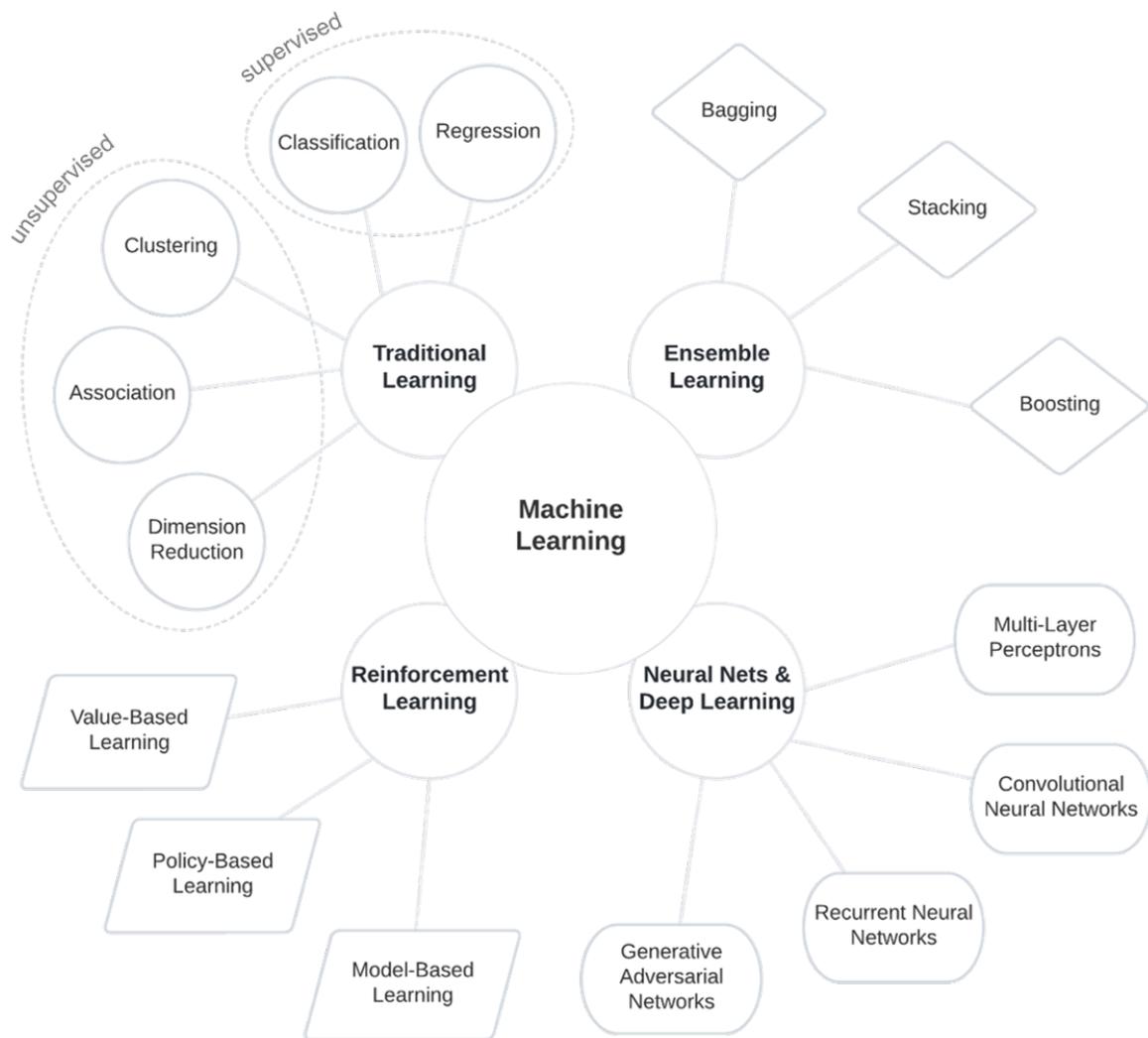


Figure 4.2. Four categories of machine learning (ML) and subtypes thereof. The dotted ovals encompassing subtypes of Traditional Learning, also referred to as Classical Learning, indicate which of these subtypes are often further described as being “supervised” or “unsupervised” depending on whether labels designating groupings (classes) of instances (e.g., samples) are or are not associated with input data, respectively.

Figure 4.3. (A) Workflow scheme of a supervised machine learning (ML) analysis. The outcome of applying the ML workflow is a dataset reduced in dimensionality such that only the attributes most determinant (i.e., those providing the most information) in the classification of instances (observations) into user-defined classes (groups) remain and a negative control for learning has been performed. Attributes are first reduced based upon Shannon’s Entropy and then again by determining the “optimal” number of attributes to yield superior classification by performing iterative runs of four orthogonal, twice cross-validated analytical algorithms on fewer and fewer attributes. Specifically, each attribute is assigned an information gain value based upon the amount of information provided to the classification of instances into classes and with higher values indicating greater information gain. Only those attributes with information gain above zero (or another user-selected cutoff) are included in the determination of the optimal dataset, which are those attributes of highest information gain value that account for (by avoiding) poor predictive ability based upon the inclusion of either too few or too many attributes (underfitting and overfitting, respectively). This optimal dataset is analysis ready (e.g., for pathway analysis).

(B) A description of applying this workflow to an example dataset of 10,000 genes identified among twenty instances (biological replicates) that are split equally into two classes, “Good” and “Bad” based upon some observable quality. The example graph depicts algorithm performance measured as percent correctly classified instances (y-axis) into classes with the exclusion of fewer attributes in each iterative run of the cross-validated algorithms from left to right (x-axis). The grey box represents the threshold of underfitting and overfitting as including the top 50 and top 500 attributes, respectively.

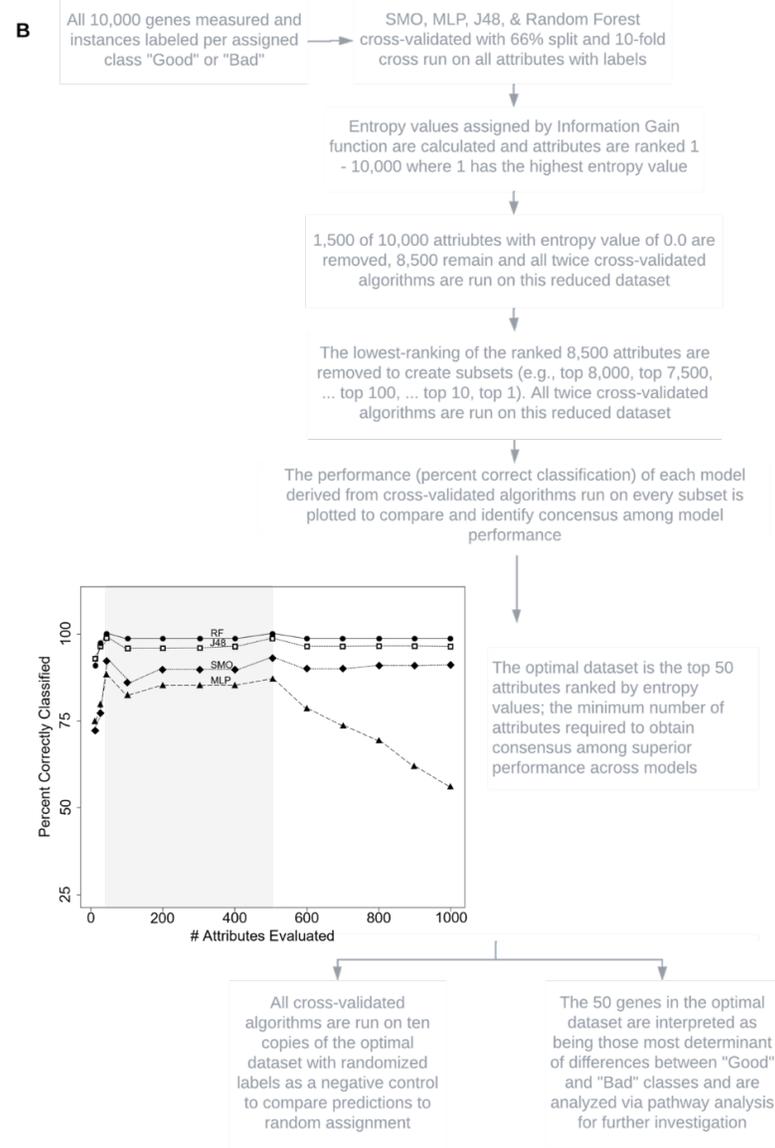
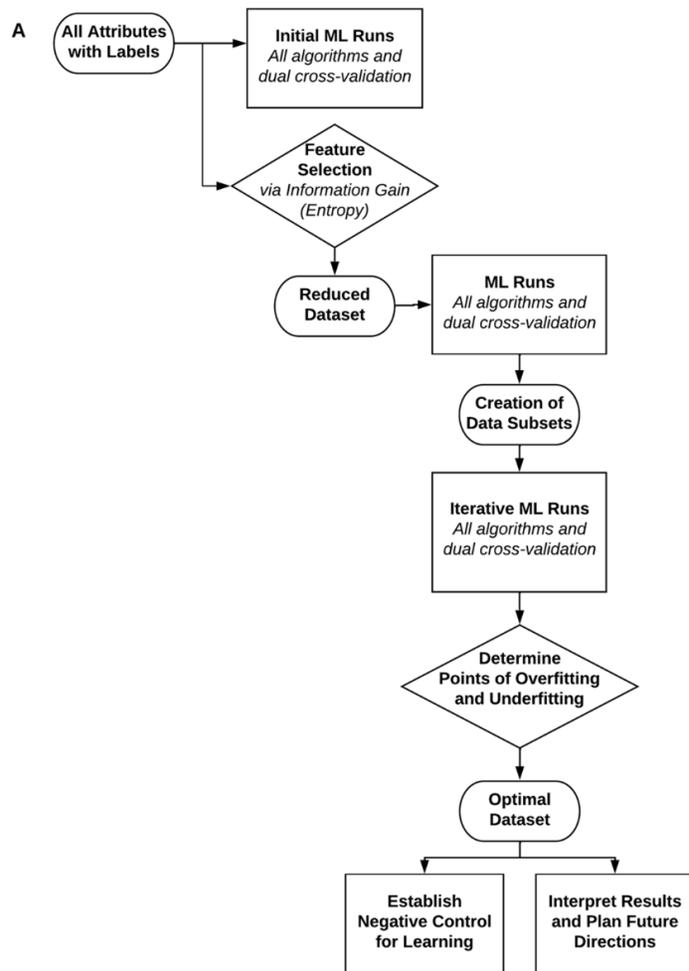


Figure 4.3.

	A	B	C	D	E	F	G	H
1	Sample ID	Quality	Gene 1	Gene 2	Gene 3	Gene 4	...	Gene 10000
2	Sample 1	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
3	Sample 2	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
4	Sample 3	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
5	Sample 4	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
6	Sample 5	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
7	Sample 6	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
8	Sample 7	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
9	Sample 8	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
10	Sample 9	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
11	Sample 10	Good	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
12	Sample 11	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
13	Sample 12	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
14	Sample 13	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
15	Sample 14	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
16	Sample 15	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
17	Sample 16	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
18	Sample 17	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
19	Sample 18	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
20	Sample 19	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!
21	Sample 20	Bad	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!

Attributes (points to columns C-H)

Instance (points to row 11)

Labels (points to column B)

Figure 4.4. Example layout of a data file formatted for Weka machine learning software (University of Waikato, New Zealand; Eibe et al., 2016, www.cs.waikato.ac.nz/ml/weka/).

CHAPTER 5. IMPLICATIONS TO RESEARCH OF STRIPED BASS AND THEIR HYBRID

Broadly, research efforts of aquaculture research are made with the long-term goal of supporting and enhancing the operational productivity of the US aquaculture industry, and specifically the HSB and SB industry here, by characterizing biological and physiological factors of interest that will inform and develop methods towards ensuring reliable production yields.

In addition to the US aquaculture industry, the outcomes of the described research and similar studies across aquaculture species will contribute to the long-range improvement and sustainability of US agriculture and food systems as they are developed with the intent of progress towards addressing challenges of local, regional, national, and global importance. The overarching challenge this research will serve to address is the promotion of sustainable expansion of aquaculture practices, especially as we face supply chain and natural resource management challenges. Continued investment and effort into improving culture methods for established aquaculture species remains crucial as they may serve as a template for breeding and production programs that are still in development worldwide. The curation of comprehensive, accessible datasets is also important for investigating breeding and rearing strategies that align with consumer trends (e.g., organic, no hormone use, no antibiotics) and potential avenues for adapting agricultural production efforts to the changing climate by mitigating impacts on animals and reducing negative outputs of production. Moreover, efforts to improve agricultural research and production are intended to be encouraging to farmers, investors, legislators, and those interested in seeking employment or establishing a production facility/operation with the ultimate goal of increasing economic prosperity, opportunity, and entrepreneurship in rural,

coastal, and food-scared communities and regions. The continuation of research on aquaculture systems will increase the visibility of aquaculture as a sector of agriculture and promote the integration of groups concerned with the management of agriculture and fisheries practices. Below is a brief overview of the genomic targets identified through the research described in **Chapter 2** for the hybrid striped bass (HSB, female white bass, WB, *Morone chrysops* x *M. saxatilis*, striped bass, SB, male) and **Chapter 3** for SB and potential future research directions.

Research Directions: Metabolism and Muscle Growth in Fishes

Metabolism broadly refers to the sum of all biochemical processes that regulate the use of energy and growth in an organism, which can be divided into two opposing types: catabolism and anabolism. Catabolic processes involve the breaking down of energy stores in the forms of carbohydrates, lipids, and proteins to provide energy. Anabolic processes are the building up, or synthesis, of these energy stores and accumulation thereof in tissues.

White skeletal muscle accounts for approximately 92.0 % of muscle mass in fishes (approximately 60.0 % of total body mass) and is a primary store of protein and glycogen that can be rapidly broken down (catabolism) for energy (Baltzegar et al., 2014; Johnston et al., 2004). The plasticity of skeletal muscle is critical in fishes as it undergoes chronic contraction in swimming and in some species, such as wild striped bass (SB, *Morone saxatilis*) as they are anadromous (Morash et al., 2014). As such, skeletal muscle fibers must ensure adequate oxygen and molecule transport for ATP production (Morash et al., 2014). Muscle homeostasis in fishes has been linked to the expression of transcription factors such as peroxisome proliferator-activated receptors (PPARs), which stimulate lipid metabolism, and AMP-activated protein kinase (AMPK), which stimulate fatty acid oxidation and glucose uptake (Morash et al., 2014).

Fish growth is indeterminate and controlled via the endocrine system, most notably through the GH-IGF axis: pituitary growth hormone (GH) and insulin-like growth factor 1 (IGF-1), as well as insulin, thyroid hormones, and other steroids to varying extents. Specifically, GH and IGF stimulate protein synthesis and this is influenced by insulin and other nutrients, amino acids posited to be the most important signals for IGF production in the liver. IGF also stimulates cell division, muscle cell differentiation, and bone growth. Well-fed fish are in an anabolic state, whereby insulin promotes lipogenesis, glycogenesis, and protein synthesis.

In the context of white skeletal muscle growth, GH regulates the expression of several genes belonging to the IGF, myostatin, myogenic regulator factors (MRFs), and atrophy systems (Fuentes et al., 2013). IGF, controlled via the Janus kinase-signal transducer and activator of transcription (JAK/STAT) signaling pathway, then stimulates muscle cell proliferation, differentiation, and protein synthesis via mitogen-activated protein kinase/Extracellular signal-regulated kinase 1/2 (MAPK/ERK) and phosphatidylinositol-3-kinase/protein kinase B/target of rapamycin (PI3K/AKT/TOR) signaling pathways, the latter of which is also used to inhibit protein degradation and atrophy (Fuentes et al., 2013).

As white skeletal muscle constitutes a major percentage of overall size and the major edible component of the fillet and therefore is the primary product of many aquaculture operations. As such, understanding the cellular mechanisms determinant of skeletal muscle growth is an important area of research in fishes, including how these growth processes interact and influence others beyond simply gene expression (i.e., along the endocrine axes, epigenetic modifications, etc.). The described study for HSB (**Chapter 2**) and SB (**Chapter 3**) identified genetic targets via machine learning (ML) analysis that are determinant of growth performance in HSB and SB, and therefore can be considered biomarkers of these traits and utilized in

selective breeding and or intervention/modification via biotechnological tools (e.g., CRISPR) in order to conduct functional genomics studies and, ultimately, leverage this information to produce a superior aquaculture product.

The transcriptomic profiles of HSB and SB indicate that the superior growth phenotypes of both HSB and SB are largely determined by the optimal functioning of metabolic processes (**Chapters 2 and 3**). Although proper metabolic function underlying growth performance is seemingly obvious, these results indicate that there is measurable dysfunction in the metabolic processing of HSB and SB that are exhibiting *inferior* growth phenotypes, and this is indicated distinctly in SB of this phenotype (referred to as “Inferior”, “Under Size” , and “Inferior Other” in **Chapter 3**). More specifically, the findings in HSB and SB indicate that the metabolic dysfunction in fish of inferior growth phenotypes is influenced by several genes and therefore the actions (or inaction, if inhibited) of many proteins and cellular pathways (e.g., signaling, activation), that this dysfunction is persistent throughout the lifecycle of these fish to some extent, is potentially heritable, and in HSB potentially heritable with key underlying genes being mitochondrially-encoded and therefore exclusively associated to the WB (maternal) parent. Specifically, lipid metabolism, oxidative phosphorylation, STAT3 and glycolysis (and JAK/STAT signaling) were of central focus in HSB and SB exhibiting superior growth traits. In SB exhibiting inferior growth traits, the most influential process seemed to be protein degradation influenced by the 26S proteasome and ubiquitin-proteasome system (UPS). A number of suggested approaches to broaden the understanding of mechanisms underlying growth in these fishes is provided below.

Numerous studies measuring tissue lipid content (typically liver and muscle) and lipid metabolism of HSB and SB either outright or as a result of dietary supplements, chronic stress

(hypoxia), and/or other modifications to rearing strategy have been conducted, including, but not limited to: Araújo et al. (2021), Beck et al. (2016), Hung et al. (1993), Gaylord and Gatlin (2000), Martin et al. (1984), Tønning et al. (2021), Twibell et al. (2000), Sealey et al. (2001), and Young and Cech (1994). Conducting additional research to examine genes specifically involved in lipid metabolism at juvenile (i.e., fingerling) age, such as those identified for HSB and SB in **Chapter 2** and **Chapter 3**, respectively, and how these gene expression profiles change overtime and/or after exposure to stressors would be particularly informative toward the determination of optimal stocking densities of these fish and in various systems (Liu et al., 2014). The catabolic state may be able to be mitigated in part through diet, and diet formulation studies that include efforts to target digestive health generally and in particularly lipid and/or carbohydrate uptake and clearing are of interest (Long et al., 2013; Xu et al., 2018).

There is evidence to suggest that the growth phenotype in HSB is underlain, in part, by mitochondrially-encoded and therefore maternally inherited genes, in particular those encoding members of the Mitochondrial complex I: NADH:ubiquinone oxidoreductase subunits (**Chapter 2**). Studies examining the gene expression of parents and offspring will be critical to make this determination and are worth examining in HSB as well as SB. Moreover, the unique MFs can be used for a more rapid and confident scanning of animal DNA at high resolution through a variety of gene expression measurement techniques including DNA sequencing, RNA sequencing, proteomic analyses, and microarrays. Individual MFs can be inserted in expression cassettes, host cells, and transgenic cells of aquatic animals to perform specific genomic enhancements in a highly directed manner. In addition to selective breeding, the DNA sequences that encode the MFs (and corresponding genes) can be used as targets for genome editing efforts. Researchers and producers of these fish can use these markers to determine if broodfish or offspring reared

for production are likely to display given phenotypic traits that may be desirable for production. Governments, prospective producers, and other stakeholder groups looking to establish aquaculture industries can use these MFs to screen for traits and evaluate hybrid striped bass and its parental fish as candidate species for culture based on the resources and external factors (e.g., climate suitability) of a specific area or region.

The advantage of the described MFs is the enablement of genomic-based breeding and screening efforts to produce a hybrid fish displaying the benefits of hybrid vigor, or heterosis. At present there are no established genetic markers associated with production traits for the hybrid striped bass or its parent species. Moreover, the described MFs are the first demonstration of elucidating patterns of expression between the parent species based on dominance (co-/over-) and presence of mitochondrial genes that are exclusively inherited from the maternal parent. Collectively these MFs provide a group of biomarkers that can be screened for to predict or ensure desirable production traits and therefore greatly reduce the amount of time previously required to identify and select such marks. The expression data of these MFs, those that are unilaterally expressed by only one of the parent species, also inform which parent species or sire strain should be used to promote the expression of genes underlying a given phenotypic trait through true-breeding or genetic modification.

Despite the multitude of genomic, transcriptomic, metabolomic, and proteomic data that have been generated for HSB, SB, and WB, there is no established method of reliably predicting the growth phenotype of the agricultural product (HSB) based on genotype. Epigenetics, the emerging field of research on changes in gene expression that are *not* caused by alterations in genetic sequence, has been heralded by many researchers as the “bridge” between genotype and phenotype. Epigenetic profiles of these fish are presently limited to methylation and sperm

motility, and conducting research in such an area would serve to expand the catalog of functional elements identified for SB, WB, and HSB (Woods et al., 2018). It is anticipated that, when considered in concert with other omics data, epigenomics data will provide a far more thorough understanding of mechanisms that may affect phenotypic outcomes of selective breeding and may also inform targeted gene editing efforts. The presence of early methyl marks, for example, may be indicative of heritable elements and their plasticity from year to year and after fertilization. These data will increase the breadth of information available for comparison between the domestic and extant congeners to inform the effect(s) of domestication on evolution that may be valuable criteria in assessments of behavior and welfare of domestic animals. Examinations of DNA methylation, fiber recruitment patterns, and fillet composition of striated muscle tissue and predictions of growth phenotype by methylation profiles will also provide information towards improving meat production. Additionally, epigenetic studies will help identify potential window(s) of time at which the animals have increased vulnerability to epigenetic modifications relevant to growth performance, whether due to stress event(s) or genome activity.

For example, **Figure 5.1** shows that motif fingerprints (MFs, **Chapter 2**) associated to Lipocalin-2 (*lcn2*) (outlined in the horizontal black box) were clearly up-regulated (indicated by red) in HSB of large (LG) grade and ultimate growth performance and down-regulated (indicated by blue to yellow color scale) in HSB of small (SM) grade and ultimate growth performance within and between HSB of the same-family (vertical dashed lines). The primary roles of *lcn2* are in energy metabolism (e.g., lipid transport) and innate immunity, however, *lcn2* has also been connected to impairment of skeletal muscle regeneration (Rebalka et al., 2018) and its methylation status has been linked to cellular division and growth related to oncogenic processes

in human tissues (Dokum et al., 2008; Rodvold et al., 2012; Wang et al., 2014). The role of *lcn2* and the up-and down-regulation in HSB of differing growth performance mirrors findings of increased bile acid concentrations (indicative of metabolic syndrome) and inferior growth in poor-growing HSB (Ducharme, 2020). Thus, it is possible that an epigenetic modification such as DNA methylation is potentially underlying the down-regulation of *lcn2* in muscle tissue of SM HSB and that LG HSB have distinctly different patterns of methylation on this loci and hundreds of others that we have identified as predictive of grade, ultimate growth phenotype, or both in the prior research described here for HSB. To test this, the presence and conservation of Differentially Methylated Regions (DMRs) in WB and SB broodfish gametes over time (year to year) and in fertilized HSB embryos produced from them, as well as in muscle tissue of HSB exhibiting superior and inferior growth phenotype at critical time points of production and their correlation to physiological pathways (metabolic syndrome) could be evaluated via approaches such as Reduced Representation Bisulfite Sequencing (RRBS). The RRBS approach combined with appropriate application of methylated adapters (e.g., Solexa adapters, Illumina, San Diego, CA, USA) and bisulfite conversion steps will allow for the generation of methylation profiles of specific portions of the animal genomes (i.e., those corresponding to genes of interest), which has proven to be an effective means of analysis of methylation in fishes such as rainbow trout (*Oncorhynchus mykiss*) (Gavery et al., 2018) and threespine stickleback (*Gasterosteus aculeatus*) (Metzger & Schulte, 2017). Ultimately, methylation patterns or other epigenomic features identified may be instrumental in characterizing the relationship between growth, metabolic syndrome, and the plasticity of an epigenetic modification that underlies these processes throughout the HSB life cycle as well as SB.

References

- Araújo, B.C., Rodriguez, M., Honji, R.M., Rombenso, A.N., del Rio-Zaragoza, O.B., Cano, A., Tinajero, A., Mata-Sotres, J.A. & Viana, M.T. (2021). Arachidonic acid modulated lipid metabolism and improved productive performance of striped bass (*Morone saxatilis*) juvenile under sub-to optimal temperatures. *Aquaculture*, 530, p.735939. DOI: 10.1016/j.aquaculture.2020.735939.
- Baltzegar, D. A., Reading, B. J., Douros, J. D., & Borski, R. J. (2014). Role for leptin in promoting glucose mobilization during acute hyperosmotic stress in teleost fishes. *J Endocrinol*, 220(1), 61-72. DOI: 10.1530/JOE-13-0292.
- Beck, B.H., Fuller, S.A., Li, C., Green, B.W., Zhao, H., Rawles, S.D., Webster, C.D. & Peatman, E. (2016). Hepatic transcriptomic and metabolic responses of hybrid striped bass (*Morone saxatilis* × *Morone chrysops*) to acute and chronic hypoxic insult. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 18, pp.1-9. DOI: 10.1016/j.cbd.2016.01.005.
- Fuentes, E. N., Valdés, J. A., Molina, A., & Björnsson, B. T. (2013). Regulation of skeletal muscle growth in fish by the growth hormone–insulin-like growth factor system. *General and Comparative Endocrinology*, 192, 136-148. DOI: 10.1016/j.ygcen.2013.06.009.
- Gavery, M.R., Nichols, K.M., Goetz, G.W., Middleton, M.A. & Swanson, P. (2018). Characterization of Genetic and Epigenetic Variation in Sperm and Red Blood Cells from Adult Hatchery and Natural-Origin Steelhead, *Oncorhynchus mykiss*. *G3: Genes, Genomes, Genetics*, 8(11), pp.3723-3736. DOI: 10.1534/g3.118.200458.

- Gaylord, T.G. & Gatlin III, D.M. (2000). Dietary lipid level but not l-carnitine affects growth performance of hybrid striped bass (*Morone chrysops* × *M. saxatilis*). *Aquaculture*, 190(3-4), pp.237-246. DOI: 10.1016/S0044-8486(00)00404-X.
- Hung, S.S., Conte, F.S. & Hallen, E.F. (1993). Effects of feeding rates on growth, body composition and nutrient metabolism in striped bass (*Morone saxatilis*) fingerlings. *Aquaculture*, 112(4), pp.349-361. DOI: 10.1016/0044-8486(93)90395-F.
- Johnston, I. A. (1999). Muscle development and growth: potential implications for flesh quality in fish. *Aquaculture*, 177(1-4), 99-115. DOI: 10.1016/S0044-8486(99)00072-1.
- Johnston, I. A., Abercromby, M., Vieira, V. L., Sigursteindóttir, R. J., Kristjánsson, B. K., Sibthorpe, D., & Skúlason, S. (2004). Rapid evolution of muscle fibre number in post-glacial populations of Arctic charr *Salvelinus alpinus*. *Journal of Experimental Biology*, 207(25), 4343-4360. DOI: 10.1242/jeb.01292.
- Liu, S., Gao, G., Palti, Y., Cleveland, B.M., Weber, G.M. & Rexroad III, C.E. (2014). RNA-seq analysis of early hepatic response to handling and confinement stress in rainbow trout. *Plos one*, 9(2), p.e88492. DOI: 10.1371/journal.pone.0088492.
- Long, X., Wang Q., Han X., Zhang X., & Deng J. (2013). Research advances on the mechanism of growth-promoting effects of dietary cholesterol in soybean meal-based diets of fish. *J. Anhui Agr. Sci.* 2954–2955. DOI: 10.3969/j.issn.0517-6611.2013.07.054.
- Martin, F.D., Wright, D.A. & Means, J.C. (1984). Fatty acids and starvation in larval striped bass (*Morone saxatilis*). *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 77(4), pp.785-790. DOI: 10.1016/0305-0491(84)90313-4.
- Metzger, D.C. & Schulte, P.M. 2017. Persistent and plastic effects of temperature on DNA methylation across the genome of threespine stickleback (*Gasterosteus aculeatus*).

- Proceedings of the Royal Society B: Biological Sciences, 284(1864), p.20171667. DOI: 10.1098/rspb.2017.1667.
- Morash, A. J., Vanderveken, M., & McClelland, G. B. (2014). Muscle metabolic remodeling in response to endurance exercise in salmonids. *Frontiers in physiology*, 5, 452. DOI: 10.3389/fphys.2014.00452.
- Sealey, W.M., Craig, S.R. & Gatlin, D.M. (2001). Dietary cholesterol and lecithin have limited effects on growth and body composition of hybrid striped bass (*Morone chrysops* × *M. saxatilis*). *Aquaculture Nutrition*, 7(1), pp.25-31. DOI: 10.1046/j.1365-2095.2001.00159.x.
- Tonning, K.A., Budge, S.M. & Tyedmers, P. (2021). Overwinter Changes in the Lipid Profile of Young-of-the-Year Striped Bass (*Morone saxatilis*) in Freshwater Ponds. *Biomolecules*, 11(11), p.1678. DOI: 10.3390/biom11111678.
- Twibell, R.G., Watkins, B.A., Rogers, L. & Brown, P.B. (2000). Effects of dietary conjugated linoleic acids on hepatic and muscle lipids in hybrid striped bass. *Lipids*, 35(2), pp.155-161. DOI: 10.1007/BF02664765.
- Woods III L.C., Li Y., Ding Y., Liu J., Reading B.J., Fuller S.A., & Song J. (2018). DNA methylation profiles correlated to striped bass sperm fertility. *BMC Genomics* 19:244. DOI: 10.1186/s12864-018-4548-6.
- Xu, C., Li, E., Xu, Z., Su, Y., Lu, M., & Qin, J. Chen, L., & Wang, X. (2018). Growth and Stress Axis Responses to Dietary Cholesterol in Nile Tilapia (*Oreochromis niloticus*) in Brackish Water. *Frontiers In Physiology*, 9. DOI: 10.3389/fphys.2018.00254.
- Young, P.S., & Cech Jr, J.J. (1994). Effects of different exercise conditioning velocities on the energy reserves and swimming stress responses in young-of-the-year striped bass

(*Morone saxatilis*). Canadian Journal of Fisheries and Aquatic Sciences, 51(7), 1528-1534.

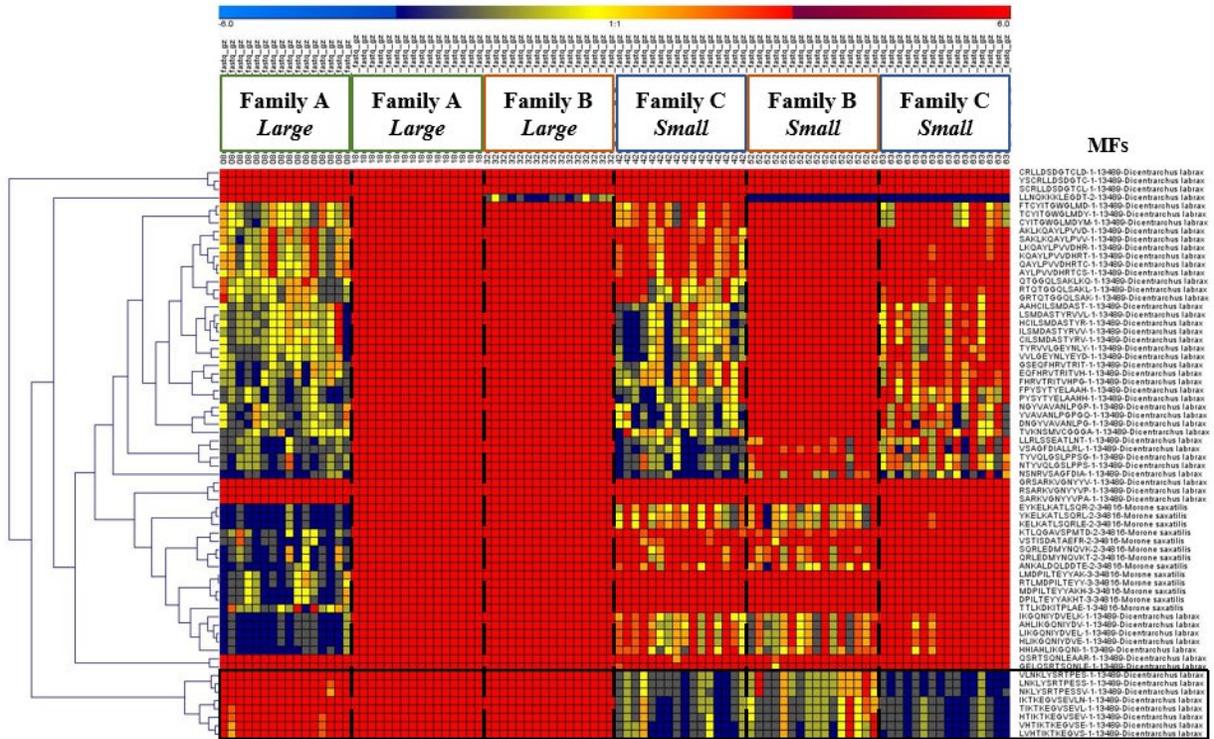


Figure 5.1. Hierarchical clustering heat-map of differences in expression of motif-fingerprints (MFs, protein fragments associated to specific genes) between hybrid striped bass (HSB) of different grade and ultimate growth performance (Large and Small) and from three different families (A through C). Up-regulation is represented in red and down-regulation is represented on the blue to yellow color scale. The eight MFs outlined in black at the bottom of the heat-map are specific to Lipocalin-2 (*lcn2*) a gene involved in energy metabolism, innate immunity, and skeletal muscle regeneration.

APPENDICES

Appendix A

Gene transcripts quantitated in striped bass (SB, *Morone saxatilis*) using CLC Genomics Workbench (v.21.0.3, Qiagen Inc., Hilden, Germany) RNA-Seq Analysis tool and reference genome sequence assembly for these fish published through National Institutes of Health National Center for Biotechnology Information (NCBI) GenBank® (NCSU_SB_2.0, accession: GCA_004916995.1). These transcripts were identified as the optimal dataset for the comparison of SB into groups based on growth performance, Superior and Inferior, via application of a machine learning (ML) pipeline whereby attributes (gene transcripts) are assigned an information gain value (Column: Info Gain), or the amount of information the inclusion of data associated to a given attribute provides the ML model for the classification of instances (individuals) into comparison classes (groups). Attributes were assigned a rank based upon the information gain value such that the first attribute (#1) has the highest information gain, or is the most informative to the classification of instances into groups (Column: Rank). Ranked attributes were then processed through four ML algorithms each twice cross-validated and whereby bottom-ranking attributes were eliminated in a recursive manner to determine points of improvement and degradation of model performance. The optimal dataset is that which yielded optimum performance of sorting individual SB into groups from as many cross-validated models as possible. SB gene and gene transcript information output from CLC Genomics Workbench is provided in columns: SB Gene, which is the SB gene symbol followed by an underscore and the transcript number; SB Gene Name; SB Gene ID, which is the NCBI ID number for a given gene. The IDs of orthologs of a given gene for zebrafish (*Danio rerio*) (Column: ZFIN ID) and humans (*Homo sapiens*) are also provided (Column: HGNC Symbol, HGNC ID). The log₂ fold change (Column: Log₂ FC) is the log₂ of the fold change values calculated as the maximum mean between the groups divided by the minimum mean, and multiplied by -1 if the maximum value was the second in order from left to right between observations. Statistical tests were performed as a component of the CLC Genomics Workbench Differential Expression for RNA-Seq analysis tool, which employs the Wald test (“All group pairs”) for comparisons of two groups. The false discovery rate-corrected p-value (Column: FDR P-Value) was used in subsequent pathway analyses.

SB Gene	SB Gene Name	SB Gene ID	ZFIN ID	HGNC Symbol	HGNC ID	Rank	Info Gain	Log ₂ FC	FDR P-Value
mrpl13_1	mitochondrial ribosomal protein L13	118336417	ZDB-GENE-050522-167	MRPL13	14278	1	0.3662	-0.06	0.95
asns_1	asparagine synthetase	118342428	ZDB-GENE-040426-1091	ASNS	753	2	0.3518	-0.46	8.72E-03
msh3_1	mutS homolog 3	118325540	ZDB-GENE-060526-307	MSH3	7326	3	0.344	-0.26	0.40
riox1_1	ribosomal oxygenase 1	118340642	ZDB-GENE-041008-221	RIOX1	20968	4	0.3191	-0.24	0.34
selenoo2_1	selenoprotein O2	118340647	ZFIN:ZDB-GENE-041210-110	SELENOO	30395	5	0.3132	-0.29	0.30
ccng2_1	cyclin G2	118332416	ZDB-GENE-021016-1	CCNG2	1593	6	0.313	0.25	0.53

Appendix A (continued).

acot11a_1	acyl-CoA thioesterase 11a	118335084	ZFIN:ZDB-GENE-050522-538	ACOT11	18156	7	0.3096	-0.32	0.92
dars1_1	aspartyl-tRNA synthetase 1	118338819	ZDB-GENE-061110-135	DARS1	2678	8	0.2965	6.09E-03	1.00
fut9a_1	fucosyltransferase 9a	118337648	ZFIN:ZDB-GENE-080723-75	FUT9	4020	9	0.2948	-0.64	0.09
ttc4_1	tetratricopeptide repeat domain 4	118322348	ZDB-GENE-040426-1021	TTC4	12394	10	0.2907	-0.22	0.44
st13_1	ST13 Hsp70 interacting protein	118323218	ZDB-GENE-030131-5395	ST13	11343	11	0.2896	-0.05	0.96
LOC118332666_1	eukaryotic translation initiation factor 4E-binding protein 1-like	118332666	-	EIF4EBP1	3288	12	0.2896	-0.3	0.54
mak16_1	MAK16 homolog	118325976	ZDB-GENE-020419-35	MAK16	13703	13	0.2884	-0.29	0.27
klhl7_1	kelch-like family member 7	118342724	ZDB-GENE-130212-1	KLHL7	15646	14	0.2868	-0.22	0.39
rnf11a_1	ring finger protein 11a	118327031	ZFIN:ZDB-GENE-040426-1277	RNF11	10056	15	0.2838	-0.21	0.44
bxdc2_1	brix domain containing 2	118332963	ZDB-GENE-060518-1	BRIX1	24170	16	0.2815	-0.33	0.31
LOC118328461_1	cytochrome c1, heme protein, mitochondrial	118328461	ZFIN:ZDB-GENE-031105-2	CYC1	2579	17	0.2803	0.07	0.92
LOC118342946_1	elongation factor 1-alpha-like	118342946	-	EEF1A1	3189	18	0.2769	-0.15	0.74
hars_1	histidyl-tRNA synthetase	118333619	ZDB-GENE-040912-152	HARS1	4816	19	0.2769	-0.11	0.81
LOC118327918_1	cytokine-inducible SH2-containing protein-like, mRNA	118327918	-	CISH	1984	20	0.2769	-0.65	3.52E-04
LOC118338989_1	cytochrome c-b	118338989	-	CYCS	19986	21	0.2692	-0.24	0.67
zpr1_1	ZPR1 zinc finger	118327934	ZDB-GENE-040426-2110	ZPR1	13051	22	0.2692	-0.29	0.29
lemd1_1	LEM domain containing 1	118324763	ZDB-GENE-040718-350	LEMD1	18725	23	0.2687	-0.2	0.63
LOC118332608_1	serine/threonine-protein phosphatase 6 catalytic subunit-like	118332608	-	PPP6C	9323	24	0.2687	-0.03	0.97

Appendix A (continued).

enpp6_1	ectonucleotide pyrophosphatase/phosphodiesterase 6	118333847	ZDB-GENE-031205-1	ENPP6	23409	25	0.2687	-0.38	0.11
polr1g_1	RNA polymerase I subunit G	118341334	ZDB-GENE-101202-2	POLR1G	24219	26	0.2649	-0.35	0.11
chchd4a_1	coiled-coil-helix-coiled-coil-helix domain containing 4a	118328695	ZFIN:ZDB-GENE-040801-46	CHCHD4	26467	27	0.2649	-0.27	0.25
LOC118335538_2	CTP synthase 1-like	118335538	-	CTPS1	2519	28	0.2643	-0.39	0.10
prmt3_1	protein arginine methyltransferase 3	118321154	ZDB-GENE-041105-1	PRMT3	30163	29	0.2643	-0.2	0.65
si:ch73-267c23.10_1	serine incorporator 1	118328318	ZFIN:ZDB-GENE-160114-59	SERINC3	11699	30	0.2643	-0.49	8.50E-04
grwd1_1	glutamate-rich WD repeat containing 1	118335870	ZDB-GENE-030131-9844	GRWD1	21270	31	0.262	-0.29	0.29
chac2_1	ChaC, cation transport regulator homolog 2	118329427	ZDB-GENE-050706-146	CHAC2	32363	32	0.262	-0.29	0.26
fgf2_1	fibroblast growth factor 2	118333801	ZDB-GENE-040318-1	FGF2	3676	33	0.262	-0.13	0.63
abitram_1	actin binding transcription modulator	118342785	ZDB-GENE-060503-602	ABITRAM	1364	34	0.262	-0.28	0.57
LOC118327556_1	transcription factor IIIA-like, mRNA	118327556	-	GTF3A	4662	35	0.262	-0.25	0.59
elac2_1	elaC ribonuclease Z 2	118323232	ZDB-GENE-041111-227	ELAC2	14198	36	0.2533	-0.23	0.32
zcchc4_1	zinc finger, CCHC domain containing 4	118338349	ZDB-GENE-080204-65	ZCCHC4	22917	37	0.2533	-0.2	0.58
LOC118329379_1	serine/threonine-protein kinase 32C-like	118329379	-	STK32C	21332	38	0.2533	-0.57	8.73E-03
LOC118335005_1	transcription factor HES-1-B-like	118335005	-	HES1	5192	39	0.2533	-0.86	9.24E-03
det1_2	DET1 partner of COP1 E3 ubiquitin ligase	118330094	ZDB-GENE-030131-2809	DET1	25477	40	0.2507	-0.64	0.75
rfoxp1_1	regulatory factor X-associated protein	118331395	ZDB-GENE-091204-321	RFXAP	9988	41	0.2484	-0.19	0.69
dusp11_1	dual specificity phosphatase 11 (RNA/RNP complex 1-interacting)	118333397	ZDB-GENE-050417-226	DUSP11	3066	42	0.2484	-0.44	0.50

Appendix A (continued).

usp14_1	ubiquitin specific peptidase 14 (tRNA-guanine transglycosylase)	118334598	ZDB-GENE-030131-7676	USP14	12612	43	0.2484	-0.26	0.30
ppid_1	peptidylprolyl isomerase D	118333298	ZDB-GENE-040625-34	PPID	9257	44	0.2452	-0.22	0.44
bud23_1	BUD23 rRNA methyltransferase and ribosome maturation factor	118341232	ZDB-GENE-070410-68	BUD23	16405	45	0.2452	-0.09	0.90
eif3ba_1	eukaryotic translation initiation factor 3, subunit Ba	118326373	ZFIN:ZDB-GENE-030131-2748	EIF3B	3280	46	0.2452	-0.1	0.87
ttl12_1	tubulin tyrosine ligase-like family, member 12	118341811	ZDB-GENE-040426-697	TTLL12	28974	47	0.2452	-0.32	0.12
LOC118325343_1	gamma-crystallin M2-like	118325343	-	CRYGC	2410	48	0.2452	-0.19	0.58
LOC118334992_1	nucleolin-like	118334992	-	NCL	7667	49	0.2426	-0.5	7.27E-03
gnl3_1	guanine nucleotide binding protein-like 3 (nucleolar)	118328734	ZDB-GENE-030131-616	GNL3	29931	50	0.2426	-0.07	0.94
ipo4_1	importin 4	118324296	ZDB-GENE-041014-307	IPO4	19426	51	0.2426	-0.56	2.54E-03
tff1.6_1	transcription termination factor 1, tandem duplicate 6	118325629	ZFIN:ZDB-GENE-041008-214	TTF1	12397	52	0.2426	-0.27	0.32
nol10_1	nucleolar protein 10	118337522	ZDB-GENE-040426-764	NOL10	25862	53	0.2426	-0.23	0.33
gpt2l_3	glutamic pyruvate transaminase (alanine aminotransferase) 2, like	118334633	ZFIN:ZDB-GENE-050302-11	GPT	4552	54	0.241	-0.21	0.47
LOC118333491_1	heterogeneous nuclear ribonucleoprotein A0-like	118333491	-	HNRNPA0	5030	55	0.241	-0.16	0.56
LOC118340847_1	26S proteasome regulatory subunit 4	118340847	-	PSMC1	9547	56	0.241	-0.24	0.33
angel1_1	angel homolog 1 (Drosophila)	118337362	ZDB-GENE-111020-12	ANGEL1	19961	57	0.241	-0.28	0.11
rpp25l_1	ribonuclease P/MRP 25 subunit-like	118336486	ZDB-GENE-040718-275	RPP25L	19909	58	0.241	-0.25	0.25
rrp12_1	ribosomal RNA processing 12 homolog	118329138	ZDB-GENE-050706-182	RRP12	29100	59	0.241	-0.3	0.19

Appendix A (continued).

c1qbp_1	complement component 1, q subcomponent binding protein	118331922	ZDB-GENE-050417-408	C1QBP	1243	60	0.2399	-0.11	0.83
rrs1_1	ribosome biogenesis regulator 1 homolog	118324391	ZDB-GENE-030131-9345	RRS1	17083	61	0.2399	-0.15	0.58
aimp2_1	aminoacyl tRNA synthetase complex interacting multifunctional protein 2	118322933	ZDB-GENE-040426-2652	AIMP2	20609	62	0.2377	-0.03	0.97
hsp90aa1.2_2	heat shock protein 90, alpha (cytosolic), class A member 1, tandem duplicate 2	118340391	ZFIN:ZDB-GENE-031001-3	HSP90AA1	5253	63	0.2377	-0.39	0.03
hnmt_1	histamine N-methyltransferase	118338631	ZDB-GENE-040801-157	HNMT	5028	64	0.2377	-0.61	0.08
qdptra_1	quinoid dihydropteridine reductase a	118333940	ZFIN:ZDB-GENE-070705-197	QDPR	9752	65	0.2377	-0.13	0.73
LOC118340060_1	cytochrome c oxidase assembly factor 6 homolog	118340060	ZFIN:ZDB-GENE-070705-152	COA6	18025	66	0.2377	-0.1	0.89
ube2g1b_1	ubiquitin-conjugating enzyme E2G 1b (UBC7 homolog)	118332502	ZFIN:ZDB-GENE-040426-2939	UBE2G1	12482	67	0.2377	-0.25	0.33
dynlrb2_1	dynein light chain roadblock-type 2	118330419	ZDB-GENE-120709-101	DYNLRB2	15467	68	0.2368	-0.48	0.24
psmd6_1	proteasome 26S subunit, non-ATPase 6	118328818	ZDB-GENE-040426-1038	PSMD6	9564	69	0.2368	-0.2	0.50
pwp1_1	PWP1 homolog, endonuclease	118342770	ZDB-GENE-040426-1049	PWP1	17015	70	0.2368	-0.23	0.25
klhl23_1	kelch-like family member 23	118339246	ZDB-GENE-081104-458	KLHL23	27506	71	0.2368	-0.35	0.13
LOC118337907_1	Y-box-binding protein 2-A-like	118337907	-	YBX2	17948	72	0.2358	-0.25	0.42
LOC118323545_1	deoxyhypusine synthase-like, transcript variant X3, mRNA	118323545	-	DHPS	2869	73	0.2358	-0.36	0.26
tsr1_1	TSR1 ribosome maturation factor	118341868	ZDB-GENE-030131-3762	TSR1	25542	74	0.231	-0.26	0.41
LOC118342612_1	tubulin--tyrosine ligase-like protein 12	118342612	-	TTLL12	28974	75	0.231	-0.33	0.20

Appendix A (continued).

herc4_1	HECT and RLD domain containing E3 ubiquitin protein ligase 4	118329204	ZDB-GENE-070801-1	HERC4	24521	76	0.231	-0.25	0.16
dus4l_1	dihydrouridine synthase 4-like	118341357	ZDB-GENE-040801-233	DUS4L	21517	77	0.2286	-0.18	0.83
kbtbd12_1	kelch repeat and BTB (POZ) domain containing 12	118328853	ZDB-GENE-050904-1	KBTBD12	25731	78	0.2286	-0.22	0.40
tp1b_1	triosephosphate isomerase 1b	118336118	ZFIN:ZDB-GENE-020416-4	TPI1	12009	79	0.2286	0.5	0.14
rsl1d1_1	ribosomal L1 domain containing 1	118335066	ZDB-GENE-130603-67	RSL1D1	24534	80	0.2227	0.04	0.97
dke1_1	dyskeratosis congenita 1, dyskerin	118337478	ZDB-GENE-031118-120	DKC1	2890	81	0.2227	-0.15	0.67
LOC118326533_1	multidrug and toxin extrusion protein 1-like, mRNA	118326533	-	SLC47A1	25588	82	0.2227	-0.22	0.34
hdhd5_1	haloacid dehalogenase like hydrolase domain containing 5	118330274	ZDB-GENE-080220-59	HDHD5	1843	83	0.2215	-0.24	0.18
ints11_1	integrator complex subunit 11	118324839	ZDB-GENE-050522-13	INTS11	26052	84	0.2215	-0.24	0.14
ntmt1_1	N-terminal Xaa-Pro-Lys N-methyltransferase 1	118332685	ZDB-GENE-040426-2055	NTMT1	23373	85	0.2215	-0.31	0.12
ddx52_1	DEAD (Asp-Glu-Ala-Asp) box polypeptide 52	118331797	ZDB-GENE-060623-1	DDX52	20038	86	0.2215	-0.15	0.68
pes_1	pescadillo	118326120	ZDB-GENE-990415-206	PES1	8848	87	0.2215	-0.19	0.55
ptcd3_1	pentatricopeptide repeat domain 3	118334089	ZDB-GENE-030131-6849	PTCD3	24717	88	0.2208	-0.07	0.90
scyl3_1	SCY1-like, kinase-like 3	118322100	ZDB-GENE-030131-2559	SCYL3	19285	89	0.2208	-0.45	0.08
psme3_1	proteasome activator subunit 3	118327615	ZDB-GENE-991110-19	PSME3	9570	90	0.2208	-0.18	0.57
polr3d_2	polymerase (RNA) III (DNA directed) polypeptide D	118325689	ZDB-GENE-040426-1017	POLR3D	1080	91	0.2208	-0.08	0.92
pdc11_3	programmed cell death 11	118326580	ZDB-GENE-030131-4076	PDCD11	13408	92	0.2208	-0.4	0.25

Appendix A (continued).

znf740b_3	zinc finger protein 740b	118324505	ZFIN:ZDB-GENE-060929-660	ZNF740	27465	93	0.2187	-1.35	0.38
rrp9_2	ribosomal RNA processing 9, U3 small nucleolar RNA binding protein	118328331	ZDB-GENE-060427-1	RRP9	16829	94	0.2172	-0.21	0.63
rpa3_1	replication protein A3	118343258	ZDB-GENE-040426-977	RPA3	10291	95	0.2172	-0.19	0.71
mrpl9_1	mitochondrial ribosomal protein L9	118342573	ZDB-GENE-070717-4	MRPL9	14277	96	0.2172	-0.13	0.83
bola1_1	bolA family member 1	118336514	ZDB-GENE-040801-76	BOLA1	24263	97	0.2172	-0.24	0.50
gpcpd1_5	glycerophosphocholine phosphodiesterase 1	118340803	ZDB-GENE-060503-472	GPCPD1	26957	98	0.2172	-0.43	0.25
spsb3b_2	sp1A/ryanodine receptor domain and SOCS box containing 3b	118331133	ZFIN:ZDB-GENE-121226-2	SPSB3	30629	99	0.217	-0.56	0.00
nop58_1	NOP58 ribonucleoprotein homolog	118340073	ZDB-GENE-040426-2140	NOP58	29926	100	0.217	-0.29	0.17
ccpg1_5	cell cycle progression 1	118343517	ZDB-GENE-041114-136	CCPG1	24227	101	0.217	-0.69	0.57
psmb5_1	proteasome 20S subunit beta 5	118334629	ZDB-GENE-990415-215	PSMB5	9542	102	0.217	-0.23	0.28
mtx2_1	metaxin 2	118329731	ZDB-GENE-021210-2	MTX2	7506	103	0.217	-0.2	0.50
flvcr2b_2	FLVCR heme transporter 2b	118340638	ZFIN:ZDB-GENE-041210-312	FLVCR2	20105	104	0.217	-0.35	0.19
cct8_2	chaperonin containing TCP1, subunit 8 (theta)	118331637	ZDB-GENE-040426-876	CCT8	1623	105	0.217	-0.08	0.91
LOC118327901_1	mitochondrial RNA pseudouridine synthase rpusd4-like	118327901	-	RPUSD4	25898	106	0.217	-0.31	0.28
cct8_1	chaperonin containing TCP1, subunit 8 (theta)	118331637	ZDB-GENE-040426-876	CCT8	1623	107	0.2134	-0.05	0.95
cltb_4	clathrin, light chain B	118333437	ZDB-GENE-101005-2	CLTB	2091	108	0.2134	-0.19	0.46
tarbp1_1	TAR (HIV-1) RNA binding protein 1	118329670	ZDB-GENE-090313-346	TARBP1	11568	109	0.2134	-0.47	0.07

Appendix A (continued).

LOC118342385_2	voltage-dependent calcium channel gamma-6 subunit-like	118342385	-	CACNG6	13625	110	0.2134	-0.43	0.09
ddx54_1	DEAD (Asp-Glu-Ala-Asp) box polypeptide 54	118332610	ZDB-GENE-021220-2	DDX54	20084	111	0.2134	-0.15	0.63
kat14_1	lysine acetyltransferase 14	118326661	ZDB-GENE-040718-452	KAT14	15904	112	0.2134	-0.17	0.60
wdr3_1	WD repeat domain 3	118339512	ZDB-GENE-030131-9830	WDR3	12755	113	0.2134	-0.15	0.68
LOC118322991_1	very-long-chain enoyl-CoA reductase-like	118322991	-	TECR	4551	114	0.2134	-0.3	0.12
wdr46_1	WD repeat domain 46	118324871	ZDB-GENE-040426-1264	WDR46	13923	115	0.2117	-0.36	0.05
ube3a_1	ubiquitin protein ligase E3A	118321655	ZDB-GENE-041114-190	UBE3A	12496	116	0.2117	-0.34	0.10
npepl1_1	aminopeptidase like 1	118328725	ZDB-GENE-050417-177	NPEPL1	16244	117	0.2099	0.01	0.99
aprt_1	adenine phosphoribosyltransferase	118329423	ZDB-GENE-040426-1492	APRT	626	118	0.2099	-0.24	0.55
LOC118336003_2	protein LYRIC-like	118336003	-	MTDH	29608	119	0.2099	-0.14	0.76
tmem41b_1	transmembrane protein 41B	118320697	ZDB-GENE-061215-138	TMEM41B	28948	120	0.2099	-0.33	0.02
rrp8_1	ribosomal RNA processing 8	118341308	ZDB-GENE-030131-8898	RRP8	29030	121	0.2099	-0.23	0.55
psmc4_1	proteasome 26S subunit, ATPase 4	118335944	ZDB-GENE-030131-5083	PSMC4	9551	122	0.2099	-0.19	0.59
trappc6bl_1	trafficking protein particle complex 6b-like	118341276	ZFIN:ZDB-GENE-040426-1602	TRAPPC6A	23069	123	0.2099	-0.21	0.33
prep_1	prolyl endopeptidase	118340874	ZDB-GENE-050522-14	PREP	9358	124	0.2099	-0.23	0.54
zgc:86609_1	TMCO1/EMC3 family protein	118324059	ZFIN:ZDB-GENE-050102-7	EMC3	23999	125	0.2099	-0.02	0.98
utp15_2	UTP15 small subunit processome component	118325671	ZDB-GENE-030131-3831	UTP15	25758	126	0.2099	-0.32	0.15
aadat_2	aminoadipate aminotransferase	118338366	ZDB-GENE-111229-1	AADAT	17929	127	0.2099	-0.28	0.10
rraga_1	Ras-related GTP binding A	118331987	ZDB-GENE-030131-4083	RRAGA	16963	128	0.2095	-0.08	0.89

Appendix A (continued).

cdh19_1	cadherin 19, type 2	118343284	ZDB-GENE-130530-766	CDH19	1758	129	0.2095	-0.25	0.37
blmh_1	bleomycin hydrolase	118342159	ZDB-GENE-030131-8485	BLMH	1059	130	0.2095	-0.13	0.79
ccnq_1	cyclin Q	118324755	ZDB-GENE-050522-495	CCNQ	28434	131	0.2095	-0.16	0.61
txnl1_1	thioredoxin-like 1	118332523	ZDB-GENE-040426-701	TXNL1	12436	132	0.2095	-0.11	0.83
slc38a10_1	solute carrier family 38 member 10	118327950	ZDB-GENE-050309-21	SLC38A10	28237	133	0.2095	0.67	0.22
parla_1	presenilin associated, rhomboid-like a	118334720	ZFIN:ZDB-GENE-050327-8	PARL	18253	134	0.2095	2.23E-03	1.00
LOC118330677_1	methionine-R-sulfoxide reductase B3-like	118330677	-	MSRB3	27375	135	0.2095	-0.23	0.29
skp1_1	S-phase kinase-associated protein 1	118331882	ZDB-GENE-040426-1707	SKP1	10899	136	0.2095	-0.16	0.64
LOC118329621_1	cytochrome c-like	118329621	-	CYCS	19986	137	0.2095	-0.23	0.48
cldn19_1	claudin 19	118328135	ZDB-GENE-050417-242	CLDN19	2040	138	0.2095	-0.4	0.04
zgc:153733_1	uncharacterized LOC118323678	118323678	ZFIN:ZDB-GENE-060825-273	PHETA1	26509	139	0.2095	-0.68	0.02
fbl_1	fibrillarlin	118335580	ZDB-GENE-040426-1936	FBL	3599	140	0.2016	-0.32	0.10
nhp2_1	NHP2 ribonucleoprotein homolog	118331471	ZDB-GENE-030131-533	NHP2	14377	141	0.2016	-0.27	0.41
dph2_1	diphthamide biosynthesis 2	118328022	ZDB-GENE-030219-100	DPH2	3004	142	0.2016	-0.27	0.32
naxe_1	NAD(P)HX epimerase	118343071	ZDB-GENE-040718-362	NAXE	18453	143	0.2016	-0.29	0.31
ruvbl2_1	RuvB-like AAA ATPase 2	118331119	ZDB-GENE-030109-1	RUVBL2	10475	144	0.2016	-0.27	0.46
LOC118320856_1	eukaryotic translation initiation factor 5A-1-like	118320856	ZDB-GENE-040426-2229	EIF5A	3300	145	0.2016	-0.02	0.98
LOC118340698_1	serine incorporator 1-like	118340698	-	SERINC1	13464	146	0.2016	-0.45	0.02
LOC118329129_1	dnaJ homolog subfamily C member 7-like	118329129	-	DNAJC7	12392	147	0.2016	-0.14	0.75
LOC118328132_1	kelch repeat and BTB domain-containing protein 13	118328132	ZDB-GENE-140106-122	KBTBD13	37227	148	0.2016	-0.2	0.48

Appendix A (continued).

LOC118321211_1	partitioning defective 3 homolog, mRNA	118321211	ZDB-GENE-170531-1	PARD3	16051	149	0.2016	0.56	0.04
fktn_1	fukutin	118332243	ZDB-GENE-070410-96	FKTN	3622	150	0.2016	-0.05	0.94
bysl_1	bystin-like	118328122	ZDB-GENE-040426-1287	BYSL	1157	151	0.2016	-0.16	0.66
strap_1	serine/threonine kinase receptor associated protein	118337952	ZDB-GENE-040426-1110	STRAP	30796	152	0.2014	-0.06	0.92
ppm1aa_2	protein phosphatase, Mg ²⁺ /Mn ²⁺ dependent, 1Aa	118337055	ZFIN:ZDB-GENE-991102-15	PPM1A	9275	153	0.2014	0.51	0.04
LOC118327173_1	forkhead box protein D1-like, mRNA	118327173	-	FOXD1	3802	154	0.2014	-0.33	0.75
tk2_1	thymidine kinase 2	118329813	ZDB-GENE-040718-243	TK2	11831	155	0.2014	-0.16	0.81
snx15_1	sorting nexin 15	118338371	ZDB-GENE-040426-1377	SNX15	14978	156	0.2014	-0.2	0.67
nup133_1	nucleoporin 133	118327465	ZDB-GENE-040426-2941	NUP133	18016	157	0.2014	-0.11	0.84
LOC118335643_1	carboxypeptidase Q-like	118335643	-	CPQ	16910	158	0.2014	-0.2	0.47
olfml2ba_1	olfactomedin-like 2Ba	118321986	ZFIN:ZDB-GENE-041014-329	OLFML2B	24558	159	0.2013	0.68	2.23E-03
eef1e1_1	eukaryotic translation elongation factor 1 epsilon 1	118341986	ZDB-GENE-030131-4949	EEF1E1	3212	160	0.2013	-0.11	0.84
hax1_1	HCLS1 associated protein X-1	118342995	ZDB-GENE-040718-26	HAX1	16915	161	0.2013	-0.27	0.15
LOC118340573_1	ankyrin repeat and SOCS box protein 2-like	118340573	-	ASB2	16012	162	0.2013	-0.43	0.05
eif2b3_1	eukaryotic translation initiation factor 2B, subunit 3 gamma	118335075	ZDB-GENE-040426-1039	EIF2B3	3259	163	0.2013	-0.11	0.82
LOC118338975_1	zinc-binding protein A33-like	118338975	-	TRIM27	9975	164	0.2013	-0.31	0.13
LOC118331977_1	fibroblast growth factor 13-like	118331977	-	FGF13	3670	165	0.1976	1.02	7.97E-03
LOC118332097_2	F-box only protein 40-like	118332097	-	FBXO40	29816	166	0.1976	-0.04	0.97
utp23_1	UTP23 small subunit processome component	118343036	ZDB-GENE-050417-353	UTP23	28224	167	0.1976	-0.15	0.81

Appendix A (continued).

LOC118340145_1	DDRGK domain-containing protein 1-like	118340145	-	DDRGK1	16110	168	0.1976	-0.13	0.59
gkap1_3	G kinase anchoring protein 1	118326161	ZDB-GENE-040426-2485	GKAP1	17496	169	0.1976	-0.23	0.58
LOC118327282_1	transcription elongation factor 1 homolog, mRNA	118327282	-	ELOF1	28691	170	0.1976	-0.21	0.56
LOC118322895_1	parvalbumin beta	118322895	ZDB-GENE-040625-48	PVALB	9704	171	0.1976	0.27	0.67
mrpl23_1	mitochondrial ribosomal protein L23	118330533	ZDB-GENE-040625-12	MRPL23	10322	172	0.1976	-0.13	0.72
LOC118326917_1	histone chaperone asf1b-B-like, mRNA	118326917	-	ASF1B	20996	173	0.1976	-0.3	0.44
six2a_1	SIX homeobox 2a	118329721	ZFIN:ZDB-GENE-010412-1	SIX2	10888	174	0.1957	-0.26	0.43
si:dkey-19b23.15_1	diamine acetyltransferase 2	118338323	ZFIN:ZDB-GENE-160728-17	SAT2	23160	175	0.1957	-0.89	8.50E-04
heatr3_1	HEAT repeat containing 3	118320872	ZDB-GENE-040426-1876	HEATR3	26087	176	0.1957	-0.31	0.06
LOC118340243_2	microtubule-associated protein RP/EB family member 3-like	118340243	-	MAPRE3	6892	177	0.1957	-0.14	0.70
epdr1_1	ependymin related 1	118334359	ZDB-GENE-040718-113	EPDR1	17572	178	0.1957	-0.12	0.71
dnajb12b_1	DnaJ heat shock protein family (Hsp40) member B12b	118334834	ZFIN:ZDB-GENE-070410-128	DNAJB12	14891	179	0.1957	-0.15	0.74
LOC118324273_4	rap1 GTPase-activating protein 1-like	118324273	-	RAP1GAP	9858	180	0.1944	0.97	0.65
rpf2_1	ribosome production factor 2 homolog	118341033	ZDB-GENE-040426-2501	RPF2	20870	181	0.1917	-0.13	0.86
wdr73_1	WD repeat domain 73	118336468	ZDB-GENE-050417-126	WDR73	25928	182	0.1917	-0.12	0.85
zak_3	sterile alpha motif and leucine zipper containing kinase AZK	118338786	ZDB-GENE-070912-386	MAP3K20	17797	183	0.1917	-0.97	8.50E-04
sdad1_1	SDA1 domain containing 1	118333011	ZDB-GENE-021213-1	SDAD1	25537	184	0.1917	-0.17	0.58
polr1b_1	RNA polymerase I subunit B	118329194	ZDB-GENE-040426-1598	POLR1B	20454	185	0.1917	-0.34	0.13

Appendix A (continued).

g3bp1_1	GTPase activating protein (SH3 domain) binding protein 1	118333368	ZDB-GENE-030131-7452	G3BP1	30292	186	0.1917	-0.37	0.03
ehbp111b_7	EH domain binding protein 1-like 1b	118331550	ZFIN:ZDB-GENE-091204-238	EHBP1L1	30682	187	0.1917	0.78	0.38
psma6a_1	proteasome 20S subunit alpha 6a	118337454	ZFIN:ZDB-GENE-020326-1	PSMA6	9535	188	0.1909	-0.19	0.51
stoml2_1	stomatin (EPB72)-like 2	118332940	ZDB-GENE-040426-1139	STOML2	14559	189	0.1909	-0.13	0.73
paip2b_2	poly(A) binding protein interacting protein 2B	118338031	ZDB-GENE-040426-1736	PAIP2B	29200	190	0.1909	-0.2	0.53
psmc2_1	proteasome 26S subunit, ATPase 2	118337318	ZDB-GENE-040426-1327	PSMC2	9548	191	0.1909	-0.18	0.54
znhit3_1	zinc finger, HIT-type containing 3	118331354	ZDB-GENE-040426-1114	ZNHIT3	12309	192	0.1909	-0.24	0.73
pfkfb4b_1	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4b	118328162	ZFIN:ZDB-GENE-031031-4	PFKFB4	8875	193	0.1909	-0.57	0.15
LOC118331364_2	core histone macro-H2A.1	118331364	ZFIN:ZDB-GENE-060421-4796	MACROH2A1	4740	194	0.1909	0.69	0.07
LOC118327907_2	WD repeat-containing protein 82-like	118327907	-	WDR82	28826	195	0.1909	-0.19	0.67
LOC118339996_1	beta-enolase	118339996	-	ENO3	3354	196	0.1909	0.66	0.03
LOC118329978_1	pseudouridylate synthase 7 homolog	118329978	-	PUS7	26033	197	0.1909	-0.18	0.58
npm1a_1	nucleophosmin 1a	118332112	ZFIN:ZDB-GENE-021028-1	NPM1	7910	198	0.1909	-0.13	0.78
LOC118330390_1	protein piccolo-like	118330390	-	PCLO	13406	199	0.1909	-1.56	0.08
LOC118330242_1	elongin-B-like	118330242	-	ELOB	11619	200	0.1909	-0.11	0.79
srm_1	spermidine synthase	118324894	ZDB-GENE-040426-1183	SRM	11296	201	0.1909	-0.32	0.19
mrpl28_1	mitochondrial ribosomal protein L28	118342233	ZDB-GENE-050522-113	MRPL28	14484	202	0.1909	-0.27	0.63
fbxl31_1	F-box and leucine-rich repeat protein 3, like	118333337	ZFIN:ZDB-GENE-130530-598	FBXL21P	13600	203	0.1909	-0.73	0.42
sdhaf2_1	succinate dehydrogenase complex assembly factor 2	118320774	ZDB-GENE-030131-7564	SDHAF2	26034	204	0.1909	-0.04	0.96
sdhaf3_1	succinate dehydrogenase complex assembly factor 3	118342430	ZDB-GENE-041010-94	SDHAF3	21752	205	0.1909	-0.15	0.82

Appendix A (continued).

zfyve27_2	zinc finger, FYVE domain containing 27	118326759	ZDB-GENE-061013-502	ZFYVE27	26559	206	0.19	-0.34	0.30
heatr1_1	HEAT repeat containing 1	118325688	ZDB-GENE-030131-6378	HEATR1	25517	207	0.19	-0.37	0.12
emc8_1	ER membrane protein complex subunit 8	118321148	ZDB-GENE-040426-2692	EMC8	7864	208	0.19	-0.2	0.58
naa15a_1	N-alpha-acetyltransferase 15, NatA auxiliary subunit a	118333569	ZFIN:ZDB-GENE-030131-2392	NAA15	30782	209	0.19	-0.09	0.87
LOC118322896_1	parvalbumin beta-1-like	118322896	ZDB-GENE-040625-48	PVALB	9704	210	0.19	-0.7	0.22
LOC118334619_1	MICOS complex subunit MIC26-like	118334619	-	APOO	28727	211	0.19	-0.13	0.79
asb14b_1	ankyrin repeat and SOCS box containing 14b	118324687	ZFIN:ZDB-GENE-090313-148	ASB14	19766	212	0.1895	-0.29	0.41
trmt61a_1	tRNA methyltransferase 61A	118336718	ZDB-GENE-040718-5	TRMT61A	23790	213	0.1895	-0.07	0.93
asl_1	argininosuccinate lyase	118332003	ZDB-GENE-040426-1152	ASL	746	214	0.1895	-0.03	0.97
dhodh_1	dihydroorotate dehydrogenase	118343269	ZDB-GENE-030131-3157	DHODH	2867	215	0.1895	-0.25	0.50
gtpbp6_1	GTP binding protein 6 (putative)	118321907	ZDB-GENE-031118-62	GTPBP6	30189	216	0.1895	-0.21	0.52
pum3_1	pumilio RNA-binding family member 3	118332424	ZDB-GENE-030131-9808	PUM3	29676	217	0.1895	-0.34	0.37
mrpl1_1	mitochondrial ribosomal protein L1	118325398	ZDB-GENE-060526-311	MRPL1	14275	218	0.1895	-0.09	0.89
atf4a_1	activating transcription factor 4a	118332414	ZFIN:ZDB-GENE-040426-2340	ATF4	786	219	0.1895	0.41	0.08
stip1_1	stress-induced phosphoprotein 1	118331123	ZDB-GENE-041121-17	STIP1	11387	220	0.1895	-0.12	0.79
psmd14_1	proteasome 26S subunit, non-ATPase 14	118321307	ZDB-GENE-070410-56	PSMD14	16889	221	0.1887	-0.3	0.21
trmt6_1	tRNA methyltransferase 6	118341018	ZDB-GENE-051120-141	TRMT6	20900	222	0.1887	-0.17	0.58
pabpc4_1	poly(A) binding protein, cytoplasmic 4 (inducible form)	118336838	ZDB-GENE-030131-9663	PABPC4	8557	223	0.1827	-0.02	0.98

Appendix A (continued).

gpatch4_1	G patch domain containing 4	118336440	ZDB-GENE-041008-158	GPATCH4	25982	224	0.1827	-0.14	0.78
wdyhv1_1	WDYHV motif containing 1	118342678	ZDB-GENE-060503-848	NTAQ1	25490	225	0.1827	-0.24	0.76
LOC118337496_1	activator of 90 kDa heat shock protein ATPase homolog 1-like	118337496	-	AHSA1	1189	226	0.1827	-0.23	0.22
glod4_1	glyoxalase domain containing 4	118341704	ZDB-GENE-040912-38	GLOD4	14111	227	0.1827	-0.31	0.09
polr3b_1	polymerase (RNA) III (DNA directed) polypeptide B	118330571	ZDB-GENE-030131-2887	POLR3B	30348	228	0.1827	-0.24	0.23
pprc1_1	PPARG related coactivator 1	118329453	ZDB-GENE-030131-9858	PPRC1	30025	229	0.1827	-0.5	0.07
antkmt_1	adenine nucleotide translocase lysine methyltransferase	118323424	ZDB-GENE-040625-59	ANTKMT	14152	230	0.1827	-0.38	0.27
rpp21_1	ribonuclease P 21 subunit	118321359	ZDB-GENE-040801-37	RPP21	21300	231	0.1827	-0.38	0.33
ufd11_1	ubiquitin recognition factor in ER associated degradation 1	118325955	ZDB-GENE-040718-150	UFD1	12520	232	0.1827	-0.04	0.95
sfxn2_3	sideroflexin 2	118330423	ZDB-GENE-040426-2831	SFXN2	16086	233	0.1827	-0.34	0.02
tegt_1	testis enhanced gene transcript (BAX inhibitor 1)	118324122	ZDB-GENE-030826-10	TMBIM6	11723	234	0.1827	-0.19	0.49
tmem182a_1	transmembrane protein 182a	118338943	ZFIN:ZDB-GENE-050306-22	TMEM182	26391	235	0.1827	-0.25	0.32
rab7a_1	RAB7a, member RAS oncogene family	118328706	ZDB-GENE-040426-1352	RAB7A	9788	236	0.1827	-0.15	0.55
klhl30_1	kelch-like family member 30	118321877	ZDB-GENE-100913-2	KLHL30	24770	237	0.1827	-0.35	0.05
adss2_1	adenylosuccinate synthase 2	118329134	ZDB-GENE-050417-337	ADSS2	292	238	0.1827	-0.34	0.05
slc26a5_1	solute carrier family 26 member 5	118337310	ZDB-GENE-030131-1566	SLC26A5	9359	239	0.1822	-0.43	0.05
LOC118343091_1	prolyl endopeptidase-like	118343091	-	PREPL	30228	240	0.1822	-0.11	0.82

Appendix A (continued).

abce1_1	ATP-binding cassette, sub-family E (OABP), member 1	118337917	ZDB-GENE-040426-1995	ABCE1	69	241	0.1822	-0.16	0.74
pop4_1	POP4 homolog, ribonuclease P/MRP subunit	118320814	ZDB-GENE-050522-505	POP4	30081	242	0.1822	-0.26	0.22
pnpo_1	pyridoxamine 5'-phosphate oxidase	118330031	ZDB-GENE-060602-2	PNPO	30260	243	0.1822	-0.12	0.73
psma8_1	proteasome 20S subunit alpha 8	118335132	ZDB-GENE-040426-2194	PSMA8	22985	244	0.1822	-0.17	0.65
LOC118340573_3	ankyrin repeat and SOCS box protein 2-like	118340573	-	ASB2	16012	245	0.1822	-0.35	0.22
prmt7_1	protein arginine methyltransferase 7	118330921	ZDB-GENE-040426-1560	PRMT7	25557	246	0.1822	-0.1	0.84
scube2_4	signal peptide, CUB domain, EGF-like 2	118321047	ZDB-GENE-050302-80	SCUBE2	30425	247	0.1822	1	0.03
hpert1_1	hypoxanthine phosphoribosyltransferase 1	118333750	ZDB-GENE-040426-1918	HPRT1	5157	248	0.1817	-0.18	0.63
cth_1	cystathionase (cystathionine gamma-lyase)	118321875	ZDB-GENE-030131-774	CTH	2501	249	0.1817	-0.17	0.63
spata5l1_1	spermatogenesis associated 5-like 1	118321321	ZDB-GENE-060929-204	SPATA5L1	28762	250	0.1817	-0.27	0.28
pcyt2_1	phosphate cytidylyltransferase 2, ethanolamine	118327034	ZDB-GENE-041010-132	PCYT2	8756	251	0.1817	-0.3	0.21
prmt2_1	protein arginine methyltransferase 2	118339068	ZDB-GENE-041104-1	PRMT2	5186	252	0.1817	-0.27	0.34
si:ch1073-358c10.1_1	mitochondrial peptide methionine sulfoxide reductase	118337601	ZFIN:ZDB-GENE-050309-123	MSRA	7377	253	0.1817	-0.24	0.62
tmem184c_1	transmembrane protein 184C	118327077	ZDB-GENE-090312-89	TMEM184C	25587	254	0.1817	-0.21	0.32
ccdc25_1	coiled-coil domain containing 25	118341092	ZDB-GENE-040426-1389	CCDC25	25591	255	0.1817	-0.13	0.69
txndc15_1	thioredoxin domain containing 15	118332115	ZDB-GENE-070615-36	TXNDC15	20652	256	0.1817	-0.03	0.97

Appendix A (continued).

LOC118330619_1	apoptosis-enhancing nuclease-like	118330619	-	AEN	25722	257	0.1817	-0.09	0.92
rars1_1	arginyl-tRNA synthetase 1	118331933	ZDB-GENE- 030131-9014	RARS1	9870	258	0.1817	-0.11	0.86
LOC118336496_1	papilin-like	118336496	-	PAPLN	19262	259	0.1817	0.68	7.66E-05
llph_1	LLP homolog, long-term synaptic facilitation factor	118339675	ZDB-GENE- 051030-30	LLPH	28229	260	0.1817	-0.04	0.97
aspg_2	asparaginase homolog	118336951	ZDB-GENE- 070820-14	ASPG	20123	261	0.1817	-0.34	0.12
pdha1a_1	pyruvate dehydrogenase E1 subunit alpha 1a	118331308	ZFIN:ZDB-GENE- 040426-2719	PDHA1	8806	262	0.1817	-0.08	0.90
LOC118334352_1	transmembrane protein 69- like	118334352	-	TMEM69	28035	263	0.1789	-0.51	0.49
rpl7l1_1	ribosomal protein L7-like 1	118341071	ZDB-GENE- 030131-970	RPL7L1	21370	264	0.1789	-0.09	0.88
ciao2b_1	cytosolic iron-sulfur assembly component 2B	118329793	ZDB-GENE- 040718-148	CIAO2B	24261	265	0.1789	-0.15	0.80
psmd13_1	proteasome 26S subunit, non-ATPase 13	118320668	ZDB-GENE- 040426-1004	PSMD13	9558	266	0.1789	-0.12	0.81
paip1_1	poly(A) binding protein interacting protein 1	118325701	ZDB-GENE- 040801-247	PAIP1	16945	267	0.1789	-0.25	0.15
alad_1	aminolevulinate dehydratase	118332570	ZDB-GENE- 050417-123	ALAD	395	268	0.1789	-0.04	0.96
LOC118324844_1	sodium-coupled neutral amino acid transporter 3- like, mRNA	118324844	-	SLC38A3	18044	269	0.1789	-0.35	0.21
rps27.2_1	ribosomal protein S27, isoform 2	118335976	ZFIN:ZDB-GENE- 040426-1735	RPS27	10416	270	0.1789	-0.08	0.93
LOC118334750_1	fizzy-related protein homolog	118334750	-	FZR1	24824	271	0.1789	-0.19	0.44
tmem70_1	transmembrane protein 70	118335171	ZDB-GENE- 070912-593	TMEM70	26050	272	0.1789	-0.16	0.72
chchd2_1	coiled-coil-helix-coiled- coil-helix domain containing 2	118341782	ZDB-GENE- 040426-1737	CHCHD2	21645	273	0.1789	-0.18	0.44
cunh9orf16_1	chromosome unknown C9orf16 homolog	118333008	ZFIN:ZDB-GENE- 111013-4	BBLN	17823	274	0.1789	-0.18	0.76
naif1_1	nuclear apoptosis inducing factor 1	118332253	ZDB-GENE- 081104-236	NAIF1	25446	275	0.1757	-0.41	0.30

Appendix A (continued).

tbl3_1	transducin beta like 3	118323115	ZDB-GENE-041114-104	TBL3	11587	276	0.1757	-0.18	0.54
dnajc30b_1	DnaJ (Hsp40) homolog, subfamily C, member 30b	118341500	ZFIN:ZDB-GENE-131121-532	DNAJC30	16410	277	0.1757	-0.13	0.74
umps_1	uridine monophosphate synthetase	118339177	ZDB-GENE-040426-785	UMPS	12563	278	0.1757	-0.44	0.01
LOC118328778_1	probable ATP-dependent RNA helicase DDX27	118328778	-	DDX27	15837	279	0.1757	-0.15	0.67
msrb3_1	methionine sulfoxide reductase B3	118338034	ZDB-GENE-040625-74	MSRB3	27375	280	0.1757	-0.27	0.37
polr2f_1	RNA polymerase II, I and III subunit F	118323693	ZDB-GENE-040718-279	POLR2F	9193	281	0.1757	-0.15	0.73
cacfd1_1	calcium channel flower domain containing 1	118332514	ZDB-GENE-120215-178	CACFD1	1365	282	0.1757	-0.22	0.29
rangap1a_1	RAN GTPase activating protein 1a	118323500	ZFIN:ZDB-GENE-060929-492	RANGAP1	9854	283	0.1757	-0.08	0.90
ccsapb_2	centriole, cilia and spindle-associated protein b	118343345	ZFIN:ZDB-GENE-081107-51	CCSAP	29578	284	0.1728	-0.52	0.26
uba5_1	ubiquitin-like modifier activating enzyme 5	118322631	ZDB-GENE-031112-2	UBA5	23230	285	0.1728	0.47	0.43
elavl4_1	ELAV like neuron-specific RNA binding protein 4	118321541	ZDB-GENE-990415-246	ELAVL4	3315	286	0.1728	1.41	0.23
exoc8_1	exocyst complex component 8	118340615	ZDB-GENE-070410-60	EXOC8	24659	287	0.1728	-0.09	0.79
egln1a_2	egl-9 family hypoxia-inducible factor 1a	118329161	ZFIN:ZDB-GENE-110408-34	EGLN1	1232	288	0.1728	-0.1	0.87
LOC118335888_1	cytochrome c oxidase assembly protein COX19	118335888	ZFIN:ZDB-GENE-070410-29	COX19	28074	289	0.1728	-0.16	0.73
LOC118340245_1	DNA (cytosine-5)-methyltransferase 3A-like	118340245	-	DNMT3A	2978	290	0.1728	0.27	0.27
LOC118322091_1	serine protease 53-like, mRNA	118322091	-	PRSS53	34407	291	0.1728	0.88	0.04
LOC118339573_3	target of Nesh-SH3	118339573	ZDB-GENE-070912-325	ABI3BP	17265	292	0.1728	0.53	0.48
zgc:154075_1	uncharacterized LOC118325005	118325005	ZFIN:ZDB-GENE-060929-910	GFOD1	21096	293	0.1728	-0.1	0.94
rufy2_1	RUN and FYVE domain containing 2	118343390	ZDB-GENE-070424-66	RUFY2	19761	294	0.1728	0.29	0.25

Appendix A (continued).

rxrgb_1	retinoid X receptor, gamma b	118321988	ZFIN:ZDB-GENE- 040718-34	RXRG	10479	295	0.1728	0.41	0.05
rbis_1	ribosomal biogenesis factor	118322240	ZDB-GENE- 030219-148	RBIS	32235	296	0.1728	-0.09	0.90
dbnnd1_1	dysbindin domain containing 1	118321128	ZDB-GENE- 121214-359	DBNDD1	28455	297	0.1728	-0.39	0.73
LOC118328559_1	aminoacylase-1-like	118328559	-	ACY1	177	298	0.1728	-0.13	0.70
calm2a_1	calmodulin 2a (phosphorylase kinase, delta)	118329268	ZFIN:ZDB-GENE- 030804-3	CALM2	1445	299	0.1728	-0.29	0.21
trim44_3	tripartite motif containing 44	118329911	ZDB-GENE- 050522-37	TRIM44	19016	300	0.1728	0.38	0.10

Appendix B

Gene transcripts quantitated in striped bass (SB, *Morone saxatilis*) using CLC Genomics Workbench (v.21.0.3, Qiagen Inc., Hilden, Germany) RNA-Seq Analysis tool and reference genome sequence assembly for these fish published through National Institutes of Health National Center for Biotechnology Information (NCBI) GenBank® (NCSU_SB_2.0, accession: GCA_004916995.1). These transcripts were identified as the optimal dataset for the comparison of SB into groups based on growth performance to Market Size (1.36 kg), the reciprocal other end of the size distribution ‘Under Size’, and the intermediate groups Inferior Other (below the median) and Superior Other (above the median). The optimal dataset was determined via application of a machine learning (ML) pipeline whereby attributes (gene transcripts) are assigned an information gain value (Column: Info Gain), or the amount of information the inclusion of data associated to a given attribute provides the ML model for the classification of instances (individuals) into comparison classes (groups). Attributes were assigned a rank based upon the information gain value such that the first attribute (#1) has the highest information gain, or is the most informative to the classification of instances into groups (Column: Rank). Ranked attributes were then processed through four ML algorithms each twice cross-validated and whereby bottom-ranking attributes were eliminated in a recursive manner to determine points of improvement and degradation of model performance. The optimal dataset is that which yielded optimum performance of sorting individual SB into groups from as many cross-validated models as possible. SB gene and gene transcript information output from CLC Genomics Workbench is provided in columns: SB Gene, which is the SB gene symbol followed by an underscore and the transcript number; SB Gene Name; SB Gene ID, which is the NCBI ID number for a given gene. The IDs of orthologs of a given gene for zebrafish (*Danio rerio*) (Column: ZFIN ID) and humans (*Homo sapiens*) are also provided (Column: HGNC Symbol, HGNC ID). The log₂ fold change (Column: Log₂ FC) is the log₂ of the fold change values calculated as the maximum mean between the groups divided by the minimum mean, and multiplied by -1 if the maximum value was the second in order from left to right between observations. Statistical tests were performed as a component of the CLC Genomics Workbench Differential Expression for RNA-Seq analysis tool, which employs the Likelihood ratio test (“Across groups (ANOVA-like)”) Likelihood ratio test (“Across groups (ANOVA-like)”) for comparisons of more than two groups. The false discovery rate-corrected p-value (Column: FDR P-Value) was used in subsequent pathway analyses.

SB Gene	SB Gene Name	SB Gene ID	ZFIN ID	HGNC Symbol	HGNC ID	Rank	Info Gain	Log ₂ FC	FDR P-Value
glod4_1	glyoxalase domain containing 4	118341704	ZDB-GENE-040912-38	GLOD4	14111	1	0.581	-0.59	0.09
LOC118322896_1	parvalbumin beta-1-like	118322896	ZDB-GENE-040625-48	PVALB	9704	2	0.431	-2.03	9.37E-03
asns_1	asparagine synthetase	118342428	ZDB-GENE-040426-1091	ASNS	753	3	0.422	-0.83	4.11E-03
mak16_1	MAK16 homolog	118325976	ZDB-GENE-020419-35	MAK16	13703	4	0.398	-0.72	0.07
LOC118338989_1	cytochrome c-b	118338989	-	CYCS	19986	5	0.388	-0.97	0.11

Appendix B (continued).

si:ch73-267c23.10_1	serine incorporator 1	118328318	ZFIN:ZDB-GENE-160114-59	SERINC3	11699	6	0.362	-0.85	2.23E-04
ttc4_1	tetratricopeptide repeat domain 4	118322348	ZDB-GENE-040426-1021	TTC4	12394	7	0.355	-0.44	0.62
LOC118335538_2	CTP synthase 1-like	118335538	-	CTPS1	2519	8	0.349	-0.57	0.29
riox1_1	ribosomal oxygenase 1	118340642	ZDB-GENE-041008-221	RIOX1	20968	9	0.345	-0.49	0.36
klhl23_1	kelch-like family member 23	118339246	ZDB-GENE-081104-458	KLHL23	27506	10	0.332	-0.74	0.07
LOC118334992_1	nucleolin-like	118334992	-	NCL	7667	11	0.332	-0.97	1.43E-03
LOC118342946_1	elongation factor 1-alpha-like	118342946	-	EEF1A1	3189	12	0.329	-0.31	1
trappc6bl_1	trafficking protein particle complex 6b-like	118341276	ZFIN:ZDB-GENE-040426-1602	TRAPPC6A	23069	13	0.326	-0.41	0.45
prmt3_1	protein arginine methyltransferase 3	118321154	ZDB-GENE-041105-1	PRMT3	30163	14	0.326	-0.55	0.47
zpr1_1	ZPR1 zinc finger	118327934	ZDB-GENE-040426-2110	ZPR1	13051	15	0.326	-0.68	0.18
pes_1	pescadillo	118326120	ZDB-GENE-990415-206	PES1	8848	16	0.322	-0.56	0.31
rnfl1a_1	ring finger protein 11a	118327031	ZFIN:ZDB-GENE-040426-1277	RNF11	10056	17	0.321	-0.35	0.72
tsr1_1	TSR1 ribosome maturation factor	118341868	ZDB-GENE-030131-3762	TSR1	25542	18	0.32	-0.61	0.37
ipo4_1	importin 4	118324296	ZDB-GENE-041014-307	IPO4	19426	19	0.32	-0.93	2.11E-03
LOC118326533_1	multidrug and toxin extrusion protein 1-like, mRNA	118326533	-	SLC47A1	25588	20	0.318	-0.44	0.41
st13_1	ST13 Hsp70 interacting protein	118323218	ZDB-GENE-030131-5395	ST13	11343	21	0.317	-0.26	1
LOC118327918_1	cytokine-inducible SH2-containing protein-like, mRNA	118327918	-	CISH	1984	22	0.316	1.11	1.58E-05
chchd4a_1	coiled-coil-helix-coiled-coil-helix domain containing 4a	118328695	ZFIN:ZDB-GENE-040801-46	CHCHD4	26467	23	0.315	-0.43	0.54
hars_1	histidyl-tRNA synthetase	118333619	ZDB-GENE-040912-152	HARS1	4816	24	0.315	-0.37	0.91

Appendix B (continued).

bxdc2_1	brix domain containing 2	118332963	ZDB-GENE-060518-1	BRIX1	24170	25	0.315	-0.76	0.19
ghrb_1	growth hormone receptor b	118332354	ZFIN:ZDB-GENE-071119-4	GHR	4263	26	0.315	-0.81	0.05
tarbp1_1	TAR (HIV-1) RNA binding protein 1	118329670	ZDB-GENE-090313-346	TARBP1	11568	27	0.314	-0.87	0.06
LOC118322991_1	very-long-chain enoyl-CoA reductase-like	118322991	-	TECR	4551	28	0.314	-0.58	0.13
selenoo2_1	selenoprotein O2	118340647	ZFIN:ZDB-GENE-041210-110	SELENOO	30395	29	0.314	-0.33	0.79
elac2_1	elaC ribonuclease Z 2	118323232	ZDB-GENE-041111-227	ELAC2	14198	30	0.312	-0.4	0.64
ppm1aa_2	protein phosphatase, Mg ²⁺ /Mn ²⁺ dependent, 1Aa	118337055	ZFIN:ZDB-GENE-991102-15	PPM1A	9275	31	0.312	0.83	0.07
gnl3_1	guanine nucleotide binding protein-like 3 (nucleolar)	118328734	ZDB-GENE-030131-616	GNL3	29931	32	0.309	-0.39	0.93
LOC118342385_2	voltage-dependent calcium channel gamma-6 subunit-like	118342385	-	CACNG6	13625	33	0.308	-0.77	0.04
LOC118340847_1	26S proteasome regulatory subunit 4	118340847	-	PSMC1	9547	34	0.308	-0.54	0.29
LOC118327556_1	transcription factor IIIA-like, mRNA	118327556	-	GTF3A	4662	35	0.307	-0.68	0.59
dcun1d2a_2	DCN1, defective in cullin neddylation 1, domain containing 2a	118338896	ZFIN:ZDB-GENE-020416-2	DCUN1D2	20328	36	0.302	-1.04	5.35E-03
grwd1_1	glutamate-rich WD repeat containing 1	118335870	ZDB-GENE-030131-9844	GRWD1	21270	37	0.301	-0.54	0.45
chac2_1	ChaC, cation transport regulator homolog 2	118329427	ZDB-GENE-050706-146	CHAC2	32363	38	0.301	-0.53	0.44
fut9a_1	fucosyltransferase 9a	118337648	ZFIN:ZDB-GENE-080723-75	FUT9	4020	39	0.3	-0.75	0.3
naxe_1	NAD(P)HX epimerase	118343071	ZDB-GENE-040718-362	NAXE	18453	40	0.3	0.58	0.19
LOC118325343_1	gamma-crystallin M2-like	118325343	-	CRYGC	2410	41	0.299	-0.48	0.64
ppid_1	peptidylprolyl isomerase D	118333298	ZDB-GENE-040625-34	PPID	9257	42	0.299	-0.58	0.25

Appendix B (continued).

LOC118332666_1	eukaryotic translation initiation factor 4E-binding protein 1-like	118332666	-	EIF4EBP1	3288	43	0.299	-0.82	0.45
herc4_1	HECT and RLD domain containing E3 ubiquitin protein ligase 4	118329204	ZDB-GENE-070801-1	HERC4	24521	44	0.298	-0.51	0.14
usp14_1	ubiquitin specific peptidase 14 (tRNA-guanine transglycosylase)	118334598	ZDB-GENE-030131-7676	USP14	12612	45	0.298	-0.63	0.14
LOC118329509_1	centromere protein F-like	118329509	-	CENPF	1857	46	0.298	-0.44	0.58
kat14_1	lysine acetyltransferase 14	118326661	ZDB-GENE-040718-452	KAT14	15904	47	0.295	0.31	0.93
dusp11_1	dual specificity phosphatase 11 (RNA/RNP complex 1-interacting)	118333397	ZDB-GENE-050417-226	DUSP11	3066	48	0.295	0.47	1
LOC118329379_1	serine/threonine-protein kinase 32C-like	118329379	-	STK32C	21332	49	0.294	-0.89	9.05E-03
nup133_1	nucleoporin 133	118327465	ZDB-GENE-040426-2941	NUP133	18016	50	0.294	-0.44	0.69
bud23_1	BUD23 rRNA methyltransferase and ribosome maturation factor	118341232	ZDB-GENE-070410-68	BUD23	16405	51	0.292	-0.39	1
rsl1d1_1	ribosomal L1 domain containing 1	118335066	ZDB-GENE-130603-67	RSL1D1	24534	52	0.291	0.55	0.63
tff1.6_1	transcription termination factor 1, tandem duplicate 6	118325629	ZFIN:ZDB-GENE-041008-214	TTF1	12397	53	0.291	-0.54	0.41
polr1g_1	RNA polymerase I subunit G	118341334	ZDB-GENE-101202-2	POLR1G	24219	54	0.291	-0.69	0.08
LOC118341538_1	neprilysin-like family with sequence similarity 98 member A	118341538	-	MME	7154	55	0.291	-1.11	6.32E-04
fam98a_1	DEAD (Asp-Glu-Ala-Asp) box polypeptide 54	118329083	ZDB-GENE-091202-6	FAM98A	24520	56	0.291	-0.49	0.08
ddx54_1	NOP58 ribonucleoprotein homolog	118332610	ZDB-GENE-021220-2	DDX54	20084	57	0.29	-0.41	0.74
nop58_1	serine incorporator 1-like	118340073	ZDB-GENE-040426-2140	NOP58	29926	58	0.289	-0.59	0.16
LOC118340698_1		118340698	-	SERINC1	13464	59	0.286	-0.83	5.35E-03

Appendix B (continued).

psmc4_1	proteasome 26S subunit, ATPase 4	118335944	ZDB-GENE-030131-5083	PSMC4	9551	60	0.285	-0.66	0.17
LOC118330728_1	protein kinase C and casein kinase substrate in neurons 2 protein-like	118330728	-	PACSIN2	8571	61	0.285	0.8	0.02
abitram_1	actin binding transcription modulator	118342785	ZDB-GENE-060503-602	ABITRAM	1364	62	0.284	0.48	0.97
nol10_1	nucleolar protein 10	118337522	ZDB-GENE-040426-764	NOL10	25862	63	0.283	-0.45	0.52
msh3_1	mutS homolog 3	118325540	ZDB-GENE-060526-307	MSH3	7326	64	0.283	-0.43	0.77
psmd6_1	proteasome 26S subunit, non-ATPase 6	118328818	ZDB-GENE-040426-1038	PSMD6	9564	65	0.283	-0.55	0.32
aimp2_1	aminoacyl tRNA synthetase complex interacting multifunctional protein 2	118322933	ZDB-GENE-040426-2652	AIMP2	20609	66	0.283	0.31	1
eif3ba_1	eukaryotic translation initiation factor 3, subunit Ba	118326373	ZFIN:ZDB-GENE-030131-2748	EIF3B	3280	67	0.283	-0.3	1
cct8_1	chaperonin containing TCP1, subunit 8 (theta)	118331637	ZDB-GENE-040426-876	CCT8	1623	68	0.279	-0.26	1
fgf2_1	fibroblast growth factor 2	118333801	ZDB-GENE-040318-1	FGF2	3676	69	0.279	-0.18	1
fbl_1	fibrillarin	118335580	ZDB-GENE-040426-1936	FBL	3599	70	0.278	-0.55	0.14
enpp6_1	ectonucleotide pyrophosphatase/phosphodiesterase 6	118333847	ZDB-GENE-031205-1	ENPP6	23409	71	0.278	-0.53	0.3
dhodh_1	dihydroorotate dehydrogenase	118343269	ZDB-GENE-030131-3157	DHODH	2867	72	0.277	-0.49	0.78
pum3_1	pumilio RNA-binding family member 3	118332424	ZDB-GENE-030131-9808	PUM3	29676	73	0.277	-0.66	0.54
LOC118333274_1	uncharacterized	118333274	-	NXF1	8071	74	0.276	-0.96	0.08
pwpl1_1	PWP1 homolog, endonuclease	118342770	ZDB-GENE-040426-1049	PWP1	17015	75	0.276	-0.39	0.45
rpp21_1	ribonuclease P 21 subunit	118321359	ZDB-GENE-040801-37	RPP21	21300	76	0.275	-0.92	0.12

Appendix B (continued).

gpib_1	glucose-6-phosphate isomerase b	118343600	ZFIN:ZDB-GENE-020513-3	GPI	4458	77	0.275	0.99	0.03
ppan_1	peter pan homolog	118323434	ZDB-GENE-030114-4	PPAN	9227	78	0.275	-0.46	0.69
hsp90aa1.2_2	heat shock protein 90, alpha (cytosolic), class A member 1, tandem duplicate 2	118340391	ZFIN:ZDB-GENE-031001-3	HSP90AA1	5253	79	0.274	-0.63	0.05
rrp8_1	ribosomal RNA processing 8	118341308	ZDB-GENE-030131-8898	RRP8	29030	80	0.274	-0.59	0.42
LOC118329129_1	dnaJ homolog subfamily C member 7-like	118329129	-	DNAJC7	12392	81	0.274	-0.42	0.89
ube2g1b_1	ubiquitin-conjugating enzyme E2G 1b (UBC7 homolog)	118332502	ZFIN:ZDB-GENE-040426-2939	UBE2G1	12482	82	0.274	-0.54	0.33
ube3a_1	ubiquitin protein ligase E3A	118321655	ZDB-GENE-041114-190	UBE3A	12496	83	0.274	-0.6	0.14
LOC118338975_1	zinc-binding protein A33-like	118338975	-	TRIM27	9975	84	0.274	0.45	0.25
dkc1_1	dyskeratosis congenita 1, dyskerin	118337478	ZDB-GENE-031118-120	DKC1	2890	85	0.273	-0.36	0.86
lemd1_1	LEM domain containing 1	118324763	ZDB-GENE-040718-350	LEMD1	18725	86	0.271	0.48	0.68
ttl12_1	tubulin tyrosine ligase-like family, member 12	118341811	ZDB-GENE-040426-697	TTLL12	28974	87	0.271	-0.49	0.27
LOC118337496_1	activator of 90 kDa heat shock protein ATPase homolog 1-like	118337496	-	AHSA1	1189	88	0.271	-0.49	0.16
rrp12_1	ribosomal RNA processing 12 homolog	118329138	ZDB-GENE-050706-182	RRP12	29100	89	0.27	-0.61	0.19
LOC118332608_1	serine/threonine-protein phosphatase 6 catalytic subunit-like	118332608	-	PPP6C	9323	90	0.269	0.11	1
zcchc4_1	zinc finger, CCHC domain containing 4	118338349	ZDB-GENE-080204-65	ZCCHC4	22917	91	0.269	-0.37	0.96
ints11_1	integrator complex subunit 11	118324839	ZDB-GENE-050522-13	INTS11	26052	92	0.269	-0.42	0.2
aspg_2	asparaginase homolog	118336951	ZDB-GENE-070820-14	ASPG	20123	93	0.267	-0.62	0.15

Appendix B (continued).

LOC118330619_1	apoptosis-enhancing nuclease-like	118330619	-	AEN	25722	94	0.267	-0.59	0.58
LOC118333681_1	inactive dual specificity phosphatase 27-like	118333681	-	DUSP3	3069	95	0.267	-0.49	0.75
LOC118342612_1	tubulin--tyrosine ligase-like protein 12	118342612	-	TTLL12	28974	96	0.267	-0.61	0.31
rpf2_1	ribosome production factor 2 homolog	118341033	ZDB-GENE-040426-2501	RPF2	20870	97	0.266	-0.5	0.74
psmb5_1	proteasome 20S subunit beta 5	118334629	ZDB-GENE-990415-215	PSMB5	9542	98	0.265	-0.36	0.61
LOC118338088_1	type I phosphatidylinositol 4,5-bisphosphate 4-phosphatase-A-like	118338088	-	PIP4P1	19299	99	0.265	-0.42	0.26
LOC118333491_1	heterogeneous nuclear ribonucleoprotein A0-like	118333491	-	HNRNPA0	5030	100	0.264	0.27	0.94
skp1_1	S-phase kinase-associated protein 1	118331882	ZDB-GENE-040426-1707	SKP1	10899	101	0.262	-0.37	0.92
LOC118337907_1	Y-box-binding protein 2-A-like	118337907	-	YBX2	17948	102	0.262	-0.46	0.66
LOC118339008_1	transmembrane protein 50B	118339008	-	TMEM50B	1280	103	0.261	0.36	0.87
LOC118341065_1	proteasome subunit alpha type-6	118341065	ZFIN:ZDB-GENE-020326-1	PSMA6	9535	104	0.261	-0.43	0.89
hdhd5_1	haloacid dehalogenase like hydrolase domain containing 5	118330274	ZDB-GENE-080220-59	HDHD5	1843	105	0.26	-0.39	0.39
LOC118323545_1	deoxyhypusine synthase-like, transcript variant X3, mRNA	118323545	-	DHPS	2869	106	0.259	-0.46	0.68
ccdc191_1	coiled-coil domain containing 191	118321052	ZDB-GENE-050102-4	CCDC191	29272	107	0.258	-1.56	0.02
rxrgb_1	retinoid X receptor, gamma b	118321988	ZFIN:ZDB-GENE-040718-34	RXRG	10479	108	0.257	0.53	0.16
psme3_1	proteasome activator subunit 3	118327615	ZDB-GENE-991110-19	PSME3	9570	109	0.257	-0.4	0.79
LOC118321211_1	partitioning defective 3 homolog, mRNA	118321211	ZDB-GENE-170531-1	PARD3	16051	110	0.256	1.07	0.02

Appendix B (continued).

g3bp1_1	GTPase activating protein (SH3 domain) binding protein 1	118333368	ZDB-GENE-030131-7452	G3BP1	30292	111	0.256	-0.63	0.05
leap2_1	liver-expressed antimicrobial peptide 2	118338304	ZDB-GENE-081022-131	LEAP2	29571	112	0.256	2.66	4.37E-05
LOC118325831_1	granzyme A-like	118325831	-	GMZA	4708	113	0.256	1.77	0.19
fndc10_1	fibronectin type III domain containing 10	118327904	ZDB-GENE-141216-142	FNDC10	42951	114	0.256	-1.04	0.2
eef1e1_1	eukaryotic translation elongation factor 1 epsilon 1	118341986	ZDB-GENE-030131-4949	EEF1E1	3212	115	0.256	-0.32	1
rangap1a_1	RAN GTPase activating protein 1a	118323500	ZFIN:ZDB-GENE-060929-492	RANGAP1	9854	116	0.255	-0.28	1
btd_1	biotinidase	118342413	ZDB-GENE-060825-186	BTD	1122	117	0.255	-0.36	1
qdpra_1	quinoid dihydropteridine reductase a	118333940	ZFIN:ZDB-GENE-070705-197	QDPR	9752	118	0.255	-0.25	1
hnmt_1	histamine N-methyltransferase	118338631	ZDB-GENE-040801-157	HNMT	5028	119	0.255	-0.83	0.29
LOC118335005_1	transcription factor HES-1-B-like	118335005	-	HES1	5192	120	0.255	-1.5	8.82E-03
rpp25l_1	ribonuclease P/MRP 25 subunit-like	118336486	ZDB-GENE-040718-275	RPP25L	19909	121	0.254	0.33	0.65
angel1_1	angel homolog 1	118337362	ZDB-GENE-111020-12	ANGEL1	19961	122	0.254	-0.3	0.42
gpt2l_3	glutamic pyruvate transaminase (alanine aminotransferase) 2, like	118334633	ZFIN:ZDB-GENE-050302-11	GPT	4552	123	0.254	-0.34	0.84
dnajb12b_1	DnaJ heat shock protein family (Hsp40) member B12b	118334834	ZFIN:ZDB-GENE-070410-128	DNAJB12	14891	124	0.254	-0.41	0.94
aadat_2	aminoadipate aminotransferase	118338366	ZDB-GENE-111229-1	AADAT	17929	125	0.254	-0.31	0.38
rfxap_1	regulatory factor X-associated protein	118331395	ZDB-GENE-091204-321	RFXAP	9988	126	0.253	0.32	1
zranb1a_3	zinc finger, RAN-binding domain containing 1a	118325531	ZFIN:ZDB-GENE-100212-3	ZRANB1	18224	127	0.253	-0.54	0.55
aclyb_1	ATP citrate lyase b	118329092	ZFIN:ZDB-GENE-090909-2	ACLY	115	128	0.253	-0.7	0.11

Appendix B (continued).

ddx52_1	DEAD (Asp-Glu-Ala-Asp) box polypeptide 52	118331797	ZDB-GENE-060623-1	DDX52	20038	129	0.252	-0.3	1
ptcd3_1	pentatricopeptide repeat domain 3	118334089	ZDB-GENE-030131-6849	PTCD3	24717	130	0.252	0.1	1
epdr1_1	ependymin related 1	118334359	ZDB-GENE-040718-113	EPDR1	17572	131	0.252	-0.28	1
npepl1_1	aminopeptidase like 1	118328725	ZDB-GENE-050417-177	NPEPL1	16244	132	0.252	0.26	1
LOC118330685_1	voltage-dependent anion-selective channel protein 2-like	118330685	-	VDAC2	12672	133	0.251	-0.21	1
rrs1_1	ribosome biogenesis regulator 1 homolog	118324391	ZDB-GENE-030131-9345	RRS1	17083	134	0.251	-0.25	0.97
LOC118336003_2	protein LYRIC-like	118336003	-	MTDH	29608	135	0.25	0.25	1
znf622_1	zinc finger protein 622	118334423	ZDB-GENE-050927-1	ZNF622	30958	136	0.25	-0.7	0.02
LOC118330114_1	pyruvate kinase PKM-like	118330114	-	PKM	9021	137	0.25	0.45	0.61
brsk2a_1	BR serine/threonine kinase 2a	118329976	ZFIN:ZDB-GENE-100422-2	BRSK2	11405	138	0.25	-2.06	0.02
banf1_1	BAF nuclear assembly factor 1	118331442	ZDB-GENE-030131-6657	BANF1	17397	139	0.25	-0.48	0.42
wdr3_1	WD repeat domain 3	118339512	ZDB-GENE-030131-9830	WDR3	12755	140	0.25	-0.43	0.72
cth_1	cystathionase (cystathionine gamma-lyase)	118321875	ZDB-GENE-030131-774	CTH	2501	141	0.25	-0.47	0.56
gprc5bb_1	G protein-coupled receptor, class C, group 5, member Bb	118326480	ZFIN:ZDB-GENE-090313-75	GPRC5B	13308	142	0.25	-2.27	0.11
dnai1.2_1	dynein, axonemal, intermediate chain 1, paralog 2	118328544	ZFIN:ZDB-GENE-070112-1302	DNAI1	2954	143	0.249	-0.75	0.21
rrp9_2	ribosomal RNA processing 9, U3 small nucleolar RNA binding protein	118328331	ZDB-GENE-060427-1	RRP9	16829	144	0.248	-0.49	0.79
serpinc1_1	serpin peptidase inhibitor, clade C (antithrombin), member 1	118321957	ZDB-GENE-030131-264	SERPINC1	775	145	0.248	-0.59	0.2

Appendix B (continued).

mylk4a_1	myosin light chain kinase family, member 4a	118334412	ZFIN:ZDB-GENE-120824-2	MYLK4	27972	146	0.248	-6.75	5.57E-03
cdh19_1	cadherin 19, type 2	118343284	ZDB-GENE-130530-766	CDH19	1758	147	0.248	-0.43	0.51
utp15_2	UTP15 small subunit processome component	118325671	ZDB-GENE-030131-3831	UTP15	25758	148	0.247	-0.53	0.23
zgc:86609_1	TMCO1/EMC3 family protein	118324059	ZFIN:ZDB-GENE-050102-7	EMC3	23999	149	0.247	0.26	1
prep_1	prolyl endopeptidase	118340874	ZDB-GENE-050522-14	PREP	9358	150	0.247	-0.68	0.27
ntmt1_1	N-terminal Xaa-Pro-Lys N-methyltransferase 1	118332685	ZDB-GENE-040426-2055	NTMT1	23373	151	0.247	-0.46	0.28
ruvbl2_1	RuvB-like AAA ATPase 2	118331119	ZDB-GENE-030109-1	RUVBL2	10475	152	0.247	0.57	0.44
dph2_1	diphthamide biosynthesis 2	118328022	ZDB-GENE-030219-100	DPH2	3004	153	0.247	-0.64	0.18
LOC118332097_2	F-box only protein 40-like	118332097	-	FBXO40	29816	154	0.247	-0.25	1
heatr1_1	HEAT repeat containing 1	118325688	ZDB-GENE-030131-6378	HEATR1	25517	155	0.246	-0.75	0.1
si:ch211-167b20.8_3	protein phosphatase 1 regulatory subunit 3A	118324234	ZFIN:ZDB-GENE-081104-147	PPP1R3A	9291	156	0.246	0.94	0.08
aprt_1	adenine phosphoribosyltransferase	118329423	ZDB-GENE-040426-1492	APRT	626	157	0.245	-0.55	0.64
psmc6_1	proteasome 26S subunit, ATPase 6	118340648	ZDB-GENE-030131-304	PSMC6	9553	158	0.245	-0.47	0.69
rnf180_1	ring finger protein 180	118332576	ZDB-GENE-081104-458	RNF180	27752	159	0.244	-0.43	0.55
zranb1a_1	zinc finger, RAN-binding domain containing 1a	118325531	ZFIN:ZDB-GENE-100212-3	ZRANB1	18224	160	0.242	-0.79	0.35
cacnb4a_1	calcium channel, voltage-dependent, beta 4a subunit	118339601	ZFIN:ZDB-GENE-080320-2	CACNB4	1404	161	0.242	-2.94	3.89E-04
ftsj3_1	FtsJ RNA 2'-O-methyltransferase 3	118321719	ZDB-GENE-030131-9828	FTSJ3	17136	162	0.242	-0.63	0.28
llph_1	LLP homolog, long-term synaptic facilitation factor	118339675	ZDB-GENE-051030-30	LLPH	28229	163	0.241	-0.34	1
pfkfb4b_1	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4b	118328162	ZFIN:ZDB-GENE-031031-4	PFKFB4	8875	164	0.241	-1.78	1.76E-04

Appendix B (continued).

umps_1	uridine monophosphate synthetase	118339177	ZDB-GENE-040426-785	UMPS	12563	165	0.24	-0.78	7.66E-03
LOC118328778_1	probable ATP-dependent RNA helicase DDX27	118328778	-	DDX27	15837	166	0.24	-0.35	0.97
LOC118340372_1	carbonic anhydrase 1	118340372	-	CA1	1368	167	0.24	0.71	0.12
psip1a_1	PC4 and SFRS1 interacting protein 1a	118327438	ZFIN:ZDB-GENE-050522-104	PSIP1	9527	168	0.239	1.06	1
mrpl28_3	mitochondrial ribosomal protein L28	118342233	ZDB-GENE-050522-113	MRPL28	14484	169	0.239	1.74	0.05
tpilb_1	triosephosphate isomerase 1b	118336118	ZFIN:ZDB-GENE-020416-4	TPI1	12009	170	0.238	0.89	0.15
dnah2_1	dynein, axonemal, heavy chain 2	118338425	ZDB-GENE-130530-910	DNAH2	2948	171	0.237	-1.32	0.17
LOC118326239_1	D-amino-acid oxidase-like	118326239	-	DAO	2671	172	0.237	1.03	0.05
trib3_1	tribbles pseudokinase 3	118328352	ZDB-GENE-040426-2609	TRIB3	16228	173	0.237	-0.52	0.51
cunh5orf22_1	chromosome unknown C5orf22 homolog	118321909	ZFIN:ZDB-GENE-040426-1078	C5orf22	25639	174	0.237	-0.36	1
pcyt2_1	phosphate cytidylyltransferase 2, ethanolamine	118327034	ZDB-GENE-041010-132	PCYT2	8756	175	0.237	-0.63	0.11
psmc2_1	proteasome 26S subunit, ATPase 2	118337318	ZDB-GENE-040426-1327	PSMC2	9548	176	0.236	-0.48	0.42
psmd12_1	proteasome 26S subunit, non-ATPase 12	118327453	ZDB-GENE-030131-617	PSMD12	9557	177	0.236	-0.56	0.23
serbp1a_2	SERPINE1 mRNA binding protein 1a	118321603	ZFIN:ZDB-GENE-030131-7389	SERBP1	17860	178	0.236	-0.15	1
rpa3_1	replication protein A3	118343258	ZDB-GENE-040426-977	RPA3	10291	179	0.236	0.43	0.96
ky_1	kyphoscoliosis peptidase	118337021	ZDB-GENE-090313-64	KY	26576	180	0.236	1.53	2.08E-04
gkap1_3	G kinase anchoring protein 1	118326161	ZDB-GENE-040426-2485	GKAP1	17496	181	0.235	0.5	0.72
LOC118331977_1	fibroblast growth factor 13-like	118331977	-	FGF13	3670	182	0.235	1.27	0.03
hppt1_1	hypoxanthine phosphoribosyltransferase 1	118333750	ZDB-GENE-040426-1918	HPRT1	5157	183	0.235	-0.36	1

Appendix B (continued).

usp12a_1	ubiquitin specific peptidase 12a	118343633	ZFIN:ZDB-GENE-060228-3	USP12	20485	184	0.235	-0.32	0.85
LOC118342439_1	caldesmon-like	118342439	-	CALD1	1441	185	0.235	1.43	0.14
LOC118333798_1	uncharacterized	118333798	-	SHTN1	29319	186	0.235	2.7	0.13
nucb1_1	nucleobindin 1	118323501	ZDB-GENE-060825-222	NUCB1	8043	187	0.235	-0.25	1
strap_1	serine/threonine kinase receptor associated protein	118337952	ZDB-GENE-040426-1110	STRAP	30796	188	0.234	-0.22	1
si:dkey-81h8.1_1	uncharacterized	118342455	ZFIN:ZDB-GENE-060503-931	MS4A15	28573	189	0.233	1.88	0.2
LOC118342455	LOC118342455								
slc38a7_1	solute carrier family 38 member 7	118320766	ZDB-GENE-040801-266	SLC38A7	25582	190	0.232	0.17	1
kbtbd12_1	kelch repeat and BTB (POZ) domain containing 12	118328853	ZDB-GENE-050904-1	KBTBD12	25731	191	0.231	-0.31	0.87
arhgef2_2	rho/rac guanine nucleotide exchange factor (GEF) 2	118335611	ZDB-GENE-030717-1	ARHGEF2	682	192	0.23	2.85	0.18
naa15a_1	N-alpha-acetyltransferase 15, NatA auxiliary subunit a	118333569	ZFIN:ZDB-GENE-030131-2392	NAA15	30782	193	0.23	-0.23	1
LOC118327282_1	transcription elongation factor 1 homolog, mRNA	118327282	-	ELOF1	28691	194	0.23	-0.45	0.83
srm_1	spermidine synthase	118324894	ZDB-GENE-040426-1183	SRM	11296	195	0.23	-0.53	0.4
scyl3_1	SCY1-like, kinase-like 3	118322100	ZDB-GENE-030131-2559	SCYL3	19285	196	0.23	-0.62	0.26
polr3d_2	polymerase (RNA) III (DNA directed) polypeptide D	118325689	ZDB-GENE-040426-1017	POLR3D	1080	197	0.23	-0.41	0.97
pdc11_3	programmed cell death 11	118326580	ZDB-GENE-030131-4076	PDCD11	13408	198	0.23	-0.56	0.62
LOC118324005_1	glycogen debranching enzyme, partial mRNA	118324005	-	AGL	321	199	0.229	0.87	0.02
dus4l_1	dihydrouridine synthase 4-like	118341357	ZDB-GENE-040801-233	DUS4L	21517	200	0.229	-0.49	1
nudc_2	nudC nuclear distribution protein	118343026	ZDB-GENE-040426-899	NUDC	8045	201	0.229	-0.46	0.67
slc16a5a_1	solute carrier family 16 member 5a	118328044	ZFIN:ZDB-GENE-121108-4	SLC16A5	10926	202	0.229	-0.67	0.25

Appendix B (continued).

gpn1_1	GPN-loop GTPase 1	118320757	ZDB-GENE-040801-154	GPN1	17030	203	0.229	0.45	0.87
kdm4c_5	lysine (K)-specific demethylase 4C	118326862	ZDB-GENE-070209-38	KDM4C	17071	204	0.229	-0.59	1
bdh1_1	3-hydroxybutyrate dehydrogenase, type 1	118343495	ZDB-GENE-070410-130	BDH1	1027	205	0.228	-0.49	0.55
LOC118341306_1	glycine--tRNA ligase-like	118341306	-	GARS1	4162	206	0.228	0.37	1
LOC118327576_1	uncharacterized	118327576	-	C17orf58	27568	207	0.228	0.61	0.54
olfml2ba_1	olfactomedin-like 2Ba	118321986	ZFIN:ZDB-GENE-041014-329	OLFML2B	24558	208	0.228	1.18	5.92E-04
LOC118326917_1	histone chaperone asf1b-B-like, mRNA	118326917	-	ASF1B	20996	209	0.227	0.33	0.98
LOC118327173_1	forkhead box protein D1-like, mRNA	118327173	-	FOXD1	3802	210	0.226	0.52	1
rnf207a_2	ring finger protein 207a	118327933	ZFIN:ZDB-GENE-120813-7	RNF207	32947	211	0.226	1.3	2.23E-04
LOC118324953_1	proteasomal ubiquitin receptor ADRM1-like	118324953	-	ADRM1	15759	212	0.226	-0.38	0.79
antkmt_1	adenine nucleotide translocase lysine methyltransferase	118323424	ZDB-GENE-040625-59	ANTKMT	14152	213	0.226	-0.53	0.61
LOC118334649_2	zinc finger protein 521	118334649	ZDB-GENE-141210-10	ZNF521	24605	214	0.226	1.21	1
rlnl_1	Ras and Rab interactor-like	118342228	ZDB-GENE-110216-3	RINL	24795	215	0.225	0.32	1
rnf165a_4	ring finger protein 165a	118332515	ZFIN:ZDB-GENE-091118-64	RNF165	31696	216	0.225	0.63	1
rraga_1	Ras-related GTP binding A	118331987	ZDB-GENE-030131-4083	RRAGA	16963	217	0.224	-0.23	1
slc38a10_1	solute carrier family 38 member 10	118327950	ZDB-GENE-050309-21	SLC38A10	28237	218	0.224	1.1	0.41
zgc:153733_1	uncharacterized	118323678	ZFIN:ZDB-GENE-060825-273	PHETA1	26509	219	0.224	-1.19	0.02
txnl1_1	thioredoxin-like 1	118332523	ZDB-GENE-040426-701	TXNL1	12436	220	0.224	-0.45	0.73
LOC118329621_1	cytochrome c-like	118329621	-	CYCS	19986	221	0.224	-0.33	0.97
mical1_1	microtubule associated monoxygenase, calponin	118324947	ZDB-GENE-081022-3	MICAL1	20619	222	0.221	5.84	1.51E-03

Appendix B (continued).

bnip3lb_1	BCL2 interacting protein 3 like b	118332314	ZFIN:ZDB-GENE-040325-1	BNIP3L	1085	223	0.221	-0.32	0.92
LOC118323206_1	dehydrogenase/reductase SDR family member 7C-B-like, mRNA	118323206	-	DHRS7C	32423	224	0.22	-0.08	1
blmh_1	bleomycin hydrolase	118342159	ZDB-GENE-030131-8485	BLMH	1059	225	0.219	-0.32	1
esrrga_1	estrogen-related receptor gamma a	118320934	ZFIN:ZDB-GENE-030821-2	ESRRB	3473	226	0.215	-0.46	1
LOC118331687_1	A disintegrin and metalloproteinase with thrombospondin motifs 2-like	118331687	-	ADAMTS2	218	227	0.214	0.76	0.96
znhit3_1	zinc finger, HIT-type containing 3	118331354	ZDB-GENE-040426-1114	ZNHIT3	12309	228	0.213	-0.7	0.79
mrpl28_1	mitochondrial ribosomal protein L28	118342233	ZDB-GENE-050522-113	MRPL28	14484	229	0.213	-0.59	0.82
psma6a_1	proteasome 20S subunit alpha 6a	118337454	ZFIN:ZDB-GENE-020326-1	PSMA6	9535	230	0.213	-0.46	0.57
sdhaf2_1	succinate dehydrogenase complex assembly factor 2	118320774	ZDB-GENE-030131-7564	SDHAF2	26034	231	0.213	-0.17	1
parla_1	presenilin associated, rhomboid-like a	118334720	ZFIN:ZDB-GENE-050327-8	PARL	18253	232	0.212	0.14	1
LOC118330677_1	methionine-R-sulfoxide reductase B3-like	118330677	-	MSRB3	27375	233	0.212	-0.38	0.61
gap43_1	growth associated protein 43	118341315	ZDB-GENE-990415-87	GAP43	4140	234	0.21	1.17	0.67
mmp19_1	matrix metalloproteinase 19	118328290	ZDB-GENE-100308-3	MMP19	7165	235	0.21	1.75	4.02E-04
cldn19_1	claudin 19	118328135	ZDB-GENE-050417-242	CLDN19	2040	236	0.21	-0.52	0.14
LOC118343104_1	gastrula zinc finger protein xFG20-1-like	118343104	-	GFI1	4237	237	0.207	0.41	0.97
LOC118343527_1	interleukin-18 receptor accessory protein-like	118343527	-	IL18RAP	5989	238	0.206	4.27	1.54E-03
LOC118338468_1	negative elongation factor A-like	118338468	-	NSMF	29843	239	0.204	-0.52	0.52
LOC118339573_3	target of Nesh-SH3	118339573	ZDB-GENE-070912-325	ABI3BP	17265	240	0.204	0.79	0.93

Appendix B (continued).

LOC118322091_1	serine protease 53-like, mRNA	118322091	-	PRSS53	34407	241	0.204	1.54	0.04
usp2a_3	ubiquitin specific peptidase 2a	118341169	ZFIN:ZDB-GENE- 041212-59	USP2	12618	242	0.201	-0.51	0.53
si:dkey- 22114.11_2	claudin-23	118337626	ZFIN:ZDB-GENE- 131127-551	CLDN23	17591	243	0.201	-1.02	0.89
LOC118331522_1	UPF0489 protein C5orf22 homolog	118331522	-	C5orf22	25639	244	0.201	-0.42	1
LOC118333599_1	magnesium transporter protein 1	118333599	-	MAGT1	28880	245	0.201	-0.35	0.4
rcn3_1	reticulocalbin 3, EF-hand calcium binding domain	118322831	ZDB-GENE- 040625-175	RCN3	21145	246	0.201	0.86	0.18
LOC118321935_1	coiled-coil domain- containing protein 106-like	118321935	ZDB-GENE- 041008-66	CCDC106	30181	247	0.199	-1.86	0.15
fyco1a_1	FYVE and coiled-coil domain autophagy adaptor 1a	118321512	ZFIN:ZDB-GENE- 110411-276	FYCO1	14673	248	0.199	-0.27	1
tango2_2	transport and golgi organization 2 homolog	118325612	ZDB-GENE- 040808-42	TANGO2	25439	249	0.199	-0.27	1
rai14_1	retinoic acid induced 14	118333089	ZDB-GENE- 040718-235	RAI14	14873	250	0.199	1.01	0.01
necap2_2	NECAP endocytosis associated 2	118327916	ZDB-GENE- 050522-67	NECAP2	25528	251	0.198	-0.72	0.72
LOC118326875_1	uncharacterized	118326875	-	PVALB	9704	252	0.198	-0.91	0.63
psma2_1	proteasome 20S subunit alpha 2	118336161	ZDB-GENE- 050522-479	PSMA2	9531	253	0.198	-0.42	0.89
ubr5_6	ubiquitin protein ligase E3 component n-recogin 5	118336001	ZDB-GENE- 030131-6559	UBR5	16806	254	0.197	2.57	0.19
dnajb1b_1	DnaJ heat shock protein family (Hsp40) member B1b	118327597	ZFIN:ZDB-GENE- 030131-5455	DNAJB1	5270	255	0.197	0.33	0.92
inpp5e_1	inositol polyphosphate-5- phosphatase E	118325341	ZDB-GENE- 050809-23	INPP5E	21474	256	0.197	0.19	1
LOC118327907_2	WD repeat-containing protein 82-like	118327907	-	WDR82	28826	257	0.197	0.41	0.86
stoml2_1	stomatin (EPB72)-like 2	118332940	ZDB-GENE- 040426-1139	STOML2	14559	258	0.197	-0.18	1
LOC118330390_1	protein piccolo-like	118330390	-	PCLO	13406	259	0.196	-1.66	0.24

Appendix B (continued).

fbxl3l_1	F-box and leucine-rich repeat protein 3, like	118333337	ZFIN:ZDB-GENE-130530-598	FBXL21P	13600	260	0.196	0.8	0.95
myo6b_1	myosin VIb	118337331	ZFIN:ZDB-GENE-030318-3	MYO6	7605	261	0.196	-0.73	0.15
LOC118337669_1	tetraspanin-8-like	118337669	-	TSPAN8	11855	262	0.194	-0.09	1
gpr156_1	G protein-coupled receptor 156	118338974	ZDB-GENE-060201-2	GPR156	20844	263	0.194	0.22	1
atpaf2_1	ATP synthase mitochondrial F1 complex assembly factor 2	118326683	ZDB-GENE-050411-18	ATPAF2	18802	264	0.194	0.06	1
rabepk_3	Rab9 effector protein with kelch motifs	118332430	ZDB-GENE-040704-52	RABEPK	16896	265	0.191	1.85	0.81
elav14_1	ELAV like neuron-specific RNA binding protein 4	118321541	ZDB-GENE-990415-246	ELAVL4	3315	266	0.19	-1.97	0.34
LOC118342280_1	U1 spliceosomal RNA	118342280	-	RNU1-1	10120	267	0.19	2.82	0.01
LOC118330559_4	cingulin-like protein 1	118330559	-	CGNL1	25931	268	0.188	-2.03	0.75
LOC118331186_1	interleukin-1 receptor type 2-like	118331186	-	IL1R2	5994	269	0.188	3.19	0.1
si:ch211-191a24.4_2	uncharacterized LOC118322462	118322462	ZFIN:ZDB-GENE-030131-4489	MARVELD3	30525	270	0.172	-2.48	0.41
ccne2_1	cyclin E2	118336365	ZDB-GENE-030131-9689	CCNE2	1590	271	0.17	1.57	0.39
adnpb_2	activity-dependent neuroprotector homeobox b	118324792	ZFIN:ZDB-GENE-030131-6385	ADNP	15766	272	0.167	-2.35	0.26
LOC118338530_1	uncharacterized	118338530	-	SAP25	41908	273	0.159	2.52	0.52
LOC118337894_6	receptor-type tyrosine-protein phosphatase delta	118337894	ZDB-GENE-101103-16	PTPRD	9668	274	0.159	5.62	3.83E-04
opn9_1	opsin 9	118336070	ZFIN:ZDB-GENE-141216-390	OPN5	19992	275	0.156	3.01	0.15
LOC118321130_2	cadherin-7-like, transcript variant X2, mRNA	118321130	ZDB-GENE-061019-3	CDH7	1766	276	0.156	-2.34	0.45
hs3st1l1_1	heparan sulfate (glucosamine) 3-O-sulfotransferase 1-like1	118329507	ZFIN:ZDB-GENE-070202-2	HS3ST1	5194	277	0.156	-2.13	0.71
LOC118324909_1	protein kinase C zeta type-like	118324909	-	PRK CZ	9412	278	0.156	1.48	1

Appendix B (continued).

gstcd_3	glutathione S-transferase, C-terminal domain containing	118327103	ZDB-GENE- 050320-41	GSTCD	25806	279	0.156	-4.99	7.09E-03
pex5la_1	peroxisomal biogenesis factor 5-like a	118322120	ZFIN:ZDB-GENE- 070705-298	PEX5L	30024	280	0.143	2.25	0.79
atf7ip2_1	activating transcription factor 7 interacting protein 2	118323627	ZDB-GENE- 131121-384	ATF7IP2	20397	281	0.143	4.43	5.68E-06
c1qtnf4_1	C1q and TNF related 4	118343581	ZDB-GENE- 050417-198	C1QTNF4	14346	282	0.143	2.36	0.38
LOC118330215_1	progesterin and adipoQ receptor family member 4- like	118330215	-	PAQR4	26386	283	0.115	1.6	1
shox2_1	short stature homeobox 2	118341895	ZDB-GENE- 040426-1457	SHOX2	10854	284	0.115	-2.13	0.71

Appendix C

Gene transcripts quantitated in striped bass (SB, *Morone saxatilis*) using CLC Genomics Workbench (v.21.0.3, Qiagen Inc., Hilden, Germany) RNA-Seq Analysis tool and reference genome sequence assembly for these fish published through National Institutes of Health National Center for Biotechnology Information (NCBI) GenBank® (NCSU_SB_2.0, accession: GCA_004916995.1). These transcripts were identified as the optimal dataset for the comparison of SB into groups based on female parentage of being produced by one of two female SB, Dam A or Dam B. The optimal dataset was determined via application of a machine learning (ML) pipeline whereby attributes (gene transcripts) are assigned an information gain value (Column: Info Gain), or the amount of information the inclusion of data associated to a given attribute provides the ML model for the classification of instances (individuals) into comparison classes (groups). Attributes were assigned a rank based upon the information gain value such that the first attribute (#1) has the highest information gain, or is the most informative to the classification of instances into groups (Column: Rank). Ranked attributes were then processed through four ML algorithms each twice cross-validated and whereby bottom-ranking attributes were eliminated in a recursive manner to determine points of improvement and degradation of model performance. The optimal dataset is that which yielded optimum performance of sorting individual SB into groups from as many cross-validated models as possible. SB gene and gene transcript information output from CLC Genomics Workbench is provided in columns: SB Gene, which is the SB gene symbol followed by an underscore and the transcript number; SB Gene Name; SB Gene ID, which is the NCBI ID number for a given gene. The IDs of orthologs of a given gene for zebrafish (*Danio rerio*) (Column: ZFIN ID) and humans (*Homo sapiens*) are also provided (Column: HGNC Symbol, HGNC ID). The log₂ fold change (Column: Log₂ FC) is the log₂ of the fold change values calculated as the maximum mean between the groups divided by the minimum mean, and multiplied by -1 if the maximum value was the second in order from left to right between observations. Statistical tests were performed as a component of the CLC Genomics Workbench Differential Expression for RNA-Seq analysis tool, which employs the Wald test (“All group pairs”) for comparisons of two groups. The false discovery rate-corrected p-value (Column: FDR P-Value) was used in subsequent pathway analyses.

SB Gene	SB Gene Name	SB Gene ID	ZFIN ID	HGNC Symbol	HGNC ID	Rank	Info Gain	Log ₂ FC	FDR P-Value
nudt19_1	nudix (nucleoside diphosphate linked moiety X)-type motif 19	118320987	ZDB-GENE-081022-80	NUDT19	32036	1	0.3268	1.49	0.45
LOC118336003_1	protein LYRIC-like	118336003	-	MTDH	29608	2	0.3215	-0.26	0.95
cunh19orf25_1	chromosome unknown C19orf25 homolog	118321916	ZFIN:ZDB-GENE-131121-81	C19orf25	26711	3	0.308	-0.35	0.85
LOC118322835_1	uncharacterized LOC118322835	118322835	ZDB-GENE-141216-161	-	-	4	0.2669	0.68	2.06E-04
fn3krp_1	fructosamine 3 kinase related protein	118322984	ZDB-GENE-041010-192	KT3K	25700	5	0.251	0.19	1.00
LOC118335407_1	uncharacterized LOC118335407	118335407	ZDB-GENE-141222-90	SLC47A1	25588	6	0.2465	1.5	1.87E-03

Appendix C (continued).

lrrc31_1	leucine rich repeat containing 31	118343482	ZDB-GENE-091204-286	LRRC31	26261	7	0.2464	-1.92	3.03E-05
LOC118342602_1	EF-hand calcium-binding domain-containing protein 6-like	118342602	-	EFCAB6	24204	8	0.2456	-1.35	0.02
lingo3a_6	leucine rich repeat and Ig domain containing 3a	118322241	ZFIN:ZDB-GENE-060503-54	LINGO3	21206	9	0.2103	-3.44	1.73E-03
copz2_1	COPI coat complex subunit zeta 2	118322629	ZDB-GENE-000406-5	COPZ2	19356	10	0.2103	0.36	0.48
si:dkey-12112.1_1	uncharacterized si:dkey-12112.1	118325131	ZFIN:ZDB-GENE-091204-232	-	-	11	0.2103	0.25	1.00
LOC118342345_1	adhesion G protein-coupled receptor B1-like	118342345	-	ADGRB1	943	12	0.2091	2.99	8.57E-04
LOC118335406_2	fuclectin-4-like	118335406	-	-	-	13	0.2079	1.74	0.06
LOC118327155_1	oocyte zinc finger protein XICOF28-like	118327155	-	ZNF28	13073	14	0.2073	1.68	0.22
nudt19_3	nudix (nucleoside diphosphate linked moiety X)-type motif 19	118320987	ZDB-GENE-081022-80	NUDT19	32036	15	0.2004	-2.67	1.87E-03
LOC118331156_1	protein FAM111A-like	118331156	-	FAM111A	24725	16	0.1927	2.04	0.04
LOC118336722_1	coagulation factor IX-like	118336722	-	F9	3551	17	0.1904	-2.51	2.38E-03
LOC118335406_1	fuclectin-4-like	118335406	-	-	-	18	0.189	1.44	0.02
kcj19b_1	potassium inwardly rectifying channel subfamily J member 19b	118335852	ZFIN:ZDB-GENE-121214-60	KCNJ2	6263	19	0.1876	0.78	0.26
LOC118321221_1	uncharacterized LOC118321221, ncRNA	118321221	ZDB-GENE-040426-1383	KCMF1	20589	20	0.1862	0.51	0.74
LOC118336366_1	uncharacterized LOC118336366	118336366	-	TP53INP1	18022	21	0.1814	-1.15	0.01
LOC118331157_1	protein FAM111A-like	118331157	-	FAM111A	24725	22	0.1812	2.28	0.01
LOC118332202_2	guanine nucleotide-binding protein G(q) subunit alpha-like	118332202	-	GNAQ	4390	23	0.1812	1.08	0.45
LOC118327865_1	endoplasmic reticulum protein SC65-like	118327865	-	P3H4	16946	24	0.181	0.66	0.53
LOC118336859_1	uncharacterized LOC118336859	118336859	-	-	-	25	0.1799	-0.3	0.40
tmem269_4	transmembrane protein 269	118324579	ZDB-GENE-090313-227	TMEM269	52381	26	0.1748	0.73	1.00

Appendix C (continued).

hapln1a_1	hyaluronan and proteoglycan link protein 1a	118325400	ZFIN:ZDB-GENE-050302-175	HAPLN1	2380	27	0.1736	0.44	0.20
LOC118324455_1	uncharacterized LOC118324455	118324455	-	TPD52L2	12007	28	0.1723	0.42	0.89
si:ch73-116o1.2_2	ras-related protein M-Ras	118339011	ZFIN:ZDB-GENE-110411-67	MRAS	7227	29	0.1723	0.42	1.00
cryl1_1	crystallin, lambda 1	118339577	ZDB-GENE-060810-7	CRYL1	18246	30	0.1723	-0.45	0.52
LOC118335493_2	protein rapunzel-like	118335493	ZDB-GENE-070117-651	-	-	31	0.1722	0.68	1.00
LOC118342530_1	NXPE family member 3-like	118342530	-	NXPE3	28238	32	0.1722	1.39	0.48
nelfcd_1	negative elongation factor complex member C/D	118328722	ZDB-GENE-040426-720	NELFCD	15934	33	0.1718	0.34	0.92
si:ch73-70k4.1_1	uncharacterized si:ch73-70k4.1	118324599	ZFIN:ZDB-GENE-030131-3788	FAAP20	26428	34	0.17	0.55	1.00
si:ch211-151p13.8_1	UPF0687 protein C20orf27 homolog	118327544	ZFIN:ZDB-GENE-141212-230	C20orf27	15873	35	0.1697	1.21	0.67
LOC118328220_1	RNA-binding protein 39-like	118328220	-	RBM39	15923	36	0.1681	0.54	0.47
fam234a_3	family with sequence similarity 234 member A	118335361	ZDB-GENE-091204-17	F234A	14163	37	0.1676	1.19	0.87
atl2_1	atlastin GTPase 2	118329286	ZDB-GENE-030131-6505	ATL2	24047	38	0.1634	-0.18	1.00
wbp4_1	WW domain binding protein 4	118340046	ZDB-GENE-050522-342	WBP4	12739	39	0.1615	2.12	0.23
LOC118338792_1	obg-like ATPase 1	118338792	-	OLA1	28833	40	0.1591	-0.31	0.68
acsf3_3	acyl-CoA synthetase family member 3	118320756	ZDB-GENE-121026-3	ACSF3	27288	41	0.1591	-0.62	0.41
b3galt6_2	UDP-Gal:betaGal beta 1,3-galactosyltransferase polypeptide 6	118327855	ZDB-GENE-101104-13	B3GALT6	17978	42	0.1585	0.58	1.00
LOC118320856_1	eukaryotic translation initiation factor 5A-1-like	118320856	ZDB-GENE-040426-2229	EIF5A	3300	43	0.1585	0.31	0.69
cdin1_2	CDAN1 interacting nuclease 1	118337500	ZDB-GENE-060929-922	CDIN1	26929	44	0.1575	1.6	0.25
LOC118330690_1	B-cadherin-like	118330690	-	CDH26	15902	45	0.1575	0.8	0.78

Appendix C (continued).

LOC118340337_1	uncharacterized LOC118340337	118340337	-	TAF11	11544	46	0.1575	-0.36	1.00
obscnb_4	obscurin, cytoskeletal calmodulin and titin- interacting RhoGEF b	118334967	ZFIN:ZDB-GENE- 070119-5	OBSCN	15719	47	0.1575	2.56	0.22
LOC118321291_1	BMP and activin membrane-bound inhibitor homolog, mRNA	118321291	ZDB-GENE- 010416-1	BAMBI	30251	48	0.1575	0.26	1.00
exoc1_1	exocyst complex component 1	118321951	ZDB-GENE- 030131-1057	EXOC1	30380	49	0.1575	0.43	1.00
shmt1_1	serine hydroxymethyltransferase 1 (soluble)	118322772	ZDB-GENE- 040426-1558	SHMT1	10850	50	0.1575	-0.21	1.00
mastl_2	microtubule associated serine/threonine kinase- like	118320747	ZDB-GENE- 040801-128	MASTL	19042	51	0.1566	-1.6	0.47
si:ch1073- 358c10.1_1	mitochondrial peptide methionine sulfoxide reductase	118337601	ZFIN:ZDB-GENE- 050309-123	MSRA	7377	52	0.1547	-0.32	1.00
eml2_2	EMAP like 2	118341662	ZDB-GENE- 050706-71	EML2	18035	53	0.1547	-0.52	0.05
capns1a_2	calpain, small subunit 1 a	118331408	ZFIN:ZDB-GENE- 030113-3	CAPNS1	1481	54	0.1547	-0.29	1.00
mbnl2_1	muscleblind-like splicing regulator 2	118339856	ZDB-GENE- 030131-9582	MBNL2	16746	55	0.1547	-0.41	0.82
nup42_2	nucleoporin 42	118342886	ZDB-GENE- 040426-2292	NUP42	17010	56	0.1537	-0.71	0.66
LOC118329757_1	dual specificity protein phosphatase 13-like	118329757	-	DUSP13	19681	57	0.1537	0.6	0.59
mat2b_1	methionine adenosyltransferase II, beta	118341134	ZDB-GENE- 030131-786	MAT2B	6905	58	0.1537	0.29	0.76
admb_1	adrenomedullin b	118330777	ZFIN:ZDB-GENE- 120221-6	ADM	259	59	0.1536	2.16	0.05
LOC118322644_1	meiosis inhibitor protein 1	118322644	mei1	MEI1	28613	60	0.1536	1.45	0.22
LOC118334479_1	RING finger protein 212B- like	118334479	-	RNF212B	20438	61	0.1536	1.62	0.10
LOC118332659_1	uncharacterized LOC118332659	118332659	-	SLC9A3R2	11076	62	0.1536	0.88	0.67

Appendix C (continued).

LOC118338748_1	sterol 26-hydroxylase, mitochondrial	118338748	ZFIN:ZDB-GENE-081104-519	CYP27A1	2605	63	0.1536	0.85	1.00
arfp2b_10	ADP-ribosylation factor interacting protein 2b	118331386	ZFIN:ZDB-GENE-050417-131	ARFIP2	17160	64	0.1505	0.31	1.00
dhrs11a_4	dehydrogenase/reductase (SDR family) member 11a	118332067	ZFIN:ZDB-GENE-060929-324	DHRS11	28639	65	0.1505	0.57	1.00
LOC118330114_1	pyruvate kinase PKM-like	118330114	-	PKM	9021	66	0.1505	0.1	1.00
zgc:113276_1	uncharacterized	118321178	ZFIN:ZDB-GENE-050522-7	-	-	67	0.1479	-0.51	0.05
cst3_1	LOC118321178 cystatin C (amyloid angiopathy and cerebral hemorrhage)	118329191	ZDB-GENE-030131-373	CST3	2475	68	0.1479	-0.33	0.82
mtfr1_1	mitochondrial fission regulator 1	118323612	ZDB-GENE-041212-86	MTFR1	29510	69	0.1479	0.56	0.76
rab11fip4a_1	RAB11 family interacting protein 4 (class II) a	118322579	ZFIN:ZDB-GENE-040718-266	RAB11FIP4	30267	70	0.1479	1.8	0.05
si:ch73-264p11.1_1	uncharacterized	118326457	ZFIN:ZDB-GENE-120703-35	SH2D1B	30416	71	0.1463	0.71	0.92
LOC118331195_1	LOC118331195 uncharacterized	118331195	-	TENT5C	24712	72	0.1463	-1.23	0.58
LOC118327601_1	lysozyme g-like	118327601	-	LYG1	27014	73	0.1459	0.51	0.80
pip4p2_1	phosphatidylinositol-4,5-bisphosphate 4-phosphatase 2	118342645	ZDB-GENE-060503-854	PIP4P2	25452	74	0.1458	-0.33	0.63
trmt10a_1	tRNA methyltransferase 10A	118338328	ZDB-GENE-130530-594	TRMT10A	28403	75	0.1458	-0.44	0.64
LOC118343084_2	formyl peptide receptor 2-like	118343084	-	FPR2	3827	76	0.1458	0.89	1.00
asip1_2	agouti signaling protein 1	118328502	ZFIN:ZDB-GENE-060215-1	ASIP	745	77	0.1438	0.84	0.92
crtc1a_9	CREB regulated transcription coactivator 1a	118334565	ZFIN:ZDB-GENE-061027-225	CRTC1	16062	78	0.1438	-1.94	0.24
tpte_1	transmembrane phosphatase with tensin homology	118331449	ZDB-GENE-030131-5503	TPTE	12023	79	0.1434	0.52	0.90
snx8a_2	sorting nexin 8a	118323507	ZFIN:ZDB-GENE-031202-1	SNX8	14972	80	0.1434	0.29	1.00
efcab14_1	EF-hand calcium binding domain 14	118334262	ZDB-GENE-040426-2469	EFCAB14	29051	81	0.1414	0.19	1.00

Appendix C (continued).

rbpjl_1	recombination signal binding protein for immunoglobulin kappa J region-like	118328757	ZDB-GENE-050307-4	RBPJL	13761	82	0.1414	0.74	0.46
etfbkmt_1	electron transfer flavoprotein subunit beta	118330633	ZDB-GENE-111107-1	ETKMT	28739	83	0.1414	-0.76	0.78
stat1a_3	lysine methyltransferase signal transducer and activator of transcription 1a	118339891	ZFIN:ZDB-GENE-980526-499	STAT1	11362	84	0.1414	-1.22	0.48
pmt_1	phosphoethanolamine methyltransferase	118333199	ZFIN:ZDB-GENE-060929-740	PEMT	8830	85	0.1414	-2.5	4.74E-04
nbn_2	nibrin	118335919	ZDB-GENE-041008-35	NBN	7652	86	0.1406	0.75	1.00
pum2_3	pumilio RNA-binding family member 2	118340632	ZDB-GENE-081031-76	PUM2	14958	87	0.1406	1.86	0.60
slc37a3_2	solute carrier family 37 member 3	118341307	ZDB-GENE-170302-2	SLC37A3	20651	88	0.1406	-0.25	1.00
zgc:101731_1	SNARE_SNAP25N and SNARE_SNAP23C domain-containing protein	118341476	ZFIN:ZDB-GENE-040912-57	SNAP25	11132	89	0.1394	2.33	6.13E-03
cpeb4b_1	cytoplasmic polyadenylation element binding protein 4b	118331227	ZFIN:ZDB-GENE-170601-167	CPEB4	21747	90	0.1388	0.36	1.00
LOC118327626_4	beta-1,3-galactosyltransferase 2-like	118327626	-	B3GALT2	917	91	0.1388	0.86	1.00
coil_1	coilin p80	118323593	ZDB-GENE-000330-8	COIL	2184	92	0.1388	0.15	1.00
ift140_1	intraflagellar transport 140 homolog (Chlamydomonas)	118329018	ZDB-GENE-040724-165	IFT140	29077	93	0.1388	0.96	0.20
LOC118338835_1	titin-like	118338835	-	TTN	12403	94	0.1388	0.27	1.00
dock7_4	dedicator of cytokinesis 7	118321379	ZDB-GENE-050302-15	DOCK7	19190	95	0.1388	0.32	1.00
arfip1_2	ADP-ribosylation factor interacting protein 1 (arfaptin 1)	118326904	ZDB-GENE-040426-2690	ARFIP1	21496	96	0.1388	-0.19	1.00
clta_2	clathrin, light chain A	118326474	ZDB-GENE-040426-1986	CLTA	2090	97	0.1381	0.12	1.00

Appendix C (continued).

bhlhe41_1	basic helix-loop-helix family, member e41	118330410	ZDB-GENE-050419-146	BHLHE41	16617	98	0.1381	-0.93	1.00
mab21l2_1	mab-21-like 2	118326870	ZDB-GENE-011101-3	MAB21L2	6758	99	0.1381	-0.78	1.00
LOC118333712_1	adenosine kinase-like	118333712	-	ADK	257	100	0.1381	-0.48	0.20
LOC118323242_2	oxysterol-binding protein-related protein 7-like, transcript variant X2, mRNA	118323242	-	OSBPL7	16387	101	0.1381	-1.74	0.08
LOC118327915_2	rho-related GTP-binding protein RhoA-B	118327915	-	RHOA	667	102	0.1381	-0.51	0.69
LOC118332997_4	uncharacterized LOC118332997	118332997	-	-	-	103	0.1381	0.29	1.00
LOC118331229_1	oligodendrocyte-myelin glycoprotein-like	118331229	-	MOG	7197	104	0.1344	-1.85	0.12
arhgap9_3	Rho GTPase activating protein 9	118328538	ZDB-GENE-110510-2	ARHGAP9	14130	105	0.1326	0.92	1.00
zc3h14_1	zinc finger CCCH-type containing 14	118340867	ZDB-GENE-041014-257	ZC3H14	20509	106	0.13	-0.29	0.82
LOC118330714_1	uncharacterized LOC118330714	118330714	-	NUP160	18017	107	0.13	-0.24	1.00
zbtb25_1	zinc finger and BTB domain containing 25	118340567	ZDB-GENE-041001-125	ZBTB25	13112	108	0.13	0.81	0.86
dnaaf6_2	dynein axonemal assembly factor 6	118334620	ZDB-GENE-040722-2	DNAAF6	28570	109	0.13	0.51	1.00
chrna11_1	cholinergic receptor, nicotinic, alpha 11	118343170	ZFIN:ZDB-GENE-060503-606	CHRFAM7A	15781	110	0.13	0.58	1.00
pir_1	pirin	118321179	ZDB-GENE-040718-288	PIR	30048	111	0.13	-0.61	0.87
si:rp71-1d10.8_1	uncharacterized si:rp71-1d10.8	118331567	ZFIN:ZDB-GENE-141216-274	TIGD1	14523	112	0.13	0.3	1.00
LOC118328040_1	vimentin-like	118328040	-	VIM	12692	113	0.13	1.14	0.52
gale_1	UDP-galactose-4-epimerase	118337530	ZDB-GENE-060421-6479	GALE	4116	114	0.1273	1.82	0.30
rhot1b_1	ras homolog family member T1	118327217	ZFIN:ZDB-GENE-061009-52	RHOT1	21168	115	0.1242	-2.47	0.22
cep19_1	centrosomal protein 19	118321626	ZDB-GENE-050913-81	CEP19	28209	116	0.1242	-2.78	0.07

Appendix C (continued).

LOC118331103_1	small nucleolar RNA SNORA3/SNORA45 family	118331103	-	SNORA3B	32638	117	0.1237	1.71	0.69
LOC118338556_1	small nucleolar RNA U3	118338556	-	SNORD3A	33189	118	0.1237	-0.12	1.00
LOC118339926_1	cytochrome b-245 heavy chain	118339926	ZFIN:ZDB-GENE- 040426-1380	CYBB	2578	119	0.1237	0.29	1.00
dcst2_1	DC-STAMP domain containing 2	118336547	ZDB-GENE- 070809-6	DCST2	26562	120	0.1237	1.01	0.73
imp3_1	IMP U3 small nucleolar ribonucleoprotein 3	118334613	ZDB-GENE- 040426-1062	IMP3	14497	121	0.1237	-0.28	1.00
crtc1a_4	CREB regulated transcription coactivator 1a	118334565	ZFIN:ZDB-GENE- 061027-225	CRTC1	16062	122	0.1237	4.22	1.02E-05
amigo3_1	adhesion molecule with Ig- like domain 3	118328215	ZDB-GENE- 050208-449	AMIGO3	24075	123	0.1237	0.88	0.83
kitb_1	KIT proto-oncogene, receptor tyrosine kinase b	118326556	ZFIN:ZDB-GENE- 050916-2	KIT	6342	124	0.1237	3.05	2.24E-03
syn2b_1	synapsin IIb	118328145	ZFIN:ZDB-GENE- 051127-49	SYN2	11495	125	0.1237	1.52	0.92
aurka_2	aurora kinase A	118327710	ZDB-GENE- 040801-161	AURKA	11393	126	0.1237	0.52	1.00
LOC118321072_1	proteolipid protein 2	118321072	under "cmtm"	CKLF	13253	127	0.1237	1.03	0.31
rwdd2b_2	RWD domain containing 2B	118341543	ZDB-GENE- 041001-119	RWDD2B	1302	128	0.1237	0.29	1.00
mybl1_2	v-myb avian myeloblastosis viral oncogene homolog-like 1	118325162	ZDB-GENE- 041111-281	MYBL1	7547	129	0.1237	0.31	1.00
rxrb_3	retinoid x receptor, beta b	118335657	ZFIN:ZDB-GENE- 990415-242	RXRB	10478	130	0.1237	0.86	0.89
pdc11_4	programmed cell death 11	118326580	ZDB-GENE- 030131-4076	PDCD11	13408	131	0.1237	1.99	0.74
mpc2b_1	mitochondrial pyruvate carrier 2b	118338939	ZFIN:ZDB-GENE- 030131-330	MPC2	24515	132	0.1237	-0.06	1.00
ehbp1_4	EH domain binding protein 1	118329407	ZDB-GENE- 091006-4	EHBP1	29144	133	0.1237	4.41	2.60E-05
si:ch211- 261d7.6_1	zinc finger protein 91	118335720	ZFIN:ZDB-GENE- 141216-266	ZNF91	13166	134	0.1237	0.25	1.00
LOC118334640_2	WD repeat-containing protein 37	118334640	-	WDR37	31406	135	0.1237	1.15	1.00

Appendix C (continued).

LOC118328428_1	potassium voltage-gated channel subfamily B member 1-like	118328428	-	KCNB1	6231	136	0.1237	1.9	0.56
colgalt2_1	collagen beta(1-O)galactosyltransferase 2	118321751	ZDB-GENE-070222-1	COLGALT2	16790	137	0.1237	0.34	1.00
LOC118342719_1	probable thiopurine S-methyltransferase	118342719	-	TPMT	12014	138	0.1237	0.28	1.00
asic1c_1	acid-sensing (proton-gated) ion channel 1c	118334460	ZFIN:ZDB-GENE-040513-3	ASIC1	100	139	0.1237	-0.66	0.83
LOC118342540_2	progranulin-like	118342540	-	GRN	4601	140	0.1237	1.08	1.00
LOC118342539_1	NACHT, LRR and PYD domains-containing protein 12-like	118342539	-	NLRP12	22938	141	0.1237	1.02	0.24
nap1l1_1	nucleosome assembly protein 1-like 1	118337203	ZDB-GENE-030516-2	NAP1L1	7637	142	0.1237	0.19	1.00
lipt1_1	lipoyltransferase 1	118342398	ZDB-GENE-060929-112	LIPT1	29569	143	0.1237	-0.00358	1.00
irak3_1	interleukin-1 receptor-associated kinase 3	118337895	ZDB-GENE-060503-710	IRAK3	17020	144	0.1237	0.58	0.26
tmem167a_1	transmembrane protein 167A	118325835	ZDB-GENE-050320-29	TMEM167A	28330	145	0.1237	0.06	1.00
si:dkey-40m6.8_1	uncharacterized si:dkey-40m6.8	118335758	ZFIN:ZDB-GENE-100812-13	LMTK3	19295	146	0.1237	1.59	0.37
eps15_2	epidermal growth factor receptor pathway substrate 15	118321661	ZDB-GENE-081104-264	EPS15	3419	147	0.1237	0.65	1.00
poc5_2	POC5 centriolar protein homolog (Chlamydomonas)	118326292	ZDB-GENE-060526-135	POC5	26658	148	0.1237	1.4	0.92
LOC118335389_1	glucocorticoid modulatory element-binding protein 1-like	118335389	-	GMEB1	4370	149	0.1237	0.94	0.88
spef1_1	sperm flagellar 1	118329506	ZDB-GENE-040426-1639	SPEF1	15874	150	0.1237	0.83	0.80

Appendix D

Gene transcripts quantitated in striped bass (SB, *Morone saxatilis*) using CLC Genomics Workbench (v.21.0.3, Qiagen Inc., Hilden, Germany) RNA-Seq Analysis tool and reference genome sequence assembly for these fish published through National Institutes of Health National Center for Biotechnology Information (NCBI) GenBank® (NCSU_SB_2.0, accession: GCA_004916995.1). These transcripts were identified as the optimal dataset for the comparison of SB into groups based on individual SB sire determined to be the parent of an individual through microsatellite genotyping. The optimal dataset was determined via application of a machine learning (ML) pipeline whereby attributes (gene transcripts) are assigned an information gain value (Column: Info Gain), or the amount of information the inclusion of data associated to a given attribute provides the ML model for the classification of instances (individuals) into comparison classes (groups). Attributes were assigned a rank based upon the information gain value such that the first attribute (#1) has the highest information gain, or is the most informative to the classification of instances into groups (Column: Rank). Ranked attributes were then processed through four ML algorithms each twice cross-validated and whereby bottom-ranking attributes were eliminated in a recursive manner to determine points of improvement and degradation of model performance. The optimal dataset is that which yielded optimum performance of sorting individual SB into groups from as many cross-validated models as possible. SB gene and gene transcript information output from CLC Genomics Workbench is provided in columns: SB Gene, which is the SB gene symbol followed by an underscore and the transcript number; SB Gene Name; SB Gene ID, which is the NCBI ID number for a given gene. The IDs of orthologs of a given gene for zebrafish (*Danio rerio*) (Column: ZFIN ID) and humans (*Homo sapiens*) are also provided (Column: HGNC Symbol, HGNC ID). The log₂ fold change (Column: Log₂ FC) is the log₂ of the fold change values calculated as the maximum mean between the groups divided by the minimum mean, and multiplied by -1 if the maximum value was the second in order from left to right between observations. Statistical tests were performed as a component of the CLC Genomics Workbench Differential Expression for RNA-Seq analysis tool, which employs the Likelihood ratio test (“Across groups (ANOVA-like)”) Likelihood ratio test (“Across groups (ANOVA-like)”) for comparisons of more than two groups. The false discovery rate-corrected p-value (Column: FDR P-Value) was used in subsequent pathway analyses.

SB Gene	SB Gene Name	SB Gene ID	ZFIN ID	HGNC Symbol	HGNC ID	Rank	Info Gain	Log ₂ FC	FDR P-Value
zgc:103759_2	U8 snoRNA-decapping enzyme	118324394	ZFIN:ZDB-GENE-041010-100	NUDT16	26442	1	0.822	-10.91	3.22E-07
pum3_2	pumilio RNA-binding family member 3	118332424	ZDB-GENE-030131-9808	PUM3	29676	2	0.683	11.02	0.00
LOC118342532_1	glutathione S-transferase A-like	118342532	ZFIN:ZDB-GENE-071004-68	GST	4628	3	0.641	-1.16	2.82E-05
LOC118322835_1	uncharacterized LOC118322835	118322835	ZDB-GENE-141216-161	-	-	4	0.628	1.46	2.19E-06
LOC118331157_1	protein FAM111A-like	118331157	-	FAM111A	24725	5	0.622	-7.58	2.69E-11

Appendix D (continued).

xrcc4_1	X-ray repair complementing defective repair in Chinese hamster cells 4	118333082	ZDB-GENE-040426-1755	XRCC4	12831	6	0.608	-1.94	9.93E-03
LOC118331210_1	uncharacterized LOC118331210	118331210	-	SMARCAL1	11102	7	0.601	8.85	0.00
LOC118341073_1	androgen-induced gene 1 protein-like	118341073	-	AIG1	21607	8	0.567	1.22	0.12
cipca_2	CLOCK-interacting pacemaker a	118337364	ZFIN:ZDB-GENE-120928-5	CIPC	20365	9	0.566	-4.14	0.45
mrps24_2	mitochondrial ribosomal protein S24	118333129	ZDB-GENE-050522-393	MRPS24	14510	10	0.557	10.15	0.00
LOC118331156_1	protein FAM111A-like	118331156	-	FAM111A	24725	11	0.557	-6.82	8.11E-11
lrrc31_1	leucine rich repeat containing 31	118343482	ZDB-GENE-091204-286	LRRC31	26261	12	0.547	-10.34	0.00
mfsd4ab_1	major facilitator superfamily domain containing 4Ab	118324758	ZFIN:ZDB-GENE-060810-38	MFSD4A	25433	13	0.53	-2.81	5.03E-04
LOC118331024_1	uncharacterized LOC118331024	118331024	-	IMMP2L	14598	14	0.516	1.84	0.61
cmtr1_1	cap methyltransferase 1	118340617	ZDB-GENE-040426-696	CMTR1	21077	15	0.512	-2.76	0.10
pelo_1	pelota mRNA surveillance and ribosome rescue factor	118330127	ZDB-GENE-040426-1074	PELO	8829	16	0.51	1.04	0.53
LOC118328008_3	calglandulin-like	118328008	-	CALML6	24193	17	0.504	4.22	0.00125
LOC118327915_3	rho-related GTP-binding protein RhoA-B	118327915	-	RHOA	667	18	0.497	1.73	5.79E-03
selenoo2_1	selenoprotein O2	118340647	ZFIN:ZDB-GENE-041210-110	SELENOO	30395	19	0.489	1.21	6.98E-03
mettl27_1	methyltransferase like 27	118331511	ZDB-GENE-060421-5918	METTL27	19068	20	0.487	-1.74	1.52E-03
LOC118333040_1	inactive phospholipid phosphatase 7	118333040	ZDB-GENE-040426-1822	PLPP7	28174	21	0.48	12.61	1.14E-07
si:ch211-266g18.10_22	trichohyalin	118337429	ZFIN:ZDB-GENE-131127-188	TCHH	11791	22	0.471	2.92	0.40
nat10_1	N-acetyltransferase 10	118330657	ZDB-GENE-040426-1543	NAT10	29830	23	0.456	1.5	0.12
dnajc4_1	DnaJ (Hsp40) homolog, subfamily C, member 4	118333426	ZDB-GENE-030131-3093	DNAJC4	5271	24	0.433	-1.17	0.76

Appendix D (continued).

LOC118324184_1	YTH domain-containing family protein 1-like	118324184	-	YTHDF1	15867	25	0.432	1.59	1.85E-05
LOC118334072_1	sequestosome-1-like	118334072	ZDB-GENE-040426-2204	SQSTM1	11280	26	0.428	2.31	0.52
dennd2da_1	DENN/MADD domain containing 2Da	118323837	ZDB-GENE-061013-542	DENND2D	26192	27	0.426	-7	0.02
LOC118324589_1	thyrotropin-releasing hormone receptor-like	118324589	ZDB-GENE-100922-18	TRHR	12299	28	0.419	3.53	0.09
nudt19_2	nudix (nucleoside diphosphate linked moiety X)-type motif 19	118320987	ZDB-GENE-081022-80	NUDT19	32036	29	0.417	8.83	3.83E-12
sccpdhb_1	saccharopine dehydrogenase b	118340807	ZDB-GENE-041010-211	SCCPDH	24275	30	0.407	1.62	9.70E-03
mrpl28_1	mitochondrial ribosomal protein L28	118342233	ZDB-GENE-050522-113	MRPL28	14484	31	0.372	1.79	0.29
LOC118335617_10	aryl hydrocarbon receptor nuclear translocator-like	118335617	ZDB-GENE-060126-7	ARNT	700	32	0.365	4.85	0.42
LOC118338698_1	contactin-associated protein-like 5	118338698	-	CNTNAP5	18748	33	0.35	-4.44	5.09E-03
LOC118332479_2	spindlin-1-like	118332479	-	SPIN1	11243	34	0.308	9.62	0.00
ltbp3_5	latent transforming growth factor beta binding protein 3	118331421	ZDB-GENE-060526-130	LTBP3	6716	35	0.25	9.72	0.76

Appendix E

Gene transcripts quantitated in striped bass (SB, *Morone saxatilis*) using CLC Genomics Workbench (v.21.0.3, Qiagen Inc., Hilden, Germany) RNA-Seq Analysis tool and reference genome sequence assembly for these fish published through National Institutes of Health National Center for Biotechnology Information (NCBI) GenBank® (NCSU_SB_2.0, accession: GCA_004916995.1). These transcripts were identified as the optimal dataset for the comparison of SB into groups based on whether the SB sires they were produced from belonged to a group, Large or Small, whereby sires significantly differed in weight and length and were crossed with the same female SB. The optimal dataset was determined via application of a machine learning (ML) pipeline whereby attributes (gene transcripts) are assigned an information gain value (Column: Info Gain), or the amount of information the inclusion of data associated to a given attribute provides the ML model for the classification of instances (individuals) into comparison classes (groups). Attributes were assigned a rank based upon the information gain value such that the first attribute (#1) has the highest information gain, or is the most informative to the classification of instances into groups (Column: Rank). Ranked attributes were then processed through four ML algorithms each twice cross-validated and whereby bottom-ranking attributes were eliminated in a recursive manner to determine points of improvement and degradation of model performance. The optimal dataset is that which yielded optimum performance of sorting individual SB into groups from as many cross-validated models as possible. SB gene and gene transcript information output from CLC Genomics Workbench is provided in columns: SB Gene, which is the SB gene symbol followed by an underscore and the transcript number; SB Gene Name; SB Gene ID, which is the NCBI ID number for a given gene. The IDs of orthologs of a given gene for zebrafish (*Danio rerio*) (Column: ZFIN ID) and humans (*Homo sapiens*) are also provided (Column: HGNC Symbol, HGNC ID). The log₂ fold change (Column: Log₂ FC) is the log₂ of the fold change values calculated as the maximum mean between the groups divided by the minimum mean, and multiplied by -1 if the maximum value was the second in order from left to right between observations. Statistical tests were performed as a component of the CLC Genomics Workbench Differential Expression for RNA-Seq analysis tool, which employs the Wald test (“All group pairs”) for comparisons of two groups. The false discovery rate-corrected p-value (Column: FDR P-Value) was used in subsequent pathway analyses.

SB Gene	SB Gene Name	SB Gene ID	ZFIN ID	HGNC Symbol	HGNC ID	Rank	Info Gain	Log ₂ FC	FDR P-Value
LOC118327915_3	rho-related GTP-binding protein RhoA-B	118327915	-	RHOA	667	1	0.3745	-0.68	1.60E-03
LOC118340196_1	uncharacterized	118340196	-	FGFR3	3690	2	0.3663	1.34	0.24
mrps24_2	mitochondrial ribosomal protein S24	118333129	ZDB-GENE-050522-393	MRPS24	14510	3	0.312	3.88	1.07E-05
xrcc4_1	X-ray repair complementing defective repair in Chinese hamster cells 4	118333082	ZDB-GENE-040426-1755	XRCC4	12831	4	0.3074	-0.74	4.25E-03

Appendix E (continued).

LOC118336868_1	carcinoembryonic antigen-related cell adhesion molecule 1-like	118336868	-	CEACAM1	1814	5	0.2953	0.41	0.81
pum3_2	pumilio RNA-binding family member 3	118332424	ZDB-GENE-030131-9808	PUM3	29676	6	0.2799	2.53	0.09
mrpl9_1	mitochondrial ribosomal protein L9	118342573	ZDB-GENE-070717-4	MRPL9	14277	7	0.2781	-0.31	0.46
pdc2l_1	programmed cell death 2-like	118330172	ZDB-GENE-040426-967	PDCD2L	28194	8	0.275	-0.58	0.04
mrps24_1	mitochondrial ribosomal protein S24	118333129	ZDB-GENE-050522-393	MRPS24	14510	9	0.2421	-0.74	0.01
cytip_2	cytohesin 1 interacting protein	118339416	ZDB-GENE-140106-76	CYTIP	9506	10	0.2392	0.68	0.76
LOC118339360_1	trans-1,2-dihydrobenzene-1,2-diol dehydrogenase-like	118339360	-	DHDH	17887	11	0.233	-0.94	0.11
LOC118321935_1	coiled-coil domain-containing protein 106-like	118321935	ZDB-GENE-041008-66	CCDC106	30181	12	0.2247	1.55	0.04
mpeg1.1_1	macrophage expressed 1, tandem duplicate 1	118326099	ZFIN:ZDB-GENE-030131-7347	MPEG1	29619	13	0.2199	0.52	0.39
csad_1	cysteine sulfenic acid decarboxylase	118324877	ZDB-GENE-041114-36	CSAD	18966	14	0.2185	-4.36	1.85E-06
mak16_1	MAK16 homolog	118325976	ZDB-GENE-020419-35	MAK16	13703	15	0.2168	-0.23	0.71
si:dkey-4e7.3_2	uncharacterized LOC118339323	118339323	ZFIN:ZDB-GENE-070912-542	UCK1	14859	16	0.2163	-0.83	0.23
pum3_1	pumilio RNA-binding family member 3	118332424	ZDB-GENE-030131-9808	PUM3	29676	17	0.2156	-0.5	0.11
ch25h_1	cholesterol 25-hydroxylase	118325894	ZDB-GENE-041212-81	CH25H	1907	18	0.2044	0.82	0.27
rbp5_1	retinol binding protein 1a, cellular	118336129	ZDB-GENE-020320-1	RBP5	15847	19	0.1997	0.46	0.49
gyg1b_2	glycogenin 1b	118324940	ZFIN:ZDB-GENE-040625-30	GYG1	4699	20	0.1995	-0.34	0.38
phrf1_3	PHD and ring finger domains 1	118329968	ZDB-GENE-030131-624	PHRF1	24351	21	0.1992	0.95	0.75
LOC118338698_1	contactin-associated protein-like 5	118338698	-	CNTNAP5	18748	22	0.1992	2.29	0.01
LOC118333303_1	protocadherin gamma-A2	118333303	-	PCDHGA2	8700	23	0.1992	1.87	0.07

Appendix E (continued).

LOC118331228_1	uncharacterized	118331228	-	PRPF4B	17346	24	0.1982	0.95	0.24
nipsnap3a_1	nipsnap homolog 3A (C. elegans)	118333403	ZDB-GENE-040426-1037	NIPSNAP3 A	23619	25	0.1975	-0.8	0.53
LOC118339886_1	ras-related protein Rab-9A-like	118339886	-	RAB9A	9792	26	0.1969	-0.4	0.39
cytip_1	cytohesin 1 interacting protein	118339416	ZDB-GENE-140106-76	CYTIP	9506	27	0.1933	-2.58	5.42E-03
LOC118338164_1	uncharacterized	118338164	-	TBC1D14	29246	28	0.1933	-1.22	0.05
psmg2_1	proteasome (prosome, macropain) assembly chaperone 2	118342587	ZDB-GENE-040426-1972	PSMG2	24929	29	0.193	-0.18	0.79
sidt2_1	SID1 transmembrane family, member 2	118341626	ZDB-GENE-030131-7356	SIDT2	24272	30	0.193	0.31	0.32
aqp9b_1	aquaporin 9b	118320889	ZFIN:ZDB-GENE-070911-1	AQP9	643	31	0.1843	1.16	0.18
kat14_1	lysine acetyltransferase 14	118326661	ZDB-GENE-040718-452	KAT14	15904	32	0.1843	-0.23	0.54
thap12b_2	THAP domain containing 12b	118341287	ZFIN:ZDB-GENE-040718-429	THAP12	9440	33	0.1817	0.63	0.34
fabp3_1	fatty acid binding protein 3, muscle and heart	118342717	ZDB-GENE-020318-2	FABP3	3557	34	0.1817	-0.55	0.04
cryba11_1	crystallin, beta A1, like 1	118333190	ZFIN:ZDB-GENE-050417-249	CRYBA1	2394	35	0.1817	0.88	0.46
tdrd15_1	tudor domain containing 15	118341103	ZDB-GENE-041014-303	TDRD15	45037	36	0.1812	-1.84	0.02
LOC118343217_1	retinoblastoma-like protein 2	118343217	-	RBL2	9894	37	0.1793	0.49	0.42
LOC118338081_2	dynactin subunit 1-like	118338081	-	DCTN1	2711	38	0.1793	0.51	0.92
LOC118338531_1	complement C1q-like protein 2	118338531	-	C1QL2	24181	39	0.1793	0.87	0.69
LOC118342792_1	uncharacterized	118342792	-	ELMO1	16286	40	0.1793	0.66	0.59
LOC118328289_1	transmembrane protein 198-like	118328289	-	TMEM198	33704	41	0.1778	0.99	0.64
cbfa2t2_1	CBFA2/RUNX1 partner transcriptional co-repressor 2	118328474	ZDB-GENE-070209-1	CBFA2T2	1536	42	0.1774	0.37	0.54
LOC118330380_1	transient receptor potential cation channel subfamily M member 1	118330380	ZDB-GENE-070112-1372	TRPM1	7146	43	0.1772	1.46	0.16

Appendix E (continued).

LOC118338255_2	protein phosphatase 1K, mitochondrial-like	118338255	-	PPM1K	25415	44	0.1751	0.63	0.17
LOC118340214_1	protein FAM107B-like	118340214	-	FAM107B	23726	45	0.1751	0.44	0.27
arhgap12b_4	Rho GTPase activating protein 12b	118342647	ZFIN:ZDB-GENE-040426-1727	ARHGAP12	16348	46	0.175	1.29	0.62
LOC118330175_1	lathosterol oxidase-like	118330175	-	SC5D	10547	47	0.1745	-0.6	0.74
LOC118321646_1	transmembrane protein 106A-like, mRNA	118321646	ZDB-GENE-040718-231	NXPH1	20693	48	0.1745	-0.29	0.62
mbnl2_11	muscleblind-like splicing regulator 2	118339856	ZDB-GENE-030131-9582	MBNL2	16746	49	0.1745	0.6	0.44
pdss2_1	prenyl (decaprenyl) diphosphate synthase, subunit 2	118337410	ZDB-GENE-040718-43	PDSS2	23041	50	0.1745	-0.19	0.80
riox2_1	ribosomal oxygenase 2	118331974	ZDB-GENE-040426-1283	RIOX2	19441	51	0.172	-0.37	0.32
mapk14b_2	mitogen-activated protein kinase 14b	118328003	ZDB-GENE-021007-1	MAPK14	6876	52	0.1698	0.15	0.86
ksr1a_2	kinase suppressor of ras 1a	118331852	ZDB-GENE-091113-34	KSR1	6465	53	0.1698	1.12	0.61
ngfrb_1	nerve growth factor receptor b	118335116	ZFIN:ZDB-GENE-070606-2	NGFR	7809	54	0.1691	0.96	0.01
pusl1_1	pseudouridine synthase like 1	118327994	ZDB-GENE-101110-1	PUSL1	26914	55	0.1687	-0.33	0.69
LOC118340618_1	trans-L-3-hydroxyproline dehydratase	118340618	ZDB-GENE-170412-1	L3HYPDH	20488	56	0.1687	-0.52	0.04
znf740a_5	zinc finger protein 740a	118328611	ZDB-GENE-060929-660	ZNF740	27465	57	0.1687	2.09	0.03
ebf2_1	EBF transcription factor 2	118326114	ZDB-GENE-990715-11	EBF2	19090	58	0.1687	0.99	0.54
abhd14b_1	abhydrolase domain containing 14B	118327789	ZDB-GENE-040426-1342	ABHD14B	28235	59	0.1682	0.7	0.43
tbc1d25_1	TBC1 domain family, member 25	118324155	ZDB-GENE-041111-25	TBC1D25	8092	60	0.1682	0.36	0.56
LOC118322369_1	E3 ubiquitin-protein ligase rnf213-alpha-like	118322369	ZDB-GENE-050302-100	RNF213	14539	61	0.1663	0.16	0.98
si:ch1073-143i10.2_1	autophagy-related protein 16	118329375	ZDB-GENE-050417-401	ATG16L1	21498	62	0.1663	-0.6	0.22
si:dkey-117m1.4_1	serine/arginine repetitive matrix protein 1	118326922	ZDB-GENE-040426-2789	SRRM1	16638	63	0.1662	0.36	0.67

Appendix E (continued).

gfm1_1	G elongation factor, mitochondrial 1	118341836	ZDB-GENE-061013-79	GFM1	13780	64	0.1662	-0.28	0.45
LOC118342532_1	glutathione S-transferase A-like	118342532	ZFIN:ZDB-GENE-071004-68	GST	4628	65	0.1662	0.3	0.62
skp1_1	S-phase kinase-associated protein 1	118331882	ZDB-GENE-040426-1707	SKP1	10899	66	0.1657	-0.17	0.79
LOC118343336_1	cytochrome P450 27C1	118343336	ZDB-GENE-080204-68	CYP27C1	33480	67	0.1657	0.83	0.45
LOC118340221_1	phosphatidylserine decarboxylase proenzyme, mitochondrial-like	118340221	ZDB-GENE-061215-46	PISD	8999	68	0.1643	-1.15	0.26
LOC118321438_1	uncharacterized LOC118321438	118321438	ZDB-GENE-090312-97	INPP4B	6075	69	0.1643	0.43	0.83
dnajb12b_1	DnaJ heat shock protein family (Hsp40) member B12b	118334834	ZFIN:ZDB-GENE-070410-128	DNAJB12	14891	70	0.1643	-0.23	0.66
LOC118332045_1	pterin-4-alpha-carbinolamine dehydratase 2-like	118332045	ZDB-GENE-070719-10	PCBD2	24474	71	0.1643	-0.47	0.23
pigo_2	phosphatidylinositol glycan anchor biosynthesis, class O	118332938	ZDB-GENE-091204-80	PIGO	23215	72	0.1643	0.31	0.91
lnx2b_2	ligand of numb-protein X 2b	118333512	ZDB-GENE-060228-2	LNX2	20421	73	0.1643	0.47	0.65
LOC118335406_1	fucolectin-4-like	118335406	-	-	-	74	0.1641	1.25	0.03
LOC118327707_1	deoxyribonuclease-1-like	118327707	ZDB-GENE-040426-2170	DNASE1	2956	75	0.1641	0.45	0.70

Appendix F

Molecules identified through Qiagen Ingenuity Pathway Analysis (Qiagen, Hilden, Germany) as the top analysis-ready molecules for comparisons of genes expressed in striped bass (SB, *Morone saxatilis*) and based on comparisons of growth performance (Superior or Inferior; input genes are provided in Appendix A), growth performance to market size (Under Size, Inferior Other, Superior Other, or Market Size; input genes are provided in Appendix B), female SB parentage (Dam A or Dam B; input genes are provided in Appendix C), sire parentage (Sires 1–12; input genes are provided in Appendix D), or sire size (Large or Small; input genes are provided in Appendix E). Analysis-ready molecules are considered those that provide the most information to pathway and network building for a given comparison. The order column indicates the order of importance where #1 is the most important up- or down-regulated molecule for a given comparison group. The orthologous human (*Homo sapiens*) gene symbol is provided in the Symbol column. The log₂ fold change (Log₂ FC) is the log₂ of the fold change values calculated as the maximum mean of expression measured in RPKM between the groups divided by the minimum mean, and multiplied by -1 if the maximum value was the second in order from left to right between observations. The p-value and false discovery rate (FDR) q-value were calculated from statistical tests.

Top Analysis-Ready Molecules for Striped Bass Growth Comparison

Superior (up-regulated)					Superior (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	ELAVL4	1.5927	0.0086	0.2300	1	-	-	-	-
2	RAP1GAP	1.2534	0.1100	0.6500	2	-	-	-	-
3	FGF13	0.9457	0.0000	0.0080	3	-	-	-	-
4	SCUBE2	0.9290	0.0003	0.0300	4	-	-	-	-
5	PRSS53	0.8372	0.0005	0.0400	5	-	-	-	-
6	EHBP1L1	0.8010	0.0200	0.3800	6	-	-	-	-
7	OLFML2B	0.7487	0.0000	0.0022	7	-	-	-	-
8	SLC38A10	0.7345	0.0072	0.2200	8	-	-	-	-
9	UBA5	0.6545	0.0300	0.4300	9	-	-	-	-
10	MACROH2A1	0.5970	0.0010	0.0700	10	-	-	-	-
Inferior (up-regulated)					Inferior (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	PCLO	2.0271	0.0012	0.0800	1	-	-	-	-
2	ZNF740	1.4055	0.0200	0.3800	2	-	-	-	-
3	PVALB	1.0782	0.0076	0.2200	3	-	-	-	-
4	MAP3K20	1.0096	0.0000	0.0009	4	-	-	-	-
5	SAT2	0.9089	0.0000	0.0009	5	-	-	-	-
6	HES1	0.8293	0.0000	0.0092	6	-	-	-	-
7	FBXL21P	0.8232	0.0300	0.4200	7	-	-	-	-
8	PHETA1	0.7469	0.0001	0.0200	8	-	-	-	-

9	CISH	0.7305	0.0000	0.0004	9	-	-	-	-
10	DET1	0.7147	0.2000	0.7500	10	-	-	-	-

Top Analysis-Ready Molecules for Striped Bass Market Size Comparison

Under Size (up-regulated)					Under Size (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	MYLK4	7.0022	0.0001	0.0056	1	MICAL1	-7.5798	0.0000	0.0015
2	GSTCD	5.7406	0.0001	0.0071	2	IL1R2	-4.2457	0.0025	0.1000
3	CACNB4	2.9772	0.0000	0.0004	3	RNU1-1	-3.4437	0.0002	0.0100
4	GPRC5B	2.7384	0.0029	0.1100	4	SHTN1	-3.2087	0.0041	0.1300
5	PVALB	2.2495	0.0001	0.0094	5	ARHGEF2	-3.0219	0.0067	0.1800
6	PCLO	2.1232	0.0100	0.2400	6	OPN5	-2.8722	0.0052	0.1500
7	CGNL1	2.0537	0.1000	0.7500	7	LEAP2	-2.8193	0.0000	0.0000
8	PFKFB4	1.7725	0.0000	0.0002	8	RABEPK	-1.9909	0.1200	0.8100
9	CCDC191	1.6455	0.0003	0.0200	9	MMP19	-1.9165	0.0000	0.0004
10	HES1	1.4304	0.0001	0.0088	10	MRPL28	-1.8569	0.0010	0.0500

Inferior Other (up-regulated)					Inferior Other (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	UBR5	2.5872	0.0074	0.1900	1	MARVELD3	-2.8070	0.0300	0.4100
2	MS4A15	2.0403	0.0082	0.2000	2	BRSK2	-2.3192	0.0003	0.0200
3	GZMA	1.9122	0.0071	0.1900	3	ELAVL4	-2.3081	0.0200	0.3400
4	CISH	1.1369	0.0000	0.0000	4	CCDC106	-2.1696	0.0053	0.1500
5	PSIP1	0.8992	0.3600	1.0000	5	CGNL1	-2.0537	0.1000	0.7500
6	FOXD1	0.6684	0.3400	1.0000	6	DNAH2	-1.4404	0.0059	0.1700
7	C17orf58	0.6246	0.0500	0.5400	7	MYO6	-0.7458	0.0048	0.1500
8	NAXE	0.6078	0.0073	0.1900	8	KDM4C	-0.5705	0.2600	1.0000
9	DUSP11	0.6011	0.2800	1.0000	9	NSMF	-0.5463	0.0400	0.5200
10	LEMD1	0.5107	0.0800	0.6800	10	PIP4P1	-0.4189	0.0100	0.2600

Superior Other (up-regulated)					Superior Other (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	IL18RAP	4.7576	0.0000	0.0015	1	UBR5	-2.5872	0.0074	0.1900
2	IL1R2	4.2457	0.0025	0.1000	2	ADNP	-2.5758	0.0100	0.2600
3	RNU1-1	3.4437	0.0002	0.0100	3	ADAMTS2	-0.8673	0.2100	0.9600
4	ARHGEF2	3.0219	0.0067	0.1800	4	CACNG6	-0.8422	0.0008	0.0400
5	ELAVL4	2.3081	0.0200	0.3400	5	FBXL21P	-0.8396	0.2000	0.9500
6	MMP19	1.9165	0.0000	0.0004	6	FOXD1	-0.6684	0.3400	1.0000
7	CCNE2	1.7232	0.0300	0.3900	7	ANTKMT	-0.6564	0.0600	0.6100
8	ZNF521	1.3120	0.2700	1.0000	8	DUSP11	-0.6011	0.2800	1.0000
9	RNF165	0.6237	0.6100	1.0000	9	CLDN19	-0.5311	0.0043	0.1400

Market Size (up-regulated)					Market Size (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
10	PARL	-0.1182	0.8700	1.0000	10	DHPS	-0.5280	0.0800	0.6800
1	MICAL1	7.5798	0.0000	0.0015	1	MYLK4	-7.0022	0.0001	0.0056
2	SHTN1	3.2087	0.0041	0.1300	2	GSTCD	-5.7406	0.0001	0.0071
3	OPN5	2.8722	0.0052	0.1500	3	IL18RAP	-4.7576	0.0000	0.0015
4	LEAP2	2.8193	0.0000	0.0000	4	CACNB4	-2.9772	0.0000	0.0004
5	MARVELD3	2.8070	0.0300	0.4100	5	GPRC5B	-2.7384	0.0029	0.1100
6	ADNP	2.5758	0.0100	0.2600	6	PVALB	-2.2495	0.0001	0.0094
7	BRSK2	2.3192	0.0003	0.0200	7	PCLO	-2.1232	0.0100	0.2400
8	CCDC106	2.1696	0.0053	0.1500	8	MS4A15	-2.0403	0.0082	0.2000
9	RABEPK	1.9909	0.1200	0.8100	9	GZMA	-1.9122	0.0071	0.1900
10	MRPL28	1.8569	0.0010	0.0500	10	PFKFB4	-1.7725	0.0000	0.0002

Top Analysis-Ready Molecules for Striped Bass Dam Comparison

Dam A (up-regulated)					Dam A (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	LINGO3	3.5999	0.0000	0.0017	1	-	-	-	-
2	CEP19	3.2447	0.0002	0.0700	2	-	-	-	-
3	RHOT1	2.7817	0.0018	0.2200	3	-	-	-	-
4	NUDT19	2.7189	0.0000	0.0019	4	-	-	-	-
5	F9	2.6437	0.0000	0.0024	5	-	-	-	-
6	PEMT	2.4159	0.0000	0.0005	6	-	-	-	-
7	OSBPL7	2.1175	0.0003	0.0800	7	-	-	-	-
8	CRTC1	2.0225	0.0023	0.2400	8	-	-	-	-
9	MOG	2.0199	0.0006	0.1200	9	-	-	-	-
10	LRRC31	1.6454	0.0000	0.0000	10	-	-	-	-
Dam B (up-regulated)					Dam B (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	CRTC1	4.6718	0.0000	0.0000	1	-	-	-	-
2	KIT	4.4424	0.0000	0.0022	2	-	-	-	-
3	SYN2	4.3820	0.0400	0.9200	3	-	-	-	-
4	SNORA3B	4.3736	0.0200	0.6900	4	-	-	-	-
5	ADGRB1	4.0108	0.0000	0.0009	5	-	-	-	-
6	SNAP25	2.7956	0.0000	0.0061	6	-	-	-	-
7	OBSN	2.6080	0.0020	0.2200	7	-	-	-	-
8	FAM111A	2.5532	0.0000	0.0100	8	-	-	-	-
9	ADM	2.4812	0.0002	0.0500	9	-	-	-	-
10	WBP4	2.3912	0.0021	0.2300	10	-	-	-	-

Top Analysis-Ready Molecules for Striped Bass Sire Comparison

Sire (up-regulated)					Sire (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	PLPP7	6230.0800	0.0000	0.0000	1	NUDT16	-1920.2700	0.0000	0.0000
2	PUM3	2081.6600	0.0000	0.0000	2	LRRC31	-1298.3300	0.0000	0.0000
3	MRPS24	1132.2900	0.0000	0.0000	3	FAM111A	-191.9700	0.0000	0.0000
4	LTBP3	844.7600	0.1400	0.7600	4	DENND2D	-127.6200	0.0005	0.0200
5	SPIN1	788.2000	0.0000	0.0000	5	CNTNAP5	-21.6500	0.0001	0.0051
6	SMARCAL1	460.4400	0.0000	0.0000	6	CIPC	-17.5900	0.0500	0.4500
7	NUDT19	455.8500	0.0000	0.0000	7	MFSD4A	-7.0200	0.0000	0.0005
8	ARNT	28.7500	0.0500	0.4200	8	CMTR1	-6.7600	0.0055	0.1000
9	CALML6	18.6500	0.0000	0.0013	9	XRCC4	-3.8400	0.0002	0.0099
10	TRHR	11.5700	0.0043	0.0900	10	METTL27	-3.3500	0.0000	0.0015

Top Analysis Ready Molecules for Striped Bass Sire Size Comparison

Large Sires (up-regulated)					Large Sires (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	MRPS24	3.9354	0.0000	0.0000	1	-	-	-	-
2	CNTNAP5	3.0580	0.0000	0.0100	2	-	-	-	-
3	PUM3	2.9549	0.0006	0.0900	3	-	-	-	-
4	PCDHGA2	2.7146	0.0004	0.0700	4	-	-	-	-
5	ZNF740	2.2891	0.0001	0.0300	5	-	-	-	-
6	CCDC106	1.9296	0.0002	0.0400	6	-	-	-	-
7	TRPM1	1.7496	0.0015	0.1600	7	-	-	-	-
8	FGFR3	1.4937	0.0028	0.2400	8	-	-	-	-
9	AQP9	1.3696	0.0017	0.1800	9	-	-	-	-
10	ARHGAP12	1.3036	0.0300	0.6200	10	-	-	-	-

Small Sires (up-regulated)					Small Sires (down-regulated)				
Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value	Order	Symbol	Log ₂ FC	P-Value	FDR Q-Value
1	CSAD	4.7550	0.0000	0.0000	1	-	-	-	-
2	CYTIP	2.6184	0.0000	0.0054	2	-	-	-	-
3	TDRD15	1.8225	0.0000	0.0200	3	-	-	-	-
4	PISD	1.1483	0.0032	0.2600	4	-	-	-	-
5	TBC1D14	1.0664	0.0003	0.0500	5	-	-	-	-
6	DHDH	0.8623	0.0007	0.1100	6	-	-	-	-
7	UCK1	0.7622	0.0026	0.2300	7	-	-	-	-
8	NIPSNAP3A	0.7388	0.0200	0.5300	8	-	-	-	-
9	RHOA	0.6303	0.0000	0.0016	9	-	-	-	-
10	XRCC4	0.5938	0.0000	0.0043	10	-	-	-	-

Appendix G

Top upstream regulators identified through Qiagen Ingenuity Pathway Analysis (Qiagen, Hilden, Germany) for comparisons of genes expressed in striped bass (SB, *Morone saxatilis*) and based on comparisons growth performance (Superior or Inferior; input genes are provided in Appendix A), growth performance to market size (Under Size, Inferior Other, Superior Other, or Market Size; input genes are provided in Appendix B), female SB parentage (Dam A or Dam B; input genes are provided in Appendix C), sire parentage (Sires 1–12; input genes are provided in Appendix D), or sire size (Large or Small; input genes are provided in Appendix E). Upstream regulators are those molecules and other elements (e.g., pathways) that are predicted to underlie the patterns of gene expression observed in the input dataset. If calculable (based on the number of associated molecules in the dataset) positive activation states are indicative of activation and negative indicative of inhibition, an activation state prediction is made if the z-score (calculated in IPA) is greater than 2.0 (“Activated”) or less than -2.0 (“Inhibited”). Upstream regulators are sorted in order of descending p-value of overlap with the most significant upstream regulators ordered as #1.

Top Upstream Regulators for Striped Bass Growth Comparison

Superior SB

Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	COL6A1	other	-	-	-	-	-	0.0000684	ENO3,TPI1
2	IKBKE	kinase	-	-	-	-	-	0.0000731	ATF4,ENO3,TPI1
3	TP53	transcription regulator	-	0.215	-	0.124	-0.105	0.0002040	ATF4,CCNG2,DNMT3A,ENO3,FGF13,MACROH2A1,PPM1A,TRIM44
4	miR-1224-3p (CCCACCU)	mature microRNA	Inhibited	-2.630	bias	-1.000	0.000	0.0003290	DNMT3A,MACROH2A1,OLFML2B,RAP1GAP,RUFY2,SLC38A10,TPI1
5	HTT	transcription regulator	-	-	-	-	-	0.0003570	ELAVL4,ENO3,RAP1GAP,RXRG,TPII
6	sodium tungstate	chemical drug	-	-	-	-	-	0.0003930	ENO3,TPI1
7	miR-3617-5p (AAGACAU)	mature microRNA	Inhibited	-2.000	bias	-1.000	0.000	0.0006220	CCNG2,ELAVL4,FGF13,PARD3
8	IL15	cytokine	-	-	-	-	-	0.0007260	ATF4,ENO3,MACROH2A1,TPI1
9	ATP5F1A	transporter	-	-	-	-	-	0.0008900	MACROH2A1
10	H2BW2	other	-	-	-	-	-	0.0008900	DNMT3A

Appendix G (continued).

Inferior SB

Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	PHF12	transcription regulator	Inhibited	-2.646	bias	-0.939	-0.160	0.0000001	BRIX1,PES1,POP4,RRP9,RRS1,TBL3,TSR1
2	1,2-dithiol-3-thione	chemical reagent	Activated	3.704	bias	0.742	0.957	0.0000001	HAX1,HPRT1,HSP90AA1,PSMA6,PSMB5,PSMC1,PSMC2,PSMD13,PSMD14,RARS1,RRS1,SERINC3,STIP1,USP14
3	ibrutinib	chemical drug	Inhibited	-3.317	bias	-0.727	-0.907	0.0000005	AIMP2,ASNS,BYSL,GNL3,HES1,IPO4,NPM1,PES1,RRP12,RRP9,SRM
4	cystemustine	chemical drug	Inhibited	-2.000	bias	-1.000	0.000	0.0000017	PSMA6,PSMB5,PSMC1,PSMC2
5	CD3	complex	Activated	4.359	bias	0.735	1.157	0.0000093	CALM1 (includes others),CISH,CYC1,CYCS,DUSP11,FBL,HAX1,HSP90AA1,NCL,NPM1,PABPC4,PCYT2,POLR2F,PREP,PSMA6,PSMC1,PUM3,RARS1,RRAGA,SLC26A5,TTF1
6	BCR (complex)	complex	Activated	3.742	bias	0.848	0.568	0.0000098	AIMP2,ASNS,BYSL,CYC1,GNL3,GRWD1,IPO4,NPM1,PES1,PPRC1,PRMT3,RRP12,RRP9,SRM
7	HNF4A	transcription regulator		1.000		0.153	0.694	0.0000100	ABCE1,ACY1,ADSS2,AHSA1,ASNS,BOLA1,BUD23,CCDC25,CCT8,CHCHD2,DDX27,DNAJC30,DUSP11,GNL3,GPT,GRWD1,HNRNPA0,NPM1,PFKFB4,PNPO,POLR1B,POLR1G,PPP6C,PRMT7,PSMB5,PSMC4,PSME3,WP1,RAB7A,RANGAP1,RRP8,RUVBL2,SAT2,SCYL3,SDHAF3,SIX2,SPATA5L1,ST13,STOML2,STRAP,TMBIM6,TRAPPC6A,TRMT6,TXNL1,UMPS,UTP23,WDR46
8	CD40	transmembrane receptor	Activated	3.742	bias	0.796	0.762	0.0000166	AIMP2,ASNS,BYSL,CYC1,GNL3,GRWD1,IPO4,NPM1,PES1,PPRC1,PRMT3,RRP12,RRP9,SRM

Appendix G (continued).

9	MYC	transcription regulator	Activated	4.503	bias	0.414	2.450	0.0000252	ABCE1,AIMP2,ASNS,C1QBP,CYCS,DDX27,DDX54,DKC1,EGLN1,EIF2B3,EIF4EBP1,EIF5A,FBL,GPT,HES1,HSP90AA1,NCL,NOP58,NPM1,PDHA1,POLR1B,POLR3D,PPID,PREP,PUS7,RARS1,RPL7L1,RPS27,RRS1,RUVBL2,SERINC3,SRM
10	CST5	other	Inhibited	-2.887	bias	-0.486	-1.205	0.0000630	BRIX1,DDX54,DKC1,GNL3,MAK16,NCL,NHP2,RPF2,RRS1,RSL1D1,TBL3,WDR3

Top Upstream Regulators for Striped Bass Market Size Comparison

Under Size SB

Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	1,2-dithiol-3-thione	chemical reagent	Activated	2.687	bias	0.508	0.809	1.050E-08	GHR,HPRT1,HSP90AA1,PSMA2,PSMA6,PSMB5,PSMC1,PSMC2,PSMC6,PSMD12,RRS1,SERINC3,TPI1,USP14
2	ibrutinib	chemical drug	Inhibited	-2.111	bias	-0.498	-0.458	9.360E-08	AIMP2,ASNS,GNL3,GPI,HES1,IPO4,PES1,RRP12,RRP9,SRM,TPI1
3	cystemustine	chemical drug	Inhibited	-2	bias	-0.686	-0.629	8.730E-07	PSMA6,PSMB5,PSMC1,PSMC2
4	bortezomib	chemical drug	Activated	2.897	-	0.079	2.625	1.230E-06	ASF1B,BNIP3L,GHR,HSP90AA1,PSMA2,PSMA6,PSMB5,PSMC1,PSMC4,PSMC6,PSMD12,PSMD6,TRIB3
5	BCR (complex)	complex	-	1.941	bias	0.582	-0.155	1.330E-06	AIMP2,ASNS,GNL3,GPI,GRWD1,IL1R2,IPO4,PES1,PRMT3,RABEPK,RRP12,RRP9,SRM,TPI1
6	CST5	other	Inhibited	-3.051	bias	-0.333	-1.851	2.090E-06	ADRM1,ARHGEF2,BRIX1,DDX54,DKC1,GNL3,MAK16,NCL,PPAN,RPF2,RRS1,RSL1D1,WDR3
7	5-fluorouracil	chemical drug	-	0		-0.124	0.466	2.190E-06	AEN,AHSA1,BNIP3L,CYCS,FBL,GAP43,GARS1,NCL,PKM,PSMA6,PSMB5,RRP8,SRM,TPI1,TRIB3

Appendix G (continued).

8	miR-4503 (UUAAGCA)	mature microRNA	-	-1.414	bias	-0.686	1.494	2.580E-06	CA1,CCNE2,CHAC2,FAM98A,FGF13,GHR,MME,MSRB3,MTDH,NOP58,PACIN2,PPID,RABEPK,RAI14,RNF11,RNF180,RRAGA,SLC47A1
9	IL2	cytokine	-	1	bias	0.387	-0.162	3.810E-06	ARHGEF2,ASNS,CCNE2,CTPS1,CYCS,DHODH,DKC1,EEF1E1,FGF13,FGF2,GARS1,GNL3,IL1R2,MME,MSRB3,PDCD11,PPP6C,SRM,TPI1,TRIB3,UMPS,WDR3
10	CD40	transmembrane receptor	-	1.941	bias	0.546	-0.027	1.180E-05	AIMP2,ASNS,GNL3,GPI,GRWD1,IP04,PES1,PRMT3,RABEPK,RRP12,RRP9,SRM,TPI1

Inferior Other SB

Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	miR-584-3p (CAGUUC)	mature microRNA	-	-0.816	bias	-0.273	-0.148	0.0001400	CISH,KDM4C,PIP4P1,PSIP1,RPA3,WDR82
2	CSF3R	transmembrane receptor	-	-	-	-	-	0.0002920	CISH,GFI1
3	miR-6768-5p (ACACAGG)	mature microRNA	-	-1.633	bias	-0.273	-0.965	0.0003030	FOXD1,GPR156,LEMD1,PIP4P1,UBR5,WDR82
4	miR-6861-3p (GGACCUC)	mature microRNA	-	-1.342	bias	-0.273	-0.732	0.0008850	FYCO1,GFI1,GPR156,PTCD3,RPP25L
5	miR-4747-5p (GGGAAGG)	mature microRNA	Inhibited	-2.333	bias	-0.273	-1.515	0.0011800	ELAVL4,GFI1,HNRNPA0,MS4A15,NAXE,RPP25L,SLC38A7,TRIM27,WDR82
6	DLAT	enzyme	-	-	-	-	-	0.0013400	CISH
7	miR-138-2-3p (CUAUUUC)	mature microRNA	-	-	-	-	-	0.0016000	GKAP1,PSIP1,TMEM50B
8	miR-3690 (CCUGGAC)	mature microRNA	-	-1.342	bias	-0.273	-0.732	0.0017100	FOXD1,FYCO1,GKAP1,GPR156,RPP25L
9	miR-1199-3p (GCGGCCG)	mature microRNA	-	-0.555	bias	-0.273	0.05	0.0021300	CCDC106,CISH,GFI1,MS4A15,NSMF
10	ARIH2	enzyme	-	-	-	-	-	0.0026700	GFI1

Appendix G (continued)

Superior Other SB

Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	IL2	cytokine	-	-0.447	bias	-0.349	0.332	0.0000987	ARHGEF2,CCNE2,CYCS,FGF2,GFI1,IL18RAP,IL1R2,PPP6C
2	guanidinopropionic acid	chemical drug	-	-	-	-	-	0.000112	ATPAF2,CYCS,ESRRB
3	miR-16-5p (AGCAGCA)	mature microRNA	Activated	2.413	bias	0.617	0.06	0.000115	AADAT,ASF1B,CCNE2,FGF2,INPP5E,KBTBD12,MMP19,PARL,PPP6C,PTCD3,RNF165,SDHAF2,SELENOO,SERBP1,SLC38A7
4	miR-3083-5p (GGCUGGG)	mature microRNA	Activated	2.292	bias	0.617	-0.308	0.00018	ADAMTS2,ANGEL1,ARHGEF2,ASF1B,BTD,CLDN19,CYCS,DHPS,ELAVL4,ESRRB,GPR156,INPP5E,KBTBD12,MMP19,NUCB1,RNF165,SLC38A7,UBR5
5	miR-5579-5p (AUGGUAC)	mature microRNA	-	1.633	bias	0.617	0.122	0.000221	ASF1B,CDH19,CYCS,GFI1,SERBP1,ZNF521
6	IL25	cytokine	-	-	-	-	-	0.000248	FGF2,GFI1,IL1R2
7	miR-760-5p (CCCUCAG)	mature microRNA	-	1.414	bias	0.617	-0.331	0.000329	ATPAF2,CACNG6,ELAVL4,GPT,KBTBD12,RNF11,RNF165,UBR5
8	miR-6824-5p (UAGGGGA)	mature microRNA	-	0.378	bias	0.617	-1.255	0.000342	ASF1B,DHPS,FGF2,IL18RAP,MMP19,NUCB1,RNF165
9	miR-4755-3p (GCCAGGC)	mature microRNA	Activated	2.111	bias	0.617	0.064	0.000503	ANGEL1,ANTKMT,ATPAF2,CYCS,GPT,MAGT1,MMP19,NUCB1,RNF165,SDHAF2,SLC38A7
10	betaine	chemical - endogenous mammalian	-	-	-	-	-	0.000658	ESRRB,FGF2

Appendix G (continued).

Market Size SB

Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	ibrutinib	chemical drug	Activated	2.309	bias	0.452	0.745	6.40E-09	AIMP2,ASNS,GNL3,GPI,GZMA,HES1,IPO4,PES1,RRP12,RRP9,SRM,TPI1
2	1,2-dithiol-3-thione	chemical reagent	Inhibited	-2.687	bias	-0.461	-0.984	7.05E-09	GHR,HPRT1,HSP90AA1,PSMA2,PSMA6,PSMB5,PSMC1,PSMC2,PSMC6,PSMD12,RRS1,SERINC3,TPI1,USP14
3	cystemustine	chemical drug	Activated	2	bias	0.622	0.757	7.71E-07	PSMA6,PSMB5,PSMC1,PSMC2
4	bortezomib	chemical drug	Inhibited	-2.354		-0.072	-2.107	8.67E-07	BNIP3L,GHR,HSP90AA1,PSMA2,PSMA6,PSMB5,PSMC1,PSMC4,PSMC6,PSMD12,PSMD6,RPA3,TRIB3
5	CST5	other	Activated	2.496	bias	0.302	1.408	1.48E-06	ADRM1,BRIX1,DDX54,DKC1,GNL3,MAK16,NCL,PPAN,RPF2,RRS1,RSL1D1,VDAC2,WDR3
6	5-fluorouracil	chemical drug	-	0		0.113	-0.422	1.49E-06	AEN,AHSA1,BNIP3L,CYCS,FBL,GAP43,GARS1,NCL,PKM,PSMA6,PSMB5,RRP8,SRM,TPI1,TRIB3
7	BCR (complex)	complex	-	-1.941	bias	-0.527	-0.04	5.09E-06	AIMP2,ASNS,GNL3,GPI,GRWD1,IPO4,PES1,PRMT3,RABEPK,RRP12,RRP9,SRM,TPI1
8	CD40	transmembrane receptor	-	-1.941	bias	-0.495	-0.157	8.47E-06	AIMP2,ASNS,GNL3,GPI,GRWD1,IPO4,PES1,PRMT3,RABEPK,RRP12,RRP9,SRM,TPI1
9	PHF12	transcription regulator	Activated	2.236	bias	0.584	0.93	1.08E-05	BRIX1,PES1,RRP9,RRS1,TSR1
10	IL2	cytokine	-	-1.784	bias	-0.351	-0.764	2.66E-05	ASNS,CISH,CTPS1,CYCS,DHODH,DKC1,EEF1E1,FGF13,GARS1,GNL3,GZMA,IL18RAP,MME,MSRB3,PDCD11,SRM,TPI1,TRIB3,UMPS,WDR3

Appendix G (continued).

Top Upstream Regulators for Striped Bass Dam Comparison

Dam A									
Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	miR-193a-3p (ACUGGCC)	mature microRNA	Inhibited	-3.317	bias	-1	0	0.000186	ACSF3,ARFIP1,BHLHE41,CRTC1,CST3,F9,MTDH,NUDT19,NUP160,NUP42,TP53INP1
2	miR-3569-3p (CAGUCUG)	mature microRNA	Inhibited	-2.985	bias	-1	0	0.000494	ADK,ARFIP1,ATL2,CST3,MAB21L2,MASTL,MOG,NUDT19,TAF11
3	miR-6481 (ACUGAAA)	mature microRNA	Inhibited	-2.236	bias	-1	0	0.000718	ADK,ATL2,ETFBKMT,NUP42,TP53INP1
4	miR-7019-3p (CACCUUG)	mature microRNA	Inhibited	-2.236	bias	-1	0	0.000831	CEP19,ETFBKMT,MPC2,MSRA,RHOT1
5	miR-5690 (CAGCUAC)	mature microRNA	Inhibited	-2.236	bias	-1	0	0.000937	CST3,ETFBKMT,F9,NUP160,RHOT1
6	miR-802-5p (CAGU AAC)	mature microRNA	Inhibited	-2.621	bias	-1	0	0.001010	F9,MBNL2,NUDT19,RHOA,STAT1,TAF11,TRMT10A
7	miR-4327 (GCUUGCA)	mature microRNA	Inhibited	-2.236	bias	-1	0	0.001020	CEP19,NUP42,RHOA,SLC37A3,TP53INP1
8	colistin	biologic drug	-	-	-	-	-	0.001040	CRYL1,CST3,RHOA
9	miR-493-5p (UGUACAU)	mature microRNA	Inhibited	-2.621	bias	-1	0	0.001080	ARFIP1,ATL2,CEP19,MBNL2,RHOT1,SLC37A3,TP53INP1
10	kanamycin A	chemical drug	-	-	-	-	-	0.001090	CRYL1,CST3,RHOA
Dam B									
Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	miR-4425 (GUUGGGA)	mature microRNA	Inhibited	-3.742	bias	-1	0	0.0000641	ARHGAP9,CDIN1,COIL,DNAAF6,EIF5A,HAPLN1,IFT140,MAT2B,PUM2,RWDD2B,SH2D1B,SNAP25,TPMT,TPTE
2	NFkB (complex)	complex	-	1.293	bias	0.856	-0.762	0.0001570	ADM,BAMBI,CAPNS1,CYBB,FPR2,KIT,MYBL1,NLRP12,TPMT,VIM
3	NEDD9	other	-	1.951	-	-0.159	2.261	0.0003450	ADM,AURKA,KIT,VIM

Appendix G (continued).

4	miR-298 (GCAGAAG)	mature microRNA	Inhibited	-3.317	bias	-1	0	0.0005380	CHRFAM7A,COPZ2,CYBB,EP515,N AP1L1,SH2D1B,SLC47A1,TMEM167 A,TPTE,VIM,ZBTB25
5	MYB	transcription regulator	-	1.999	bias	0.388	1.224	0.0005720	ADM,AURKA,CLTA,KIT,VIM
6	miR-5119 (AUCUCAU)	mature microRNA	Inhibited	-2.449	bias	-1	0	0.0006940	ARFIP2,CDIN1,FAM111A,P3H4,SNA P25,TMEM167A
7	DAB2IP	other	-	-	-	-	-	0.0007380	KIT,VIM
8	TIMP3	other	-	-	-	-	-	0.0009140	BAMBI,KIT,VIM
9	miR-4464 (AGGUUUG)	mature microRNA	Inhibited	-2.828	bias	-1	0	0.0009780	CDIN1,CHRFAM7A,CLTA,EIF5A,IR AK3,MAT2B,SLC9A3R2,ZNF28
10	OPRM1	g-protein coupled receptor	-	-	-	-	-	0.0010400	ADM,VIM

Top Upstream Regulators for Striped Bass Sire Comparison

Sire									
Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias- corrected z-score	p-value of overlap	Target molecules in dataset
1	miR-3914 (AGGAACC)	mature microRNA	-	-0.447	bias	-0.273	0.163	0.001030	CIPC,LRRC31,NUDT19,PLPP7,SPIN 1
2	gentamicin	chemical drug	-	1	-	0.051	0.897	0.001220	CIPC,PELO,RHOA,SQSTM1
3	ARHGAP26	other	-	-	-	-	-	0.001340	RHOA
4	CCM2	other	-	-	-	-	-	0.001340	RHOA
5	DAAM1	other	-	-	-	-	-	0.001340	RHOA
6	PURPL	other	-	-	-	-	-	0.001340	SQSTM1
7	RHPN2	other	-	-	-	-	-	0.001340	RHOA
8	RhoGap	group	-	-	-	-	-	0.001340	RHOA
9	TNS1	other	-	-	-	-	-	0.001340	RHOA
10	TPM2	other	-	-	-	-	-	0.001340	RHOA

Appendix G (continued).

Top Upstream Regulators for Striped Bass Sire Size Comparison

Large Sire									
Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	miR-344d-2-5p (GUCUGGU)	mature microRNA	Inhibited	-2.828	bias	-1	0	0.0000779	ABHD14B,CBFA2T2,CRYBA1,EBF2,FAM107B,MBNL2,SIDT2,TBC1D25
2	miR-203b-5p (AGUGGUC)	mature microRNA	Inhibited	-2.646	bias	-1	0	0.0000849	AQP9,CNTNAP5,LNX2,MPEG1,PPM1K,SRRM1,THAP12
3	miR-12192-5p (GUGGGGU)	mature microRNA	Inhibited	-3.317	bias	-1	0	0.0000878	C1QL2,CBFA2T2,DCTN1,ELMO1,FAM107B,KSR1,NGFR,PCDHGA2,SIDT2,TMEM198,ZNF740
4	miR-6825-5p (GGGGAGG)	mature microRNA	Inhibited	-4.338	bias	-1	0	0.0001180	ABHD14B,AQP9,CBFA2T2,CCDC106,CEACAM1,CNTNAP5,INPP4B,KSR1,MPEG1,NGFR,PCDHGA2,PIGO,PPM1K,PRPF4B,RBP5,SIDT2,TBC1D25,TMEM198,ZNF740
5	miR-149-3p (GGGAGGG)	mature microRNA	Inhibited	-3.873	bias	-1	0	0.0002100	ABHD14B,C1QL2,CBFA2T2,EBF2,FGFR3,KSR1,NGFR,PCDHGA2,PIGO,PPM1K,RBP5,SIDT2,TBC1D25,TMEM198,ZNF740
6	miR-3649 (GGGACCU)	mature microRNA	Inhibited	-2.646	bias	-1	0	0.0003900	ABHD14B,DCTN1,ELMO1,PCDHGA2,PIGO,TMEM198,ZNF740
7	miR-637 (CUGGGGG)	mature microRNA	Inhibited	-3.719	bias	-1	0	0.0006790	ABHD14B,AQP9,CCDC106,CEACAM1,DCTN1,FGFR3,LNX2,MBNL2,NGFR,PCDHGA2,RBP5,SIDT2,TBC1D25,ZNF740
8	miR-1298-3p (AUCUGGG)	mature microRNA	Inhibited	-2.985	bias	-1	0	0.0007140	ABHD14B,ARHGAP12,CEACAM1,CYTIP,FAM107B,MPEG1,RBL2,RBP5,TBC1D25
9	miR-3179 (GAAGGGG)	mature microRNA	Inhibited	-3	bias	-1	0	0.0007210	CEACAM1,DCTN1,EBF2,ELMO1,MPEG1,NGFR,PCDHGA2,RBP5,ZNF740
10	miR-185-3p (GGGGCUG)	mature microRNA	Inhibited	-3.838	bias	-1	0	0.0007720	CBFA2T2,CEACAM1,DCTN1,EBF2,ELMO1,KSR1,LNX2,MAPK14,MPEG1,NGFR,PCDHGA2,PIGO,RBP5,SIDT2,TMEM198

Appendix G (continued).

Small Sires									
Order	Upstream Regulator	Molecule Type	Predicted State	Activation z-score	Notes	Bias Term	Bias-corrected z-score	p-value of overlap	Target molecules in dataset
1	miR-4519 (AGCAGUG)	mature microRNA	Inhibited	-2.646	bias	-1	0	0.0000317	GYG1,NIPSNAP3A,PSMG2,PUSL1,RHOA,SKP1,UCK1
2	miR-1927 (ACCUCUG)	mature microRNA	Inhibited	-2	bias	-1	0	0.0003320	CYTIP,MRPS24,PSMG2,TBC1D14
3	let-7c-1-3p (UGUACAA)	mature microRNA	Inhibited	-2.63	bias	-1	0	0.0003560	FABP3,NXPH1,PDSS2,PISD,PSMG2,PUSL1,XRCC4
4	miR-344d-3p (AUAUAAC)	mature microRNA	-	-1.982	bias	-1	0	0.0003780	ATG16L1,MRPL9,PSMG2,SKP1
5	miR-8066 (AAUGUGA)	mature microRNA	Inhibited	-2.236	bias	-1	0	0.0004220	NIPSNAP3A,PCBD2,TDRD15,UCK1,XRCC4
6	miR-6844 (UCUUUGU)	mature microRNA	Inhibited	-2.236	bias	-1	0	0.0005280	CSAD,CYTIP,PSMG2,SC5D,XRCC4
7	miR-154-3p (AUCAUAC)	mature microRNA	Inhibited	-2.213	bias	-1	0	0.0008470	PCBD2,PSMG2,RHOA,SC5D,SKP1
8	miR-467e-5p (UAAGUGU)	mature microRNA	Inhibited	-2	bias	-1	0	0.0009440	GYG1,L3HYPDH,NIPSNAP3A,NXP H1
9	miR-6974-3p (CUCCACU)	mature microRNA	Inhibited	-2.236	bias	-1	0	0.0011600	CYTIP,MAK16,PDSS2,PUSL1,RAB9 A
10	ARHGAP26	other	-	-	-	-	-	0.0012900	RHOA

Appendix H

Top causal networks identified through Qiagen Ingenuity Pathway Analysis (Qiagen, Hilden, Germany) for comparisons of genes expressed in striped bass (SB, *Morone saxatilis*) and based on comparisons growth performance (Superior or Inferior; input genes are provided in Appendix A), growth performance to market size (Under Size, Inferior Other, Superior Other, or Market Size; input genes are provided in Appendix B), female SB parentage (Dam A or Dam B; input genes are provided in Appendix C), sire parentage (Sires 1–12; input genes are provided in Appendix D), or sire size (Large or Small; input genes are provided in Appendix E). Causal networks explore the causal relationships associated with the experimental data and upstream regulators thereof (Table 3.13) to other master regulators. If calculable (based on the number of associated molecules in the dataset) positive activation states are indicative of activation and negative indicative of inhibition, an activation state prediction is made if the z-score (calculated in IPA) is greater than 2.0 (“Activated”) or less than -2.0 (“Inhibited”). Upstream regulators are sorted in order of descending p-value of overlap with the most significant upstream regulators ordered as #1.

Top Causal Networks for Striped Bass Growth Comparison

Superior SB									
Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	USP15	peptidase	-	-	0.333	0.0000236	0.0005	CTNNB1,RORC,TP53,USP15	ATF4,CCNG2,DNMT3A,ENO3,FGF13,MACROH2A1,PAR3,PPM1A,TPI1
2	NEK10	kinase	-	-	0.707	0.0000457	0.0003	MAP2K1,NEK10,TP53	ATF4,CCNG2,DNMT3A,ENO3,FGF13,MACROH2A1,PPM1A,RAP1GAP
3	BUB1B	kinase	-	-	0	0.0000467	0.0003	APC,BUB1B,TP53	ATF4,CCNG2,DNMT3A,ENO3,FGF13,MACROH2A1,PAR3,PPM1A
4	RFFL	enzyme	-	-	0	0.0000482	0.0006	Pkc(s),RFFL,TP53	ATF4,CCNG2,DNMT3A,ELAVL4,ENO3,FGF13,MACROH2A1,PPM1A
5	PSEN2	peptidase	-	-	0	0.0000596	0.0005	AGT,AKT1,CTNNB1,EGFR,ERK1/2,ESR1,EZH2,Jnk,MAP2K1,Pkc(s),PLA2G6,PSEN2,RXRA,TNF,TP53	ATF4,CCNG2,DNMT3A,ELAVL4,ENO3,FGF13,MACROH2A1,OLFML2B,PAR3,PPM1A,PRSS53,RAP1GAP,RXRG,SCUBE2
6	IKBKE	kinase	-	bias	1.732	0.0000597	0.0004	IKBKE	ATF4,ENO3,TPI1

Appendix H (continued).

7	Jnk dimer	complex	-	-	-0.277	0.0000676	0.0003	CTNNB1,ESR1,Jnk ,Jnk dimer,JUN,MAP2K 1,MAP2K4,MAPK8 ,MAPK9,Pkc(s),RX RA,TNF,TP53	ATF4,CCNG2,DNMT3A, ELAVL4,ENO3,FGF13,M ACROH2A1,OLFML2B,P ARD3,PPM1A,RAP1GAP ,RXRG,SCUBE2
8	COL6A1	other	-	bias	1.414	0.0000684	0.0004	COL6A1	ENO3,TPI1
9	PHF1	transcription regulator	-	-	0	0.0000693	0.0005	EZH2,PHF1,TP53	ATF4,CCNG2,DNMT3A, ENO3,FGF13,MACROH2 A1,PPM1A,RAP1GAP
10	MELK	kinase	-	-	0	0.0000704	0.0005	EZH2,MELK,TP53	ATF4,CCNG2,DNMT3A, ENO3,FGF13,MACROH2 A1,PPM1A,RAP1GAP

Inferior SB

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	PHF12	transcription regulator	Inhibited	bias	-2.646	6.06E-08	0.0001	PHF12	BRIX1,PES1,POP4,RRP9, RRS1,TBL3,TSR1
2	1,2- dithiol-3- thione	chemical reagent	Activated	bias	3.742	9.05E-08	0.0001	1,2-dithiol-3-thione	HAX1,HPRT1,HSP90AA1 ,PSMA6,PSMB5,PSMC1, PSMC2,PSMD13,PSMD1 4,RARS1,RRS1,SERINC3 ,STIP1,USP14
3	BCR (complex)	complex	Activated	bias	3.742	0.0000001 91	0.0001	BCR (complex)	AIMP2,ASNS,BYSL,CYC 1,GNL3,GRWD1,IPO4,NP M1,PES1,PPRC1,PRMT3, RRP12,RRP9,SRM

Appendix H (continued).

5	dioleoylphosphatidic acid	chemical - endogenous mammalian	Activated	-	2.5	0.000000326	0.0001	Akt,AKT1,CDKN2A,Creb,dioleoylphosphatidic acid,EIF3A,FMR1,HIF1A,MAPT,MTOR,MYC,p70S6k,PDX1,RICTOR,RPS6KB1,TP53	ABCE1,ACOT11,AEN,ASF1B,ASL,ASNS,ASPG,C1QBP,CISH,CPQ,CTH,CYC1,CYCS,DDX27,DDX54,EEF1A1,EGLN1,EIF4EBP1,EIF5A,FBL,FGF2,FKTN,GNL3,GPT,MAPRE3,MRPL1,MRPL13,NCL,NOP58,NPEPL1,NPM1,PABPC4,PCLO,PNPO,POLR3D,PPID,PPRC1,PREP,PSMA6,PSMA8,PSMB5,PSMC1,PSMC2,PSMC4,PSMD13,PSMD14,PSMD6,PSME3,PUM3,PUS7,QDPR,ARS1,RPA3,RPL7L1,RPS27,RRS1,RUVBL2,SERINC3,SRM,ST13,STIP1,TXNL1,UMPS,USP14
6	ibrutinib	chemical drug	Inhibited	bias	-3.317	0.000000423	0.0001	ibrutinib	AIMP2,ASNS,BYSL,GNL3,HES1,IPO4,NPM1,PES1,RRP12,RRP9,SRM

Appendix H (continued).

7	PP-121	chemical drug	-	bias	-1.763	0.0000006	0.0001	15	ABL1,Akt,Ap1,CDKN1B,CRTC2,CTN NB1,DNA-PK,EGFR,EIF4E,ERBB2,ERK,ESR1,estrogen receptor,ETS1,FOS,GNAS,HCK,HIF1A,HNRNPK,HRAS,HSF1,JUN,KDR,KRAS,MAPK1,MAPK9,MAPT,MKMK1,MTOR,MYC,MYD88,NEDD9,NFKBIA,NOS2,Pdgfr,PDX1,PIK3CA,PIK3CG,PKD1,PP121,PRDM1,RAS,RB1,RICTOR,SGK1,SRC,STAT3,STAT5B,STK4,TLR2,TLR3,TP53,TP63,TP73,Vegf	ABCE1,ACOT11,ACY1,AN, AHS1,AIMP2,ANGEL1,APOO,ASL,ASNS,ASPG,BLMH,BYSL,C1QP,CALM1 (includes others),CCPG1,CCT8,CDH19,CYC1,CYCS,DBND1,DDX27,DDX54,DHPS,DKC1,DUSP11,DYNLRB2,EIF5A,ELOB,ENPP6,FKTN,FOXD1,G3BP1,GFOD1,GNL3,GPT,GTFA,HAX1,HERC4,HNRNPA0,MAP3K20,MRPL1,MRPL13,MRPL9,MSH3,NCL,NHP2,NOP58,NPM1,PAIP1,PCLO,PCYT2,PES1,PNPO,POLR1B,POLR1G,POLR3B,POLR3D,PPI D,PREP,PSMA6,PSMA8,PSMB5,PSMC1,PSMC2,PSMC4,PSMD13,PSMD14,PSMD6,PSME3,PUM3,PUS7,PWP1,QDPR,RANGAP1,RARS1,RPL7L1,RPS27,RRS1,RUVBL2,SERINC1,SERINC3,SLC38A3,SLC47A1,TK2,TMBIM6,TXNL1,UBE3A,UMPS,USP14,YBX2,ZPR1
---	--------	---------------	---	------	--------	-----------	--------	----	--	---

Appendix H (continued).

8	sirolimus	chemical drug	-	-	-0.381	0.000000978	0.0001	AIFM1,Akt,AKT1, AKT2,BCL2,BIRC 5,Ca2+,Calcineurin protein(s),Calmodulin,CCND1,Cdc42,Cdk,CDK19,CDK2,CDK4,Creb,CXCL12, DDX5,DNMT3A,d oxorubicin,EIF4E,ERBB2,ERK,ETS1,EYA1,FADD,FMR1, FOXA2,GATA4,Gsk3,Hdac,HIF1A,HS D17B12,HSF1,IGF1,IL2,Jnk,JUN,KRAS,MAF,MAP3K14, MAPK1,MAPT,Me k,mir-21,MKNK1,MMP1, MMP3,MTOR,MXD1,MYC,NCF4,NE DD9,NFkB (complex),NFKBIA ,NOS3,ODC1,OSM, OTX2,P38 MAPK,p70 S6k,PI3K (complex),PI3K (family),PITX2,PK D1,PPARG,PPIF,PRKCD,progesterone ,PTK2,RB1,reactive oxygen species,RICTOR,sir olimus,SMARCA4, STAR,STAT3,STAT4,STAT5B,sulforan, TAL1,TLR4,TN	ABCE1,ACOT11,ADSS2, AEN,AHSA1,AIMP2,ANGEL1,APOO,APRT,ASN S,ASPG,BUD23,C1QBP,C ALM1 (includes others),CCT8,CHCHD4,C PQ,CRYGC,CTH,CTPS1, CYCS,DARS1,DDX54,D HODH,DHPS,DKC1,DYN LRB2,EEF1A1,EEF1E1,ELN1,EIF5A,ELAC2,EN PP6,EPDR1,FBL,FGF2,FKTN,FLVCR2,FOXO1, F UT9,G3BP1,GFOD1,HER C4,HNMT,HSP90AA1,KL HL7,MAK16,MAP3K20, MAPRE3,MRPL1,MRPL13,MRPL23,MRPL28,MRP L9,MSH3,MSRA,NCL,NP EPL1,NPM1,PABPC4,PAI P1,PDCD11,PDHA1,POL R1B,POLR1G,POLR3B,P PID,PPP6C,PPRC1,PREP, PSMA6,PSMA8,PSMC1,PSMC2,PSMC4,PSMD13,PSMD14,PSMD6,PSME3,PTCD3,PUM3,PUS7,QDPR ,RANGAP1,RARS1,RPL7 L1,RPS27,RRP8,SDHAF3 ,SERINC1,SERINC3,SFX N2,SIX2,SKP1,SLC38A3, SLC47A1,SRM,STK32C, STRAP,TECR,TK2,TTL12, TXNL1,UBE3A,UMPS ,USP14,WDR73,WDR82, YBX2,ZPR1
---	-----------	---------------	---	---	--------	-------------	--------	---	---

Appendix H (continued).

9	Cdk	group	Activated	-	2.082	0.0000013	0.0001	Cdk,CDK19,CDK5,CDKN2A,MYC,NRF1,RB1	ABCE1,AEN,AIMP2,ASF1B,ASNS,C1QBP,CTH,DDX27,DDX54,EEF1A1,EIF4EBP1,EIF5A,FGF2,GNL3,GPT,HES1,HSP90AA1,MRPL9,NCL,NOP58,PDHA1,POLR3D,PPID,PPRC1,PREP,PSMB5,PSMC1,PSMC4,PSMD14,PSMD6,PUS7,RARS1,RPL7L1,RPS27,RRP8,RRS1,RUVBL2,SERINC3,SRM
10	CD8A	other	Activated	bias	4.796	0.0000013	0.0001	CD3,CD8A,TCR	CISH,CYC1,CYCS,DKC1,FBL,HAX1,HSP90AA1,NCL,NHP2,NOP58,NPM1,PABPC4,PCYT2,POLR2F,PREP,PSMA6,PSMC1,PU M3,RARS1,RRAGA,RRS1,SLC26A5,TTF1

Top Causal Networks for Striped Bass Market Size Comparison

Under Size SB

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	1,2-dithiol-3-thione	chemical reagent	Activated	bias	2.673	1.05E-08	0.0001	1,2-dithiol-3-thione	GHR,HPRT1,HSP90AA1,PSMA2,PSMA6,PSMB5,PSMC1,PSMC2,PSMC6,PSMD12,RRS1,SERINC3,TP11,USP14

Appendix H (continued).

2	miR-503-5p (AGCAGCG)	mature microRNA	Inhibited	bias	-2.907	0.0000000	0.0001	Akt,BRCA1,CDK2,CDKN1B,CDX2,C EBPB,CTNNB1,E2 F1,estrogen receptor,FOXO1,ID 3,MAPT,miR-503- 5p (AGCAGCG),MYC ,MYOD1,NCOA1,P ARP1,PGR,PTTG1, RB1,RBL1,SMARC A4,SP1,STAT3,TP5 3,WT1	AEN,AGL,AHSA1,AIMP 2,APRT,ASF1B,ASNS,AS PG,ATPAF2,BNIP3L,CA1 ,CCNE2,CDH19,CENPF, CLDN23,CTPS1,CYCS,D AO,DDX54,DKC1,DPH2, DUS4L,DUSP3,EEF1A1, EIF4EBP1,FBL,FGF13,F GF2,FTSJ3,GAP43,GHR, GPI,GPT,GSTCD,HES1,H NMT,KLHL23/PHOSPHO 2- KLHL23,MAGT1,MME, MSH3,NAA15,NCL,NOP 58,NUCB1,NUDC,PCLO, PES1,PFKFB4,POLR3D,P PAN,PPID,PPM1A,PRKC Z,PSMA6,PSMB5,PSMD1 2,PSME3,PUM3,PVALB, QDPR,RAI14,RANGAP1, RNF207,RRS1,RSL1D1,R UVBL2,SDHAF2,SERIN C3,SERPINC1,SRM,ST13 ,STK32C,TPI1,TSPAN8,T XNL1,UMPS,USP14,USP 2,ZNF622,ZPR1
---	-------------------------	--------------------	-----------	------	--------	-----------	--------	--	---

Appendix H (continued).

3	GNL1	other	-	-	0.25	4.06E-08	0.0001	26s Proteasome,AIFM1, Akt,CCND1,CCNE 1,CD40,CDK4/6,C DKN1A,EGR1,EIF 2AK3,EP300,ERK1 /2,FOS,GNL1,HIF1 A,HSF1,Hsp70,MA PT,Mek,MYC,NFE 2L2,NFkB (complex),NFKBIA ,PARP1,TLR2,TP53	ADRM1,AEN,AHSA1,AR HGEF2,ASPG,ATPAF2,B LMH,CCT8,CENPF,CHC HD4,CTH,DDX27,DDX5 4,EEF1A1,EIF4EBP1,FGF 13,G3BP1,GNL3,GPI,GPT ,GRWD1,HERC4,HPRT1, IL1R2,IPO4,MAGT1,MM E,MMP19,MSH3,NOP58, NUCB1,PCLO,PCYT2,PE S1,PFKFB4,POLR3D,PPI D,PPM1A,PRKCZ,PRMT 3,PSMA2,PSMA6,PSMC1 ,PSMC2,PSMC4,PSMD12 ,PSMD6,PSME3,PUM3,P VALB,QDPR,RABEPK,R ANGAP1,RRP12,RRP9,R UVBL2,SERINC1,SERIN C3,SERPINC1,SRM,ST13 ,TRIB3,UBE3A,UMPS AIMP2,ASNS,GNL3,GPI, HES1,IPO4,PES1,RRP12, RRP9,SRM,TPI1
4	ibrutinib	chemical drug	Inhibited	bias	-2.111	7.52E-08	0.0001	ibrutinib	

Appendix H (continued).

5	dioleoylphosphatidic acid	chemical - endogenous mammalian	-	-	1.05	8.06E-08	0.0001	Akt,AKT1,CDKN2A,Creb,CREB1,dioleoylphosphatidic acid,EIF4B,FMR1,HIF1A,MAPT,MTOR,MYC,p70S6k,PDX1,RICTOR,RPS6KB1,TP53	AEN,ASF1B,ASNS,ASPG,ATPAF2,BNIP3L,CENPF,CTH,CYCS,DDX27,DDX54,EEF1A1,EIF4EBP1,FBXL,FGF13,FGF2,GAP43,GHR,GNL3,GPI,GPT,IL1R2,MAGT1,MICAL1,MME,NCL,NOP58,NPEPL1,NUCB1,PCLO,POLR3D,PPI D,PPM1A,PREP,PRKCZ,PSMA2,PSMA6,PSMB5,PSMC1,PSMC2,PSMC4,PSMC6,PSMD12,PSMD6,PSME3,PUM3,QDPR,RRS1,RUVBL2,SERINC3,SERPINC1,SHTN1,SRM,ST13,TPI1,TXNL1,UMPS,USP14
---	---------------------------	---------------------------------	---	---	------	----------	--------	---	---

Appendix H (continued).

6	perilla alcohol	chemical drug	-	-	-0.831	9.18E-08	0.0001	AHR,Akt,BRCA1,C AT,CCND1,CD44, CDK2,CDKN1B,C EBPB,CTNNB1,E2 F1,EGFR,estrogen receptor,farnesyl transferase,FOS,FO XO1,G6PD,GADD 45A,GC-GCR dimer,HRAS,HSF1, hydrogen peroxide,ID3,IDH1, MAP3K14,MAPT, MYC,MYOD1,NC OA1,NFKB1,ODC1 ,perilla alcohol,PPARG,RE LA,RXRA,SMARC A4,STAT5a/b,TNF, TNFRSF1A,TP53, WT1,XDH	AEN,AGL,AHSA1,AIMP 2,APRT,ARHGEF2,ASF1 B,ASPG,ATPAF2,BDH1, BLMH,BNIP3L,BUD23,C CNE2,CCT8,CDH19,CEN PF,CTPS1,DAO,DDX52, DDX54,DUSP3,EIF4EBP 1,FBL,FGF2,G3BP1,GHR, GPI,GPRC5B,HERC4,HN MT,HPRT1,IL1R2,KDM4 C,MAGT1,MME,MTDH, NCL,NUCB1,OLFML2B, PCYT2,PES1,PKM,POLR 1G,POLR3D,PPAN,PPID, PPM1A,PSMA2,PSMA6,P SMC4,PSMD12,PSME3,P UM3,QDPR,RAI14,RAN GAP1,RCN3,RSL1D1,RU VBL2,RXRG,SERBP1,SE RINC3,SERPINC1,SLC16 A5,SRM,TSPAN8,UMPS, USP14,USP2,WDR3
---	--------------------	------------------	---	---	--------	----------	--------	---	---

Appendix H (continued).

7	procaterol	chemical drug	-	-	0.882	0.0000001	0.0001	ADRB2,Akt,CAV1, Cdc42,Creb,doxorubicin,EGFR,ERK1/2,F7,FGFR1,HRAS, Mek,MET,MYC,MYD88,NFKBIA,NO S3,POU5F1,PPAR G,procaterol,PTEN, PTGS2,PTH,RAS,RYR2,SMAD3,TERT ,TNF,TP53	AEN,ARHGFE2,ASNS,ATPAF2,BDH1,BUD23,CENPF,CGNL1,CLDN23,CTH,CTPS1,DDX27,DDX54, DHODH,DKC1,EEF1A1, FBL,FGF13,FGF2,G3BP1, GNL3,GPI,GPRC5B,MAGT1,MME,MMP19,MSH3 ,MTDH,MYLK4,NOP58, NUCB1,NUDC,NXF1,OL FML2B,PFKFB4,POLR1 G,POLR3D,PPID,PREP,PRKCZ,PSMA2,PSMD12,PSME3,PUM3,QDPR,RABEPK,RAI14,RANGAP1,R CN3,RUVBL2,SERBP1,SERINC1,SERINC3,SERPIN1,SLC16A5,SRM,ST13 ,TPI1,TSPAN8,UBE3A,UMPS,USP14,USP2
8	BCR (complex)	complex	-	bias	1.941	0.0000001	0.0001	BCR (complex)	AIMP2,ASNS,GNL3,GPI, GRWD1,IPO4,PES1,PRMT3,RABEPK,RRP12,RRP9,SRM,TPI1

Appendix H (continued).

9	L-threonine	chemical - endogenous mammalian	-	-	0.122	0.0000001	0.0001	26s Proteasome,AIFM1, Akt,BRCA1,CDKN1B,CYP19A1,EGFR,ERK,ESR2,estrogen receptor,FOS,GATA4,HSF1,HTT,IGF1,Jnk,L-threonine,MAPT,MASTL,MTOR,MYCN,NCF4,NFKBIA,NGEF,P38MAPK,p70S6k,PI3K (complex),PPARGC1A,PRDM1,RICTOR,RPS6KB1,RPTOR,RXRA,S100A9,TAL1,TNF,Vegf	ACLY,ADRM1,AHSA1,AIMP2,ANGEL1,ASPG,BANF1,BLMH,BNIP3L,BUD23,CCT8,CDH19,CGNL1,CHCHD4,CYCS,DHPS,DKC1,DNAI1,DUSP3,EEF1A1,EEF1E1,EPDR1,FGF2,FUT9,GAP43,GPI,HERC4,HES1,MAGT1,MAK16,MICAL1,MME,MSH3,MSRB3,MTDH,NUDC,OLFML2B,PACSIN2,PCYT2,PES1,PKM,POLR1G,PPAN,PRKCZ,PSMA2,PSMA6,PSMB5,PSMC1,PSMC2,PSMC4,PSMC6,PSMD12,PSMD6,PSME3,RCN3,RNF11,RRP12,RRS1,RXRG,SERBP1,SERINC1,SLC16A5,TPI1,TXNL1,UBE2G1,UBE3A,YBX2
10	Z-LLL-CHO	chemical - protease inhibitor	Activated	-	3.046	0.0000002	0.0001	26s Proteasome,Akt,ATF2,CCND1,CDX2,CFTR,DRD2,EIF2AK4,ERK1/2,FOS,HSF1,HSF2,JUN,MAPK9,MAPT,NFE2L2,p70S6k,PPARA,RGS4,STAT5a/b,STAT5B,Z-LLL-CHO	ADRM1,AHSA1,AIMP2, ARHGEF2,ASNS,ASPG, BLMH,CA1,CCT8,CTH,CYCS,EEF1A1,EIF4EBP1, FGF13,FGF2,G3BP1,GHR,GPT,HERC4,HPRT1,HSP90AA1,IL1R2,MAGT1,ME,MMP19,MSH3,NCL,NOP58,PARD3,PCLO,PCYT2,PFKFB4,PPM1A,PREP,PRKCZ,PSMA2,PSMA6,PSMC1,PSMC2,PSMC4,PSMC6,PSMD12,PSMD6,PVALB,RANGAP1,RRS1,RUVBL2,SERINC3,SLC16A5,SLC47A1,SRM,ST13,UBE2G1,UBE3A,USP14,WDR3,YBX2

Appendix H (continued).

Inferior Other SB

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	cadmium	chemical toxicant	-	-	-0.229	0.000119	0.0002	afatoxin B1,Akt,cadmium,C EBPB,CYP3A4,EG F,EIF4E,ERK1/2,E RN1,ESR1,ETS1,E ZH2,G6PD,Hdac,ID H1,IGF1,IL12 (complex),Jnk,MAP 2K1,MAPK1,MITF, MTORC1,NANOG, NFE2L2,P38 MAPK,PPARGC1A ,PPIF,PTPRR,reacti ve oxygen species,Rsk,STAT6, TFEB,TNF,TP53	CGNL1,CISH,DNAH2,D USP11,FOXD1,FYCO1,G KAP1,GPR156,KDM4C, MS4A15,MYO6,NSMF,PI P4P1,PSIP1,RPA3,TMEM 50B,TSPAN8,UBR5,WDR 82
2	miR-584-3p (CAGUUC)	mature microRNA	-	bias	-0.816	0.00014	0.0003	miR-584-3p (CAGUUC)	CISH,KDM4C,PIP4P1,PSI P1,RPA3,WDR82
3	KIN001-043	chemical drug	-	-	-1.807	0.000273	0.0003	Akt,Alpha catenin,CEBPB,CT NNB1,EIF4E,ETS1, EZH2,Gsk3,GSK3B ,HTT,IGF1,KIN001 - 043,MITF,NANOG, PIK3CG,PLD1,PPA RGC1A,SRC,STAT 3,STAT6,TFEB,TN F	CGNL1,DNAH2,ELAVL4 ,FOXD1,FYCO1,GFI1,GK AP1,HNRNPA0,MS4A15, PIP4P1,PSIP1,RPA3,TME M50B,TSPAN8,UBR5
4	miR-6768-5p (ACACAGG)	mature microRNA	-	bias	-1.633	0.000303	0.0009	miR-6768-5p (ACACAGG)	FOXD1,GPR156,LEMD1, PIP4P1,UBR5,WDR82

Appendix H (continued).

5	medroxyprogesterone acetate	chemical drug	-	-	1.698	0.000429	0.0017	Akt,Ap1,CDK4,CDK6,CEBPB,EGF,ERK,ERK1/2,ESR1,ETS1,EZH2,IGF1,JUN,medroxyprogesterone acetate,MITF,MMP9,NANOG,NFE2L2,NR3C1,PPARGC1A,PTK2,PTPRR,STAR,STAT5a/b,STAT6,TFEB,TNF	CGNL1,CISH,DNAH2,DISP11,FOXD1,FYCO1,GKAP1,GZMA,MS4A15,MYO6,NSMF,PIP4P1,PSIP1,PTCD3,RPA3,TMEM50B,TSPAN8
6	BRD2	kinase	-	-	0	0.00043	0.0004	BRD2, TOP1	CISH,MYO6
7	ingenol mebutate	chemical drug	-	-	1.069	0.0008	0.0016	CEBPB,EIF4E,FGFR2,FOXO1,Hdac,ingenol mebutate,NANOG,NFE2L2,PIK3CG,PKc(s),PRKCA,PRKCB,PRKCD,PRKE,PRKCG,PRKCH,PRKCZ,PRKG1,RAS,RB1,STAT6,TFEB,TP53	ELAVL4,FOXD1,FYCO1,GKAP1,GPR156,GZMA,HNRNPA0,MS4A15,NSMF,PIP4P1,RPA3,TMEM50B,TSPAN8,UBR5
8	miR-6861-3p (GGACCUC)	mature microRNA	-	bias	-1.342	0.000885	0.002	miR-6861-3p (GGACCUC)	FYCO1,GFI1,GPR156,PTCD3,RPP25L
9	lithium	chemical drug	-	-	-1.213	0.00107	0.0031	Akt,CEBPB,EIF4E,ERK,ETS1,EZH2,Gsk3,GSK3B,IGF1,JUN,lithium,MAPK1,MAPK3,NANOG,NR3C1,P38 MAPK,PI3K (complex),PIK3CG,PPIF,STAR,STAT6,TFEB,TNF,TP53	CGNL1,DNAH2,FOXD1,FYCO1,GFI1,GKAP1,GPR156,GZMA,HNRNPA0,MS4A15,MYO6,PIP4P1,PTCD3,RPA3,TMEM50B,UBR5,WDR82

Appendix H (continued).

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
10	miR-4747-5p (GGGAA GG)	mature microRNA	Inhibited	bias	-2.333	0.00118	0.0038	miR-4747-5p (GGGAAGG)	ELAVL4,GFI1,HNRNPA0,MS4A15,NAXE,RPP25L,SLC38A7,TRIM27,WDR82
Superior Other SB									
1	guanidino propionic acid	chemical drug	-	bias	-1.732	0.000105	0.0001	guanidinopropionic acid	ATPAF2,CYCS,ESRRB
2	miR-16-5p (AGCAG CA)	mature microRNA	Activated	bias	2.324	0.000115	0.0021	miR-16-5p (AGCAGCA)	AADAT,ASF1B,CCNE2,FGF2,INPP5E,KBTBD12,MMP19,PARL,PPP6C,PTCD3,RNF165,SDHAF2,SELENOO,SERBP1,SLC38A7
3	RASGRP2	other	-	-	0.688	0.000136	0.0013	AHR,Akt,CEBPB,EGFR,EIF4E,ESR1,Hdac,ITGB1,ITGB2,ITGB3,JUN,MAP3K14,MYC,Pkc(s),Rap1,RAS,RASGRP2,SPI1,TP53	ADAMTS2,ARHGEF2,ASF1B,ATPAF2,DUSP11,ELAVL4,ESRRB,FOXD1,GP R156,GPT,IL18RAP,IL1R2,MAGT1,MMP19,NUCB1,QDPR,SERBP1,UBR5,VDAC2
4	ruboxistaurin	chemical drug	-	bias	0.816	0.000137	0.0003	Pkc(s),PRKCB,ruboxistaurin,SPI1	CCNE2,CYCS,ELAVL4,FGF2,GFI1,IL1R2
5	tacrolimus	chemical drug	-	-	0	0.000141	0.0014	Akt,calcimycin,CTNNB1,ERK,ERK1/2,HTT,IL2,JUN,MAP2K1,MTOR,NR3C2,PLCG2,tacrolimus,Vegf	ADAMTS2,ANGEL1,ARHGEF2,CCNE2,CYCS,DHPS,ELAVL4,FGF2,GPT,IL18RAP,IL1R2,MAGT1,MMP19,PPP6C

Appendix H (continued).

6	mir-802	microRNA	-	-	0.894	0.000152	0.001	Akt,AKT1,BID,CEBPB,CLOCK,Collagen type II,CTNNB1,EGFR,ESR1,ETS2,GSK3B,HMOX1,ITGB1,MDM2,mir-802,MTOR,NDRG2,NFKBIA,PLD2,PRKCA,PRKCB,PTEN,RHOA,SATB1,SOX2,SOX9,TP53	ADAMTS2,ARHGEF2,ASF1B,ATPAF2,CCNE2,CDH19,CYCS,DUSP11,ESRRB,FGF2,GPR156,GPT,IL18RAP,IL1R2,MAGT1,MMP19,NUCB1,QDPR,RNF11,SERBP1
7	miR-3083-5p (GGCUGGG)	mature microRNA	Activated	bias	2.357	0.00018	0.0048	miR-3083-5p (GGCUGGG)	ADAMTS2,ANGEL1,ARHGEF2,ASF1B,BTD,CLDN19,CYCS,DHPS,ELAVL4,ESRRB,GPR156,INPP5E,KBTBD12,MMP19,NUCB1,RNF165,SLC38A7,UBR5
8	CDK2-CyclinE	complex	-	-	-0.5	0.000191	0.0009	Akt,CCND1,CCNE1,CDC25A,CDK2,CDK2-CyclinE,CDK4/6,CEBPA,CyclinE,EGFR,EIF4E,estrogen receptor,ID2,JUN,MYC,NPM1,Rb,RB1,RUNX2,SMAD3,Y1	ADAMTS2,ARHGEF2,CNE2,CDH19,CYCS,FGF2,GFI1,GPR156,GPT,IL1R2,MAGT1,MMP19,NUCB1,SERBP1,UBR5,VDAC2
9	IL25	cytokine	-	bias	-0.577	0.000195	0.0003	IL25	FGF2,GFI1,IL1R2
10	miR-5579-5p (AUGGUAC)	mature microRNA	-	bias	1.633	0.000221	0.0009	miR-5579-5p (AUGGUAC)	ASF1B,CDH19,CYCS,GFI1,SERBP1,ZNF521

Appendix H (continued).

Market Size SB

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	ibrutinib	chemical drug	Activated	bias	2.309	5.02E-09	0.0001	ibrutinib	AIMP2,ASNS,GNL3,GPI,GZMA,HES1,IPO4,PES1,RRP12,RRP9,SRM,TPI1
2	1,2-dithiol-3-thione	chemical reagent	Inhibited	bias	-2.673	7.05E-09	0.0001	1,2-dithiol-3-thione	GHR,HPRT1,HSP90AA1,PSMA2,PSMA6,PSMB5,PSMC1,PSMC2,PSMC6,PSMD12,RRS1,SERINC3,TP11,USP14
3	miR-503-5p (AGCAGCG)	mature microRNA	Activated	bias	2.813	8.66E-09	0.0001	Akt,BRCA1,CDK2,CDKN1B,CDX2,CBFB,CTNNB1,E2F1,estrogen receptor,FOXO1,ID3,MAPT,miR-503-5p (AGCAGCG),MYC,MYOD1,NCOA1,PARP1,PGR,PTTG1,RB1,RBL1,SMARCA4,SP1,STAT3,TP53,WT1	AEN,AGL,AHSA1,AIMP2,APRT,ASNS,ASPG,BNIP3L,CA1,CENPF,CLDN23,CTPS1,CYCS,DAO,DDX54,DKC1,DPH2,DUS4L,DUSP3,EEF1A1,EIF4EBP1,FBL,FGF13,FTSJ3,GAP43,GHR,GPI,GSTCD,GZMA,HES1,HNMT,KAT14,KLHL23/PHOSPHO2-KLHL23,MME,MSH3,MYO6,NAA15,NCL,NOP58,NSMF,NUDC,PCLO,PES1,PFKFB4,PIP4P1,POLR3D,PPAN,PPID,PPM1A,PRKCZ,PSIP1,PSMA6,PSMB5,PSMD12,PSME3,PUM3,PVALB,RAI14,RANGAP1,RNF207,RPA3,RRS1,RS11D1,RUVBL2,SERINC3,SERPINC1,SRM,ST13,STK32C,TMEM50B,TPI1,TXNL1,UMPS,USP14,USP2,VDAC2,WDR82,ZNF622,ZPR1

Appendix H (continued).

4	GNL1	other	-	-	-0.882	2.76E-08	0.0001	26s Proteasome,AIFM1, Akt,CCND1,CCNE 1,CD40,CDK4/6,C DKN1A,EGR1,EIF 2AK3,EP300,ERK1 /2,FOS,GNL1,HIF1 A,HSF1,Hsp70,MA PT,Mek,MYC,NFE 2L2,NFkB (complex),NFKBIA ,NR3C1,PARP1,ST AT,STAT3,TLR2,T LR4,TP53	ADRM1,AEN,AHSA1,AS PG,BLMH,CCT8,CENPF, CHCHD4,CISH,CTH,DD X27,DDX54,EEF1A1,EIF 4EBP1,ESRRB,FGF13,G3 BP1,GNL3,GPI,GRWD1, GZMA,HERC4,HPRT1,IP O4,MME,MSH3,MYO6,N OP58,NSMF,PCLO,PCYT 2,PES1,PFKFB4,POLR3D ,PPID,PPM1A,PRKCZ,PR MT3,PSIP1,PSMA2,PSM A6,PSMC1,PSMC2,PSMC 4,PSMD12,PSMD6,PSME 3,PUM3,PVALB,RABEP K,RANGAP1,RRP12,RRP 9,RUVBL2,SERINC1,SE RINC3,SERPINC1,SRM,S T13,TRIB3,UBE3A,UMP S,VDAC2
---	------	-------	---	---	--------	----------	--------	---	---

Appendix H (continued).

5	KEAP1	transcription regulator	-	-	-1.94	0.0000000	0.0001	26s Proteasome, Calcine urin protein(s), CEBPA, C HUK, Creb, CTNNB 1, EGFR, GLI1, Gsk3 , HSF1, HTT, IKBKG , IKK (complex), KAT5, K EAP1, MAF, MAFB, MAPT, MYC, NFE2 L2, NFKBIA, p70 S6k, PDX1, SMAD2, STAT3, TBX21, TFE B, TNF, TP73	ADRM1, AEN, AHSA1, AS NS, ASPG, BUD23, CA1, C CT8, CLDN23, CRYGC, CT PS1, DDX27, DDX54, DNA I1, EEF1A1, EIF4EBP1, ES RRB, FGF13, GHR, GKAP1 , GPI, GPRC5B, GTF3A, GZ MA, HERC4, HPRT1, HSP9 0AA1, MSH3, MTDH, NCL , NOP58, NPEPL1, NSMF, N UDC, OLFML2B, PCLO, P FKFB4, PIP4P1, POLR1G, POLR3D, PPID, PPM1A, P REP, PRKCZ, PSIP1, PSM A2, PSMA6, PSMB5, PSM C1, PSMC2, PSMC4, PSMD 12, PSMD6, PVALB, RAI14 , RANGAP1, RCN3, RRS1, RUVBL2, SERINC1, SERI NC3, SRM, ST13, TPI1, UB E3A, USP14, USP2, VDAC 2
---	-------	----------------------------	---	---	-------	-----------	--------	--	--

Appendix H (continued).

6	Z-LLL- CHO	chemical - protease inhibitor	Inhibited	-	-3.538	7.41E-08	0.0001	26s Proteasome,AIFM1, Akt,Ap1,ATF2,AT M,BCL2,beta- estradiol,CASP1,C ASP8,CAT,CCND1 ,CDK1,CDK2,CDX 2,CFTR,CXCL12,c yclooxygenase,DRD 2,EIF2AK4,ERK1/2 ,ESR1,ESR2,EZH2, FBXO32,FOS,GJA 1,H2AX,HSF1,HSF 2,HSF4,hydrogen peroxide,IGF1,IL1B ,IL33,ING1,IRS1,J AK2,Jnk,JUN,LEP R,LIPE,MAP2K4,M AP3K7,MAPK8,M APK9,MAPT,mir- 21,MMP9,MYOD1, NCF4,NFE2L2,NFk B (complex),NR1H3,P 38 MAPK,p70 S6k,PIK3CG,Pkc(s) ,PPARA,PRKACA, PRKN,PRL,PTGS1, PTPRR,RGS4,RXR A,S100A8,S100A9, SQSTM1,STAR,ST AT3,STAT4,STAT 5a/b,STAT5B,STA T6,TAL1,tetradecan oylphorbol acetate,TFEB,TGM 2,trabectedin,VASP, WT1,Z-LLL-CHO	ADNP,ADRM1,AHSA1,A IMP2,APRT,ASNS,ASPG, BANF1,BLMH,CA1,CAC NB4,CCT8,CHCHD4,CIS H,CRYGC,CTH,CYCS,D AO,DDX52,DNAH2,DNA JC7,EEF1A1,EEF1E1,EIF 3B,EIF4EBP1,ELAC2,ES RRB,FGF13,FUT9,FYCO 1,G3BP1,GHR,HERC4,H NRNPA0,HPRT1,HSP90A A1,IPO4,LEAP2,MAK16, MME,MSH3,NCL,NOP58 ,NSMF,NUP133,NXF1,PA RD3,PCLO,PCYT2,PFKF B4,PHETA1,PIP4P1,PPM 1A,PREP,PRKCZ,PSIP1,P SMA2,PSMA6,PSMC1,PS MC2,PSMC4,PSMC6,PS MD12,PSMD6,PVALB,P WP1,RANGAP1,RPF2,RR P12,RRS1,RUVBL2,RXR G,SERINC3,SLC16A5,SL C47A1,SRM,ST13,STK32 C,STRAP,TECR,TSR1,U BE2G1,UBE3A,USP12,U SP14,WDR3,YBX2
---	---------------	-------------------------------------	-----------	---	--------	----------	--------	---	--

Appendix H (continued).

7	BCR (complex)	complex	-	bias	-1.941	0.0000001 2	0.0001	BCR (complex)	AIMP2,ASNS,GNL3,GPI, GRWD1,IPO4,PES1,PRM T3,RABEPK,RRP12,RRP 9,SRM,TPI1
8	SMARCB 1	transcription regulator	-	-	1.179	0.0000002 65	0.0001	26s Proteasome,BRCA1 ,CCND1,Cdc42,CD K4/6,CDKN1A,CD KN1B,CDKN2A,C EBPA,CEBPB,CH UK,CTNNB1,DNM 2,DNMT3A,E2F1,E GFR,EIF4E,ERBB2 ,ESR2,FOS,GLI1,H 2AX,HSF1,KRAS, MAP2K1,NANOG, NFE2L2,NFKBIA,P PARGC1A,PPIF,PR KAA,PRKAA1,PR KAA2,RHOA,SMA D3,SMARCB1,STA T5a/b,TCF12,TCF2 0,TCF3,TCF4,TER T,TP53,trabectedin,t ranscription factor	ACL1,ADAMTS2,ADRM 1,AEN,AHSA1,ASNS,BA NF1,BLMH,CCT8,CENPF ,CLDN23,CTH,DKC1,DU SP3,ENPP6,ESRRB,FBL, FGF13,G3BP1,GPI,GPRC 5B,GTF3A,HERC4,HES1, HPRT1,HSP90AA1,IL18R AP,MS4A15,MSH3,MYO 6,NPEPL1,NSMF,NUDC, NXF1,PACSIN2,PARD3, PCYT2,PDCD11,PES1,PF KFB4,POLR1G,PPM1A,P PP1R3A,PSMA6,PSMC1, PSMC2,PSMC6,PSMD12, PSMD6,PSME3,PUM3,R ABEPK,RAI14,RANGAP 1,RCN3,RRP12,RSL1D1, RUVBL2,SERINC1,SERP INC1,SLC16A5,ST13,TM EM50B,TRIB3,TSR1,UB E3A,UMPS,USP14,USP2, WDR3,WDR82,ZPR1

Appendix H (continued).

9	ABL2	kinase	-	-	-0.585	0.0000003	0.0001	ABL2,CAT,CDC25 A,Cdc42,CEBPA,C REB1,CTSB,CXCL 12,cytarabine,DVL2 ,EIF4E,EPHB2,FOS ,FOXO1,GATA4,G PX1,HSF1,hydroge n peroxide,ITGB1,KR AS,LEF1,MAP2K1, MAP4K4,MAPT,mi r- 21,MMP9,MTOR, MYC,NFKBIA,PTE N,Rac,Ras homolog,RASGRF1 ,STAT3,STAT5a/b, TAFAZZIN,TGFB1 ,TNF,TP53	ADAMTS2,ADNP,ADRM 1,AGL,AHSA1,ASPG,BL MH,BNIP3L,BUD23,CCT 8,CENPF,CTPS1,DDX27, DDX52,DDX54,DHODH, DKC1,EIF4EBP1,EPDR1, ESRRB,FBL,G3BP1,GAP 43,GARS1,GKAP1,GNL3, GPRC5B,GTF3A,GZMA, HERC4,HPR1,MICAL1, MSH3,MTDH,MYO6,NO P58,OLFML2B,PCLO,PC YT2,POLR3D,PPM1A,PR EP,PSIP1,PSMA2,PSMA6 ,PSMB5,PSMC1,PSMC4, PSMC6,PSMD12,PSMD6, PSME3,PUM3,PVALB,R CN3,RSL1D1,RUVBL2,S ERINC1,SERINC3,SERPI NC1,SHTN1,SLC16A5,S RM,STRAP,TXNL1,UBE 2G1,UBE3A,UMPS,USP1 4,USP2,VDAC2,WDR3,Z PR1
---	------	--------	---	---	--------	-----------	--------	--	--

Appendix H (continued).

10	HSPBP1	other	-	-	-0.707	0.0000003	0.0001	26s Proteasome,AGT,AI FM1,Akt,APP,CCN D1,CD40,Cdc42,CR EB1,CTNNB1,EGF R,ERBB2,ERK,FO S,FOXO1,GNAS,H SF1,Hsp70,HSPA1 A/HSPA1B,HSPBP 1,JUN,KRAS,MAP 2K1,MAPK3,MAP T,Mek,MET,MKN K1,NFE2L2,p70 S6k,PARP1,SMAD 2,SREBF1,STAT5a/ b,STAT5B,STAT6, STUB1,TIMP1,TP5 3,TWIST1,WBP2,X BP1	ADAMTS2,AEN,AHSA1, ASPG,BDH1,BLMH,CCT 8,CENPF,CHCHD4,CLD N23,CYCS,DKC1,EEF1A 1,ESRRB,FBL,FYCO1,G3 BP1,GHR,GPRC5B,GRW D1,GTF3A,HERC4,HNM T,HPRT1,IPO4,MME,MS H3,MSRB3,NCL,NOP58, NSMF,NUDC,PARD3,PC LO,PCYT2,PES1,PKM,P OLR1G,POLR3D,PPM1A, PREP,PRKCZ,PRMT3,PR SS53,PSMA2,PSMA6,PS MB5,PSMC1,PSMC2,PS MC4,PSMC6,PSMD6,PS ME3,PUM3,PVALB,RAB EPK,RAI14,RANGAP1,R CN3,RRP12,RRP9,RRS1, RUVBL2,SERINC3,SERP INC1,SHTN1,SLC16A5,S LC47A1,SRM,UMPS,USP 2,WDR3
----	--------	-------	---	---	--------	-----------	--------	---	--

Top Causal Networks for Striped Bass Dam Comparison

Dam A

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	miR-193a-3p (ACUGGCC)	mature microRNA	Inhibited	bias	-3.051	0.0000986	0.0014	miR-193a-3p (ACUGGCC),MTO R	ACSF3,ARFIP1,BHLHE41,CRTC1,CST3,F9,MOG,MTDH,NUDT19,NUP160,NUP42,RHOA,TP53INP1

Appendix H (continued).

2	WAS	other	-	-	-0.943	0.000114	0.0008	Ap1,Cdc42,CDKN1B,CEBPB,Creb,DIA PH3,EDN1,ERK,ES R1,IGF1,IL5,MAP2 K1,MAPK1,MTOR, MYC,NFAT (complex),RHOA,S P1,STAT3,Tgf beta,TOB1,WAS	ADK,ASIC1,ATL2,BHLH E41,CRYL1,CST3,F9,MB NL2,MOG,MTDH,NUDT 19,NUP160,NUP42,OSBP L7,PEMT,RHOT1,SHMT 1,STAT1
3	101.10 peptide	chemical reagent	-	bias	-1.342	0.000171	0.0007	101.10 peptide,5- hydroxytryptamine, APP,CDKN1B,CEB PB,Creb,DIAPH3,E SR1,IGF1,Jnk,JUN, MAPK1,MITF,MT OR,MTORC1,MYC ,NFkB (family),NR3C1,P3 8 MAPK,PTGS2,RH OA,ROCK2,SGK1, SP1,STAT3,TCF3,T GFB1,TNF	ADK,ASIC1,ATL2,CRYL 1,CST3,LIPT1,MASTL,M BNL2,MOG,MPC2,MTD H,NUP160,NUP42,OSBP L7,PEMT,PIR,RHOT1,SH MT1,STAT1,TENT5C
4	miR- 193a-3p (ACUGG CC)	mature microRNA	Inhibited	bias	-3.317	0.000186	0.0012	miR-193a-3p (ACUGGCC)	ACSF3,ARFIP1,BHLHE4 1,CRTC1,CST3,F9,MTDH ,NUDT19,NUP160,NUP4 2,TP53INP1

Appendix H (continued).

5	6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol	chemical reagent	-	bias	1.342	0.000216	0.0008	5-hydroxytryptamine, 6-(7-nitro-2,1,3-benzoxadiazol-4-ylthio)hexanol,APP,CEBPB,Creb,ESR1,GSTP1,IGF1,Jnk,JUN,MAPK1,MITF,MTOR,MTORC1,MYC,NFkB (family),NR3C1,P38 MAPK,PTGS2,SGK1,SP1,STAT3,TCF3,TGFB1,TNF	ADK,ASIC1,ATL2,CRYL1,CST3,F9,LIPT1,MASTL,MBNL2,MOG,MPC2,MTDH,NUP160,NUP42,OSBPL7,PEMT,PIR,SHMT1,STAT1,TENT5C
6	leupeptin	chemical - protease inhibitor	-	-	0.688	0.000282	0.0015	26s Proteasome,ACE,ACE2,AGT,beta-estradiol,CTSB,ESR1,IGF1,leupeptin,MAPK1,MTORC1,NR3C1,PLG,Tgf beta,TGFB1,TP53	ACSF3,ATL2,BHLHE41,CRYL1,CST3,EML2,ETFBKMT,F9,LIPT1,MBNL2,MOG,MPC2,NUP160,NUP42,OSBPL7,PIR,RHOA,STAT1,TP53INP1
7	arsenic trioxide	chemical drug	Activated	-	2	0.000379	0.0016	arsenic trioxide	CST3,PIR,RHOA,SHMT1
8	R59949	chemical drug	-	-	0	0.000414	0.0015	AKT1,AR,CDKN1A,CDKN1B,Creb,ERBB2,ESR1,HIF1A,HMGA1,IL1B,MAPK1,NR3C1,PRKCA,R59949,STAT3,TNF,TP53,VEGFA	ACSF3,ADK,ASIC1,BHLHE41,C19orf25,CEP19,CRYL1,CST3,F9,MASTL,MBNL2,MPC2,MTDH,NUDT19,NUP160,NUP42,SHMT1,STAT1
9	miR-3569-3p (CAGUCUG)	mature microRNA	Inhibited	bias	-3	0.000494	0.0029	miR-3569-3p (CAGUCUG)	ADK,ARFIP1,ATL2,CST3,MAB21L2,MASTL,MOG,NUDT19,TAF11

Appendix H (continued).

10	bosentan	chemical drug	Inhibited	bias	-2.236	0.000516	0.0017	Ap1,APP,bosentan, CDKN1A,CEBPB, EDN1,EDNRA,EDNRB,ERBB2,ERK,ESR1,IGF1,IL1B,INSR,MAP2K1,MAPK1,MAPK8,MITF, MMP2,MTOR,MTORC1,MYC,NFkB (complex),NFkB (family),NR3C1,SP1,SPIB,STAT3,TOB1	ASIC1,ATL2,C19orf25,C RYL1,CST3,MASTL,MBNL2,MOG,MPC2,MTDH, NUDT19,NUP160,NUP42, OSBPL7,PEMT,PIR,RHO A,SHMT1,STAT1,TENT5 C
----	----------	---------------	-----------	------	--------	----------	--------	---	---

Dam B

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	HIF1AN	enzyme	-	-	0	0.0000581	0.0005	AMPK,HIF1A,HIF1AN,NFkB (complex)	ADM,AURKA,BAMBI,CAPNS1,CYBB,FPR2,KIT, NBN,PKM,RAB11FIP4, TN,VIM
2	miR-4425 (GUUGGGA)	mature microRNA	Inhibited	bias	-3.742	0.0000641	0.0007	miR-4425 (GUUGGGA)	ARHGAP9,CDIN1,COIL, DNAAF6,EIF5A,HAPLN1, IFT140,MAT2B,PUM2,R WDD2B,SH2D1B,SNAP25, TPMT,TPTE
3	Huwei1	enzyme	-	bias	0	0.000124	0.0007	ASH1L,DNMT1,FSH,Histone h2a,Histone h4,Huwei1,MACROH2A1,NFkB (complex),NLRP3, NSD1	ADM,ASIP,BAMBI,CAPNS1,CYBB,EPS15,FPR2, I RAK3,KIT,NBN,SNAP25, VIM
4	Vhl	complex	Inhibited	bias	-2.309	0.000126	0.0006	HIF1A,NFkB (complex),VHL,Vhl	ADM,AURKA,BAMBI,CAPNS1,CYBB,FPR2,KIT, NBN,PKM,RAB11FIP4, TN,VIM

Appendix H (continued).

5	TSG101	transcription regulator	-	-	0	0.000132	0.0009	AR,EGFR,NFkB (complex),NR3C1,TSG101	ADM,BAMBI,CAPNS1,CPEB4,CYBB,CYP27A1,EIF5A,EP515,FPR2,GRN,KCNJ2,KIT,MRAS,NAP1L1,NUDT19,PKM
6	neopterin	chemical - endogenous mammalian	-	-	0.302	0.000133	0.0006	EGFR,neopterin,NFkB (complex)	ADM,BAMBI,CAPNS1,CYBB,EIF5A,EP515,FPR2,GRN,KIT,PKM,VIM
7	ADAMTS 5	peptidase	-	bias	0.5	0.000154	0.0013	ACAN,ADAMTS5,Akt,collagenase,EGFR,ERK1/2,MET,NFkB (complex),RAF1,RELA,TIMP3,TLR2	ADM,AURKA,B3GALT6,BAMBI,CAPNS1,CYBB,EIF5A,EP515,FPR2,GRN,IRAK3,KIT,NBN,NLRP12,PKM,VIM
8	COPS5	transcription regulator	-	-	0.48	0.000163	0.0007	Akt,Ap1,APP,CD274,CDK2,CDKN2A,CEBPA,COPS5,EGFR,EPRS1,ERBB2,ESR1,GATA3,Hif1,HIF1A,HNRNPK,IRF1,IL10,IL1B,IRF4,Jnk,JUN,KAT6A,MAPK8,MAPK9,NFkB (complex),NFkB1,NRAS,PDCD1,PGR,PI3K (complex),PRKDC,PTEN,RELA,SMAD4,SMARCA4,STAT3,STAT5B,STAT6,STK11,TNF,TP53,TRAF2	ADGRB1,ADM,AMIGO3,ARFIP2,B3GALT6,BAMBI,CAPNS1,CHRFAM7A,CLTA,CYP27A1,DHRS11,DUSP13,EP515,FAM111A,FPR2,GNAQ,IFT140,IRAK3,KCMF1,KCNJ2,KIT,MRAS,MYBL1,NAP1L1,NBN,NLRP12,NXPE3,P3H4,PKM,PUM2,RAB11FIP4,SH2D1B,SLC47A1,SLC9A3R2,SNX8,TPD52L2,TTN,WDR37,ZBTB25

Appendix H (continued).

9	ID1	transcription regulator	-	-	0	0.000167	0.0012	Akt,AKT1,CAV1,CDK2,Cyclin E,DNMT1,EGFR,ERK1/2,FZR1,HIF1A,ID1,IL10,IL1B,IRF4,KAT6A,NCOA1,NFkB (complex),NFkB (family),NR3C1,NR4A3,RAF1,RB1,SKIL,SMARCA4,STAT5B,STAT6,TCF3,TNF,TP53,WBP2	ADGRB1,ADM,ASIP,AURKA,B3GALT6,BAMBI,CAPNS1,CHRFAM7A,CLTA,COLGALT2,CPEB4,USP13,EIF5A,EP515,FPR2,GMEB1,GNAQ,GRN,IFT140,KCMF1,KCNJ2,MYBL1,NBN,NELFCD,NLRP12,RAB11FIP4,RBM39,SH2D1B,SLC47A1,SLC9A3R2,SNAP25,SNX8,TPD52L2,TTN,VIM,WDR37
10	benzyl isothiocyanate	chemical - endogenous non-mammalian	-	-	-0.707	0.000185	0.0008	Akt,APP,BCL2,benzyl isothiocyanate,CASP3,dexamethasone,FOXO1,FOXO3,GATA3,HNRNP1K,IGF1,IL12 (family),IL1B,IRF4,Jnk,KAT6A,MAPK8,MIF,MTOR,NFE2L2,NFkB (complex),NR3C1,P38 MAPK,PI3K (family),PLAU,PRDM1,PRKCD,PTEIN,STAT3,STAT6,STATK11,TGFB1,TNF,TSC2	ADM,AMIGO3,B3GALT2,B3GALT6,BAMBI,CAPNS1,CHRFAM7A,CKLF,CLTA,CPEB4,CYP27A1,DHRS11,FAM111A,FPR2,GMEB1,GRN,IFT140,KCMF1,KIT,NBN,NXPE3,PKM,RXR,SH2D1B,SLC47A1,SNAP25,SNX8,SYN2,VIM,WDR37,ZBTB25,ZNF91

Appendix H (continued).

Top Causal Networks for Striped Bass Sire Comparison

Sire									
Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	miR-3914 (AGGAA CC)	mature microRNA	-	bias	-0.447	0.00103	0.003	miR-3914 (AGGAACC)	CIPC,LRRC31,NUDT19,P LPP7,SPIN1
2	gentamicin	chemical drug	-	-	1	0.0011	0.0014	gentamicin	CIPC,PELO,RHOA,SQSTM1
3	ARHGAP26	other	-	bias	-1	0.00134	0.0012	ARHGAP26	RHOA
4	CCM2	other	-	bias	-1	0.00134	0.0012	CCM2	RHOA
5	DAAM1	other	-	bias	1	0.00134	0.0012	DAAM1	RHOA
6	LINC01021_Nr_038848.1	other	-	bias	1	0.00134	0.0009	PURPL	SQSTM1
7	RHPN2	other	-	bias	1	0.00134	0.0012	RHPN2	RHOA
8	TPM2	other	-	bias	1	0.00134	0.0012	TPM2	RHOA
9	chondroitin sulfate proteoglycan	chemical reagent	-	bias	1	0.00134	0.0012	chondroitin sulfate proteoglycan	RHOA
10	cyclo (Arg-Gly-Asp-D-Phe-Lys)	chemical reagent	-	biased	-1	0.00134	0.0012	cyclo (Arg-Gly-Asp-D-Phe-Lys)	RHOA

Appendix H (continued).

Top Causal Networks for Striped Bass Sire Size Comparison

Large Sires

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	1-(carboxymethylthio)tetradecane	chemical drug	-	-	-0.728	0.000072	0.0009	1-(carboxymethylthio)tetradecane, ACOX1, APP, FOXO1, IL2, MTOR, NFE2L2, PPARG, PPARGC1A, progesterone, RB1, RELA, SCD, SIRT1, TP53	AQP9, ARHGAP12, CEACAM1, CH25H, CYTIP, DCTN1, DNASE1, EBF2, GSTA3, INPP4B, KSR1, MAPK14, MPEG1, MRPS24, PUM3, RBL2, TBC1D25
2	miR-344d-2-5p (GUCUGGU)	mature microRNA	Inhibited	bias	-2.828	0.0000779	0.0011	miR-344d-2-5p (GUCUGGU)	ABHD14B, CBFA2T2, CRABP1, EBF2, FAM107B, MBOAT1, SIDT2, TBC1D25
3	miR-203b-5p (AGUGGUC)	mature microRNA	Inhibited	bias	-2.646	0.0000849	0.0007	miR-203b-5p (AGUGGUC)	AQP9, CNTNAP5, LNX2, MPEG1, PPM1K, SRRM1, THAP12
4	fidarestat	chemical drug	-	bias	-0.728	0.0000864	0.0006	CASP1, CDK2, ELAVL1, fidarestat, FOXO1, HMGA1, IL1, IL1B, IL33, LYN, MTOR, PI3K (family), PPARGC1A, RELA, STAT1, SYK, TP53	C1QL2, CEACAM1, CYTIP, DCTN1, DNASE1, EBF2, ELMO1, FGFR3, INPP4B, KSR1, MAPK14, MPEG1, MRPS24, PIGO, PRPF4B, PUM3, RBL2
5	miR-12192-5p (GUGGGU)	mature microRNA	Inhibited	bias	-3.317	0.0000878	0.0027	miR-12192-5p (GUGGGU)	C1QL2, CBFA2T2, DCTN1, ELMO1, FAM107B, KSR1, NGFR, PCDHGA2, SIDT2, TMEM198, ZNF740

Appendix H (continued).

6	RSPO1	other	-	-	-0.775	0.0000965	0.001	CTNNB1,FOXO1, HIF1A,LEF1,LRP6, MAPK14,MTOR,P PARGC1A,PRKAA ,RSPO1,SIRT1,STA T1,TP53	AQP9,CEACAM1,CH25H ,DCTN1,DNASE1,EBF2,E LMO1,GSTA3,KSR1,MA PK14,MRPS24,NGFR,PU M3,RBL2,TRPM1
7	poly dA- dT	chemical reagent	Activated	-	2.324	0.000101	0.0008	AIM2 Inflammasome,CAS P1,DDX3X,IL1,IL1 B,IL33,IRF3,MTOR ,poly dA- dT,RELA,STAT1,T BK1,TP53,YAP1	C1QL2,CEACAM1,CH25 H,CYTIP,DCTN1,DNASE 1,ELMO1,FGFR3,INPP4B ,KSR1,MAPK14,MRPS24, NGFR,PRPF4B,PUM3
8	inavolisib	chemical drug	-	-	-1	0.000113	0.0014	ELAVL1,FOXO1,F OXO3,HMGA1,ina volisib,MTOR,NFK BIA,PI3K (family),PIK3CA,P PARGC1A,RELA,S P1,STAT1,TP53	CEACAM1,CH25H,DCT N1,DNASE1,EBF2,ELMO 1,FGFR3,INPP4B,KSR1, MAPK14,MPEG1,MRPS2 4,NGFR,PIGO,PUM3,RB L2
9	miR- 6825-5p (GGGGA GG)	mature microRNA	Inhibited	bias	-4.359	0.000118	0.0077	miR-6825-5p (GGGGAGG)	ABHD14B,AQP9,CBFA2 T2,CCDC106,CEACAM1, CNTNAP5,INPP4B,KSR1 ,MPEG1,NGFR,PCDHGA 2,PIGO,PPM1K,PRPF4B, RBP5,SIDT2,TBC1D25,T MEM198,ZNF740
10	NR3C1	ligand- dependent nuclear receptor	-	-	0	0.000157	0.0017	Akt,Ap1,APP,AR,C AV1,CREBBP,GSK 3B,HMGA1,IL2,Jnk ,NEUROG2,NR1H4 ,NR3C1,PPARGC1 A,PRL,STAT1,STA T5A,STAT6,TP53, TP73	ARHGAP12,CBFA2T2,C EACAM1,CH25H,DNAS E1,EBF2,ELMO1,FGFR3, GSTA3,INPP4B,KSR1,M APK14,MPEG1,MRPS24, NGFR,PIGO,PPM1K,PU M3,RBL2,TBC1D25

Appendix H (continued).

Small Sires

Order	Master Regulator	Molecule Type	Predicted State	Notes	Activation z-score	p-value of overlap	Network bias-corrected p-value	Participating regulators	Target molecules in dataset
1	FRAT1	other	-	-	1.213	0.0000163	0.0003	APP,AR,AXIN1,CTNNB1,EIF6,ESRRA,FRAT1,GSK3B,JUN,MAPT,MYC,NCOA1,PGR,PPARA,PPARG,PSEN1,RB1,SREBF1,STAT1,TCF7L2,TERT,TP53	CSAD,CYTIP,GFM1,GYG1,MRPL9,MRPS24,PDS2,PISD,PSMG2,PUM3,RAB9A,RHOA,RIOX2,RNF213,SC5D,TBC1D14,UCK1
2	miR-4519 (AGCAGUG)	mature microRNA	Inhibited	bias	-2.646	0.0000317	0.0002	miR-4519 (AGCAGUG)	GYG1,NIPSNAP3A,PSMG2,PUSL1,RHOA,SKP1,UCK1
3	SVIL	other	-	-	1.698	0.000108	0.001	ESR1,GLI1,HNF4A,IL6,MYC,NFAT5,PGR,PPARG,RAC1,RB1,STAT1,SVIL,TCF7L2,TP53	CSAD,CYTIP,FABP3,GFM1,GYG1,MRPL9,MRPS24,PCBD2,PDSS2,PISD,PUM3,RHOA,RIOX2,RNF213,SC5D,TBC1D14,UCK1
4	PARVB	other	-	-	0	0.000136	0.0008	APP,AR,EIF6,GSK3B,ILK,KLF5,NCOA1,PARVB,PGR,PPARA,PSEN1,RB1,SREBF1,STAT1,TP53	CSAD,CYTIP,FABP3,GFM1,GYG1,MRPL9,MRPS24,PDSS2,PISD,PSMG2,PUM3,RAB9A,RNF213,TBC1D14
5	TP53INP1	other	-	-	-1	0.000169	0.0009	FOXP1,GLI1,HNF4A,IL6,MYC,PGR,PPARA,PPARG,RB1,Smad2/3,TP53,TP53INP1	CSAD,CYTIP,FABP3,GFM1,GYG1,MRPL9,MRPS24,NXPH1,PCBD2,PDSS2,PISD,PUM3,RHOA,RIOX2,SC5D,UCK1
6	goserelin	biologic drug	-	-	1.732	0.000169	0.0007	FST,GNAS,GNRH1,GNRHR,goserelin,IGF1R,MYC,PGR,RB1,STAT1,TP53	CYTIP,FABP3,GFM1,MRPL9,MRPS24,PDSS2,PISD,PUM3,RAB9A,RHOA,RIOX2,RNF213

Appendix H (continued).

7	miliciclib	chemical drug	-	bias	0.258	0.00017	0.0008	APP,AR,CDK1,CDK2,CDK4,GLI1,miliciclib,MYC,NCOA1,NTRK1,PGR,PPARG,RB1,SIRT2,TP53,YAP1	CYTIP,FABP3,GFM1,GYG1,MRPL9,MRPS24,PCBD2,PDSS2,PISD,PSMG2,PUM3,RIOX2,RNF213,SC5D,UCK1
8	orotic acid	chemical - endogenous mammalian	Activated	-	2.673	0.000171	0.001	AMPK,HNF4A,LP L,orotic acid,PGR,PPARG,PRKAA,RB1,SIRT1,SREBF1,STAT1,TP53,YAP1	CSAD,CYTIP,FABP3,GFM1,GYG1,MRPL9,MRPS24,PDSS2,PISD,PUM3,RAB9A,RNF213,SC5D,UCK1
9	UBE4B	enzyme	-	-	0	0.000176	0.0011	ATXN3,GLI1,HNF4A,IL6,MAPT,MYC,PGR,PPARA,PPARG,RB1,STAT1,TP53,UBE4B,Ubiquitin	CSAD,CYTIP,FABP3,GFM1,GYG1,MRPL9,MRPS24,PCBD2,PDSS2,PISD,PUM3,RHOA,RIOX2,RNF213,SC5D,UCK1
10	CDK1/2	group	-	-	-1.069	0.000189	0.0006	APP,AR,CDK1,CDK1/2,CDK2,MTOR,MYC,NCOA1,PGR,RB1,SIRT2,SKP2,STMN1,TP53,YAP1	CYTIP,GFM1,GYG1,MRPL9,MRPS24,PDSS2,PISD,PSMG2,PUM3,RAB9A,RHOA,RIOX2,RNF213,SC5D
