ABSTRACT

BALIK-MEISNER, MICHELE ROBYN. Gene-by-Environment Interactions Associated with Inter-Individual Variation in Response to Chemical Exposure. (Under the direction of Dr. David M. Reif).

Modern societies are exposed to vast numbers of potentially hazardous chemicals. Despite demonstrated linkages between chemical exposure and severe health effects, there are limited, often conflicting, data on how adverse health effects of exposure differ across individuals. We hypothesized that population variability in response to certain chemicals could elucidate a role for gene-environment interactions (GxE) in differential susceptibility. High throughput screening (HTS) data on thousands of chemicals in genetically-heterogeneous zebrafish were leveraged to identify a candidate chemical (Abamectin) with response patterns indicative of population susceptibility differences. We tested this prediction by generating genome-wide sequence data for 276 individual Tropical 5D (T5D) zebrafish displaying susceptible ('Affected') versus resistant ('Unaffected') phenotypes following identical chemical exposure. We found GxE associated with differential susceptibility in the *sox7* promoter region, then confirmed gene expression differences between phenotypic response classes. The results demonstrate that GxE associated with naturally-occurring, population genetic variation play a significant role in mediating individual responses to chemical exposure.

Additionally, the individual sequencing data from the GxE study was used to assess whether T5D natural diversity was in line with other zebrafish lines or representative of other species. Findings from pooled samples of zebrafish support a supposition of diversity yet cannot directly measure allele frequencies for reference versus alternate alleles. Individual T5D sequences were used to compare observed population genetic variation across species (humans, mice, zebrafish), then across lines within zebrafish. We found more single nucleotide polymorphisms (SNPs) in T5D than have been reported in SNP databases for any of the WIK, TU, TL, or AB lines. We theorize that some subset of the novel SNPs may be shared with other zebrafish lines but have not been identified in other studies due to the limitations of capturing population diversity in pooled sequencing strategies. We establish T5D as a model that is representative of diversity levels within laboratory zebrafish lines and

demonstrate that experimental design and analysis can exert major effects when characterizing genetic diversity in heterogeneous populations.

Chapter 1 of the dissertation discusses computational methods for assessing individual and integrated “omic” data in a Systems Biology framework for environmental and toxicological science. Chapter 2 provides brief background on the specific topic of gene-environment interactions (GxE), as well as the major considerations for mining big data to design a GxE study in a specific population of zebrafish. Chapter 3 discusses the study design, implementation, and results in detail. Chapter 4 looks deeper into population variability that this analysis has brought to the fore. Chapter 5 addresses the impacts of experimental design choices, as well as future directions for the work.

Gene-by-Environment Interactions Associated with Inter-Individual Variation in Response to Chemical Exposure

by
Michele Balik-Meisner

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

Bioinformatics

Raleigh, North Carolina

2017

APPROVED BY:

_____	_____
David Reif Committee Chair	Jung-Ying Tzeng
_____	_____
Jeffrey Yoder	Reade Roberts

DEDICATION

To my loving wife, Colleen Balik-Meisner, and our baby on the way.

BIOGRAPHY

Michele Balik-Meisner was born in New York and raised in New Jersey. She graduated from the Medical Sciences Learning Center at Freehold High School in 2007. She then attended The College of New Jersey (TCNJ) where she graduated *magna cum laude* with a B.A. in Mathematics-Statistics and a minor in Psychology. After participating in the Summer Institute for Training in Biostatistics (SIBS) through North Carolina State University (NCSU) and Duke Clinical Research Institute during college, Michele went on to attend NCSU as a graduate student, completing her Master of Statistics with a concentration in Statistical Genetics. While working on this degree, Michele was a mentor for the SIBS program and a recipient of a National Science Foundation Graduate Research Fellowship. She then continued onto doctoral studies in Bioinformatics at NCSU under the advisement of Dr. David Reif. She has accepted a position as a Bioinformatics Analyst and will continue contributing to the field after receiving her Ph.D.

ACKNOWLEDGMENTS

This culmination of work would not have been possible without the guidance, support, and critical thinking of my advisor, Dr. David Reif. Thank you for the advice you have given me on both personal and professional matters, which has shaped who I have become through this journey toward the completion of my degree. I would also like to thank my committee members: Dr. Jung-Ying Tzeng, Dr. Jeffrey Yoder, and Dr. Reade Roberts for their commitment to my work and their guidance. I have also received continued support on my projects from Dr. Elizabeth Scholl. Additionally, I would like to thank all of the members of the Reif Lab for their help along the way: Dr. Skylar Marvel, Dr. Guozhu Zhang, Kimberly To, Kyle Roell, and Marissa Kosnik. My dissertation work has stemmed from a collaboration with Dr. Robert Tanguay and Dr. Lisa Truong at Oregon State University, from whom I have gained lots of insight about the zebrafish model.

Many others have helped me throughout my academic path at NCSU. Dr. Marie Davidian has given me support starting from my participation in the SIBS program and throughout my graduate studies. I am grateful for the support and guidance I received from Dr. Eric Stone that ultimately led to being awarded a National Sciences Foundation Graduate Research Fellowship. I also learned a lot through my rotation and extended professional and personal interactions with Dr. Alison Motsinger-Reif and Dr. Daniel Rotroff. I would like to thank Dr. Spencer Muse and Dr. David Bird, the co-directors of the Genomic Sciences Graduate Program, and Dr. Fred Wright, the director of the Bioinformatics Research Center. Additionally, Dr. Spencer Muse has been so helpful throughout my transition from Statistics to Bioinformatics. I would also like to thank Dana Ripperton and Babitha Annaji for extensive administrative support and Chris Smith and Kevin Dudley for computing and IT support.

I would have never made it to this point without the undying love and support from my parents, Debra and Arthur Meisner. My father piqued my statistical interests at a young age through his lifelong hobby of horseracing, and my mother was my primary motivator to try out the Bioinformatics I course during my first year of graduate school that solidified my

path from Statistics to its application in Genetics through the field of Bioinformatics. Mom and Dad, you have motivated, sculpted, and given me the tools to succeed in both my academic and personal life! Of course, I would be remiss not to mention my sister, Barrie, who is always there when I need a sarcastic giggle. I would also like to thank my wife, Colleen Balik-Meisner, for putting up with some long nights of analysis and writing and for motivating me to be a better student and person each and every day.

Finally, I would like to thank my entire family, who has been my rock throughout my doctoral career, and all of my friends, who have believed in me and given me strength throughout this endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Computational Methods Used in Systems Biology

Chapter 1 discusses computational methods for assessing individual and integrated "omic" data in a Systems Biology framework for environmental and toxicological science. Chapter 2 provides brief background on the specific topic of gene-environment interactions (GxE), as well as the major considerations for mining big data to design a GxE study in a specific population of zebrafish. Chapter 3 discusses the study design, implementation, and results in detail. Chapter 4 looks deeper into population variability that this analysis has brought to the fore. Chapter 5 addresses the impacts of experimental design choices, as well as future directions for the work.

This chapter contains a book chapter with minor formatting modifications from:

Meisner M, Reif DM. "Computational Methods Used in Systems Biology." *Systems Biology in Toxicology and Environmental Health.* Ed. Rebecca Fry. Academic Press, 2015.

1.1 Introduction

The overarching theme of systems biology is that of complex interactions between multi-scale systems, so it follows that computational methods used in systems biology aim to integrate data and originate from an interdisciplinary slate of scientific fields. To deal with 'omic data generation discussed in previous chapters, suitable analysis methods for systems biology must account for measurements made across scales of time, space, and biological organization. Importantly, analytical methods must first account for the specifics (and peculiarities) of individual technology platforms. For the more established platforms, such as chip-hybridization and sequencing techniques, progress in computational methods research has resulted in a trend toward standardization, where coalescence of statistical methods into powerful software packages handle early stages of analysis in a generally accepted manner. For emerging platforms, computational methods remain diffuse, although popular approaches

share many statistical similarities with more mature methods. Once individual data components have been analyzed, integration into a systems framework can begin.

In this chapter, we present computational methods used in systems biology aligned with the themes discussed above. First, we discuss basic elements of study design common across data types. Second, we present analytical considerations for individual data types organized along the dogmatic progression according to the subheadings Genetics, Epigenetics, Transcriptomics, Proteomics, and Metabolomics. We emphasize shared properties amongst computational methods and highlight data and software resources for each, when available. Third, we survey methods capable of integrating across data types. For each section, we highlight areas of active research necessary to move environmental health and toxicology into a true systems context, where susceptibility to environmental agents can become a predictive science.

1.2 Study Design for Systems Biology

Careful consideration of experimental design is especially important for studies undertaken within a systems biology framework, where the ultimate goal is to integrate across experiments. This is especially true for studies conducted on an ‘omics scale, where failure to account for multiple testing, batch effects, and other considerations associated with high-dimensional data can result in underpowered experiments. As methods mature toward standardization, the statistical rigor and computational transparency required have increased.

Sample sizes and treatments (experimental groups or conditions) will vary according to the particular goals of a given study. In the environmental health sciences (EHS), all basic experimental designs are common, including case-control, quantitative trait/outcome, case-only, observational, and natural experiments. While complete coverage of all contingencies is beyond the scope of this chapter, detailed references on power calculations and design considerations are available (see Table 1.1 for associated software references). For experiments including multiple conditions (e.g. treatments, exposure scenarios, genetic backgrounds), designs lacking adequate sample numbers within conditional “cells” will

sacrifice power. It is often the case that augmenting samples within the referent ("baseline" or "negative control") condition has the greatest effect on detection power. This phenomenon is especially useful for high-throughput screening (HTS) data, because repeated controls can be used to assess batch effects and align results across laboratories or related platforms.

Table 1.1 Software Resources

Name	Data	Function	URL
Pfam	Protein	Amino acid sequence analyses	http://pfam.xfam.org/
Interpro	Protein	Amino acid sequence analyses	http://www.ebi.ac.uk/interpro/
SMART	Protein	Amino acid sequence analyses	http://smart.embl-heidelberg.de/
AbIDconvert	Integrated	Converting between identifiers	http://bioinformatics.louisville.edu/abid/
Cufflinks (Tuxedo Suite)	RNA	Differential expression analyses	http://cufflinks.cbcb.umd.edu/
DAVID	DNA, RNA, protein	Functional annotation	http://david.abcc.ncifcrf.gov/
PLINK	DNA	GWAS analyses	http://pngu.mgh.harvard.edu/purcell/plink/
Systems Biology Workbench	Integrated	Integrated analyses	http://sbw.sourceforge.net
SciMiner	DNA, protein	Literature mining for gene or protein interactions	jdrf.neurology.med.umich.edu/SciMiner/
Chilibot	Integrated	Literature mining for key-word interactions	http://www.chilibot.net/
STRING	Protein	Literature mining for protein–protein interactions	http://string-db.org/
KEGG Mapper	Integrated	Mapping to pathway database	http://www.genome.jp/kegg/tool/map_pathway1.html
BioCyc	Integrated	Mapping to pathway database	http://biocyc.org/
MotifX	DNA, RNA, protein	Motif finding	http://motif-x.med.harvard.edu/
PhosphoMotif Finder	Protein	Motif finding	http://www.hprd.org/PhosphoMotif_finder
Cytoscape	Integrated	Network visualization	http://www.cytoscape.org/
Bioconductor packages for R	All	Omic analyses	http://www.bioconductor.org/
EnrichNet	Integrated	Pathway analysis	http://www.enrichnet.org/
G*Power	All	Power and effect size	http://www.gpower.hhu.de/
PS	All	Power and sample size	http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize
DSTPLAN	All	Power and sample size	https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=41
BLAST	DNA, RNA, protein	Protein and nucleotide searches	http://blast.ncbi.nlm.nih.gov
PICR	Protein, DNA	Protein–gene linking	http://www.ebi.ac.uk/Tools/picr/
CRONOS	Protein, DNA	Protein–gene linking	http://mips.gsf.de/genre/proj/cronos/index.html
BioMart	Protein, DNA	Protein–gene linking	http://www.biomart.org/
Bowtie (Tuxedo Suite)	RNA	Short read alignment	http://bowtie-bio.sourceforge.net/
TopHat (Tuxedo Suite)	RNA	Splice junction read mapping	http://ccb.jhu.edu/software/tophat
R	All	Statistical analyses	http://r-project.org

Using clinical GWAS as an example of a maturing method, sample sizes of well over one thousand are generally required to have sufficient statistical power to detect even modest associations, because hundreds of thousands of single nucleotide polymorphisms (SNPs) are evaluated in a single study (Witte, 2010). Depending on study type, this sample would need to contain both "affected" subjects and unaffected controls (case-control), a spectrum of individuals with different severities of the disease (for quantitative "disease" traits), or "affected" subjects having different environmental exposures or doses (case-only). Subjects should be representative of the population of interest so that valid inferences may be made about genetic associations. For many traits, an association study using unrelated subjects is the only feasible alternative, yet when possible, family-based designs may be especially useful in controlling for common environmental or other exposures. However, family-based designs may confer lower power due to shared genetics (Witte, 2010). For validation, there are genetic databases that contain documented patient information which can sometimes be used in place of a new study cohort (Luca et al., 2008). Outside of the human clinical setting, other considerations enter into design of a genetic study, such as completeness of reference sequences (especially for non-model species), unknown population structure, use of inbred strains, and evolutionary peculiarities (e.g. gene duplication events) (Bowers et al., 2012; Collaborative Cross Consortium, 2012; Ellegren, 2014; Woods et al., 2005).

For experiments measuring data beyond constitutive (static) genetic information, namely epigenetic, gene expression, proteomic, or metabolomic studies, raw sample size must be balanced with the need for measurements across time (e.g. across key points in development), space (e.g. across multiple tissues of interest), and/or dose (e.g. across multiple concentrations in an in vitro study). For these reasons, the absolute sample sizes per treatment condition used in these studies are often smaller than those for GWAS. However, as one moves along the central dogma from DNA sequence towards measures of effect, fewer layers of uncertainty exist between an inferred association and the outcome of interest (Gieger et al., 2008). For example, a misfolded protein that is targeted for intracellular destruction may represent a key step in an enzymatic pathway leading to disease yet be due

to some post-transcriptional mechanism that is unobservable through sequence alone (or at least at the given sample size, using current technologies). A well-designed proteomic study may be able to detect such an effect using a modest sample size if the appropriate temporal and spatial aspects have been covered.

Depending on the number of simultaneous measurements to be made, design considerations must also address multiple testing issues. Multiple testing issues do not come to the forefront in candidate gene studies or targeted follow-up experiments. In contrast, provisions must be made for experiments that are hypothesis-generating, exploratory, or 'omic in scale, where hundreds to millions of measurements are made in as unbiased a manner as possible. If the measurements can be considered independent tests, then simple corrections, such as the Bonferroni can be employed (Dunn, 1961). The Bonferroni significance level (α') is adjusted based on the number of tests performed (n) as $\alpha' \approx \alpha / n$, thus controlling the family-wise error rate (FWER), the probability of making at least one Type I error in the set of tests. This is a conservative approach, where the goal of decreasing the number of false positives must be balanced versus the potential increase in false negatives (i.e. missed disease associations).

For measurements with underlying dependence structure (e.g. linkage disequilibrium between genetic variants or metabolites within a biochemical pathway), more complicated corrections should be employed that attempt to adjust for correlation structure via permutation (Balding, 2006) or approaches that do not assume independence, such as controlling the false-discovery rate (FDR) (Benjamini & Hochberg, 1995). Procedures for controlling the FDR have gained traction, as they maintain higher power to detect smaller biological differences (effect sizes) versus FWER approaches.

1.3 Genetics

Genome-wide association studies (GWAS) examine the genomes of many individuals in a population in order to find some pattern associated with a disease or phenotype of interest. The first study design using the GWAS approach was published in *Nature Genetics*

in 2002 and discovered a region of the lymphotoxin-alpha gene associated with myocardial infarction (MI; "heart attack") susceptibility (Ozaki et al., 2002). The researchers were able to identify one haplotype containing five SNPs (single-nucleotide difference in a DNA sequence compared to the common sequence in the population). Another GWAS was not attempted until 2005 in *Science* that aimed to find genetic regions in humans that were linked to age-related macular degeneration (AMD), which commonly causes blindness in the elderly and is caused by a combination of genetic and environmental risk factors (Klein et al., 2005). This gap was due to the high technical demands in genotyping at least 100,000 SNPs. Technologies that enable this and high-throughput sequencing methods have propelled the field forward in the past decade.

Population association studies search for allele patterns overrepresented in diseased individuals. In searching for meaningful patterns, one must adequately weed out patterns that arise through chance in the large genome; polymorphisms that could have arisen from causal genetic variants are the only ones of interest (Balding, 2006). GWAS analyze common genetic variants to find variants associated with a trait, often looking for variants associated with a disease or drug response.

Association is based on linkage disequilibrium (LD; nonrandom association of alleles across loci). LD is a population-level measure that is highly documented in the human genome and differs for various geographical and ethnic human populations (Shifman, Kuypers, Kokoris, Yakir, & Darvasi, 2003). Recombination hotspots of high LD and low haplotype diversity lead to highly correlated neighboring SNPs in those genetic regions (International HapMap Consortium, 2005). For this reason, GWAS look at many individuals within a population to scan the genome for SNPs associated with a disease or trait. SNPs and genetic regions identified should be validated in subsequent studies and analyses to (1) home in on the specific locations and associations and (2) verify contribution to the trait of interest or LD with other loci (quantitative trait loci; QTLs) that moderate the trait. GWAS studies provide greater statistical power to detect small genomic differences (higher resolution) than linkage-based designs that use recombination rates estimated from family data.

1.3.1 Data Handling and Preprocessing

Errors in data recording, batch effects, and the like could lead to spurious results. It is imperative to check that data were collected effectively and with limited bias. A first data handling and pre-processing check is to make sure the assumptions of Hardy-Weinberg equilibrium (HWE) are met. Deviations from HWE can be due to inbreeding, population stratification, selection, mutation, nonrandom mating, etc. However deviations can also be related to disease association if they are due to a deletion polymorphism (Conrad, Andrews, Carter, Hurles, & Pritchard, 2006) or segmental duplication (Bailey & Eichler, 2006), so care must be taken when deleting significant (deviating from HWE) loci before continuing the analysis (Balding, 2006).

Typically, some flavor of the basic Pearson goodness-of-fit test (chi-square test) is used to test the hypothesis that the population has Hardy-Weinberg proportions [P(AA) = p^2, P(Aa) = $2p(1-p)$, P(aa) = $(1-p)^2$ for diploid organisms with P(A) = p]. Pearson's chi-square test statistic is

$$\chi^2 = \sum \frac{(\#\ observed - \#\ expected)^2}{\#\ expected}.$$

The observed genotypes at each locus can be plugged-in and expectations calculated based on allele frequency and sample size. The test statistic is then compared to the critical value for a chi-squared distribution with one degree of freedom. When genotype counts are low due to small sample size or rare allele frequency, the Fisher exact test is more appropriate (Balding, 2006; J. Wang & Shete, 2012). This test determines a deviation from HWE when the proportion of heterozygotes is significantly different than expected for the sample size. For chi-square or other quality control (QC) statistics from high-dimensional data, quantile-quantile (QQ) plots provide useful visualizations of deviations from expectation.

Other QC steps involve handling missing data. SNP and genotype calling from raw sequencing output is more difficult for rare alleles and heterozygous genotypes since next-generation sequencing (NGS) studies rely on low-coverage sequencing, allowing some diploid individuals to be sampled only at one chromosome of a pair (Nielsen, Paul, Albrechtsen, & Song, 2011). After genotyping many individuals will be "missing" for one or

more genotype(s). Deleting all individuals with missingness from the analysis may excessively reduce the sample size. Imputation techniques have been developed to assign values to missing genotypes. These assume missingness is independent of actual genotype and phenotype, which is generally reliable for tightly linked markers (Balding, 2006). Missing genotypes can be predicted via maximum likelihood estimates (single imputation) or a random selection from a probability distribution (multiple imputations). "Hot-deck" approaches replace the missing genotype for an individual at one locus with the genotype at that locus from an individual with the same genotypes at neighboring loci as the individual with missingness. Many nearest neighbor hot deck imputation approaches and other phasing methods have been developed (Browning & Browning, 2007; Schwender, 2012; Y. Wang et al., 2012). Current methods utilize reference panels and can predict genotypes at SNPs not directly genotyped in the study in order to increase the number of SNPs to boost statistical power to detect associations (Marchini & Howie, 2010). They are identity-by-state (IBS) based or ancestry-weighted, and software are continually evolving (Liu, Li, Wang, & Li, 2013).

1.3.2 Analysis Methods

In a case-control study design, a 2 x 3 matrix of counts (two rows for case/control; three columns for the possible homozygotic or heretozygotic genotypes) can represent a biallelic SNP. A Pearson's chi-square test with two degrees of freedom or a Fisher exact test can be used to test the hypothesis that there is no association between the rows and columns. To augment detection power for additive traits (traits in which heterozygote risk is intermediate between the two homozygote extremes), the Cochran-Armitage test can be used. It is more conservative because it does not assume HWE, but cannot detect overdominance, a phenomenon where the heterozygote phenotype is more extreme than either homozygote. The hypothesis for the Cochran-Armitage test is that the best line between the three genotype risk estimates has slope zero. A possible rule of thumb in test choice is to use Cochran-Armitage when the least common allele appears very infrequently in the population (low

minor allele frequency; MAF) and Fisher when there are enough genotype counts in each category to observe non-additivity. There are also logistic regression case-control approaches that apply $\text{logit}(\pi) = \log(\pi/(1-\pi))$ to the disease risk of each individual. Additionally a score test can be used, which is computationally simpler than logistic regression and provides similar results (Balding, 2006; Wallace, Chapman, & Clayton, 2006).

In the case of continuous outcomes in the framework of single SNP associations, linear regression or analysis of variance (ANOVA) models can be used. For categorical outcomes a multinomial regression analysis is often employed. However, for ordered categories (for example, different severities of a disease) a 'proportional odds' assumption can be applied to add more weight to the information from more severely affected individuals (O'Reilly et al., 2012).

Multiple SNP association tests are more complicated, and the LD structure comes into play. A common analysis strategy is SNP-based logistic regression, where one coefficient is used for each SNP. The degrees of freedom (df) for the basic version of this test are 2 times the number of SNPs, which quickly escalates beyond utility. However, df are limited to the number of sampled individuals, so df for a test must fall below the sample size and leave residual df for error. To address this limitation, constraints can be added. For example, $\beta_1 = \frac{(\beta_0+\beta_2)}{2}$ tests for additive effects and has df equivalent to the number of SNPs (Balding, 2006). Possible covariates (sex, age, environmental exposures, etc.) can also be added into this model, as can interactions between SNPs. Other issues with logistic regression approaches arise from high correlation between predictors. Haplotype-based methods have been developed that can take into account this correlation structure. There is still some uncertainty in these methods because haplotypes are inferred, rather than observed, in typical GWAS designs where phase is uncertain (Balding, 2006).

Given the high-dimensionality of GWAS data and the limitations of traditional parametric approaches when searching for gene-gene interactions (epistasis), a multitude of data mining and machine learning techniques have been applied to genetic data. These include techniques developed expressly for categorical variant data such as multifactor

dimensionality reduction (MDR), decision tree approaches, complex mathematical approaches such as support vector machines (SVMs), and stochastic search methods based upon evolutionary computation (Moore, Asselbergs, & Williams, 2010; Upstill-Goddard, Eccles, Fliege, & Collins, 2013). These methods are often used in combination with traditional models in multi-stage, filter-based strategies or ensemble-learner approaches.

1.3.3 Analytical Challenges and Outlook for Environmental Health and Toxicology

Despite technological and analytical progress, there remain limitations of the current GWAS approaches. GWAS determine association between variants (typically SNPs) and some phenotype of interest but do not necessarily indicate causation. The resulting SNP associations must still be analyzed further to decipher connections or disease risk levels (Witte, 2010). There have been cases of low disease heritability of discovered variants (Eichler et al., 2010), which has prompted further study of less common variants. Additionally, it is very difficult to model covariate effects meaningfully in GWAS designs (Moore, et al., 2010). Creative designs and new methodologies will be necessary to characterize environmental stressors, account for environmental effects, and elucidate gene-environment interactions (GxE) using GWAS. To gain a more complete understanding of the mechanisms of these genetic linkages, and to directly account for environmental components that are difficult to statistically model and account for in human subjects, *in vitro* methods can be utilized. Induced pluripotent stem cells (iPS cells) and population-derived cell line models, when used for GWAS, can predict drug efficacy and toxicity in certain individuals, propelling the field of personalized medicine (Hankowski, Hamazaki, Umezawa, & Terada, 2011; Jack, Rotroff, & Motsinger-Reif, 2014).

According to NHGRI Genome Sequencing Program (GSP), from September 2001 to April 2014 sequencing costs per megabase have decreased from over $5 K to $0.05 (5 orders of magnitude) and genome costs have decreased from $95 M to under $5 K (Wetterstrand). This has been a driving force toward next-generation GWAS projects like the 1000 Genomes Project (http://www.1000genomes.org), launched in January 2008, an international effort to

catalogue human genetic variation in the highest possible resolution. Cheaper, faster, and more accurate technologies will continue to develop, making the "personal genome" attainable and personalized medicine approaches feasible as long as the statistical methods to analyze sequences, SNP-disease linkages, and environmental factors continue to advance accordingly.

1.4 Epigenomics

Epigenomics studies differential gene expression triggered by chemical reactions or other stressors that do not alter the DNA sequence. An epigenome is a cell's full set of epigenetic modifications. As reviewed in (Lim, Tan, & Tong, 2010), epigenomics offers "new opportunities to further our understanding of transcriptional regulation, nuclear organization, development and disease". Epigenomic data present many similar analytical challenges as gene expression data. It is dynamic in time, through the entire course of developmental progression, and dynamic in space spanning cells, tissues, etc. Analytical methods depend on the specific type of "epigenomic" data that is collected.

1.4.1 Data Handling and Preprocessing

The 'omics-scale QC procedures for epigenomic data are relatively immature, compared with the standardization of methods seen in GWAS. While the basics of epigenetic techniques such as chromatin immunoprecipitation (ChIP) are well-characterized, the relatively recent scaling of this field into genome-wide epigenomics (Barski et al., 2007; Laird, 2010) means that QC methods are still evolving. However, for sequencing-based epigenomic technologies, QC shares many similarities with methods originating from GWAS and transcriptomic analysis (see those sections for additional detail).

1.4.2 Analysis Methods

Chromatin immunoprecipitation (ChIP) is usually utilized for the analysis of chromatin modifications. For ChIP-on-chip the goal is to create a ranked list of

overrepresented genomic regions based on raw probe intensities. The procedure involves performing quantile-normalization, conducting a Wilcoxon rank sum test on sliding windows for differential hybridization of probes, merging close significant regions, and ranking z-scores (Bock & Lengauer, 2008). ChIP-seq is used more commonly. The general procedure involves mapping reads, QC, peak detection, data normalization, and identifying significant differential enrichment (Mensaert et al., 2014). Each sequence read directly corresponds to a single chromatin fragment that was bound by the antibody during immunoprecipitation, so little normalization is needed (Bock & Lengauer, 2008).

Modern techniques for DNA methylation mapping allow sequence-level mapping. Methods for alignments, read counts, and the like use the same software as those that were developed for genetic/genomic analyses, such as the Tuxedo suite. Standard analytic tools can also be used for histone modifications, although novel, specific algorithms have been developed for nucleosome positioning and DNA methylation (Mensaert, et al., 2014).

Epigenomic procedures can be integrated within a systems framework and related to other fields in this chapter. Epigenomic data analysis is often combined with gene expression experiments, since transcriptional regulation is thought to be the proximal effect of epigenomic modifications. The underlying DNA sequence highlights potential epigenomic hotspots, low methylation regions (LMRs), and unmethylated regions (UMRs). Other analyses aim to infer epigenetic states from a DNA sequence. Some key areas have been the prediction of promoter locations, CpG islands, DNA methylation regions, and nucleosome positioning. Promoter identification methods combine DNA sequence characteristics with machine-learning algorithms (Bock & Lengauer, 2008). CpG islands mediate open chromatin structure, often overlapping with enhancers and other regulators in a sequence, so a UMR should have fewer CpG-to-TpG mutations, resulting in an overrepresentation of CpGs. Prediction of CpG islands is based on GC and CpG frequencies (Bock & Lengauer, 2008). High false positive rates in studies with analyses relying on these frequencies has led to a new definition of CpG islands proposed based on large-scale epigenome prediction using SVMs (Bock, Walter, Paulsen, & Lengauer, 2007). DNA methylation prediction has

employed sophisticated machine-learning techniques, including neural networks (NNs), SVMs, and hidden Markov models (HMMs) (Lim, et al., 2010).

1.4.3 Analytical Challenges and Outlook for Environmental Health and Toxicology

There are many directions and challenges that still need to be addressed in the realm of epigenomic analysis. Paramount among these are noisy, difficult-to-reproduce data on which it relies and the requirement for millions of cells in standard ChIP-seq. Possible solutions include ChIP-nano (Adli & Bernstein, 2011), which requires <50,000 cells, and single-molecule real-time sequencing (Roberts, Carneiro, & Schatz, 2013). Additionally, causal relationships among different epigenomic events are not well-understood (Sarda & Hannenhalli, 2014). Tying together all of the related pieces of the systems biology framework is a continued challenge and intriguing prospect.

For EHS, epigenomics provides a promising new mechanism to link environmental insults to altered physiology and health. The technical and computational advances discussed in this chapter have opened the door to the analysis of epigenetic modifications in response to a range of environmental factors. These include air toxics such as formaldehyde, where changes in miRNA patterns regulating human lung gene expression can initiate the onset of various diseases (Rager, Smeester, Jaspers, Sexton, & Fry, 2011). Epigenomic data have even been used as evidence of transgenerational inheritance (Daxinger & Whitelaw, 2012; Greer et al., 2011), where exposures in parental generations can manifest in health consequences for offspring. Designing studies to address these questions in human populations is exceedingly difficult, although several applications have made use of umbilical cord blood (Laubenthal et al., 2012; Soubry et al., 2013). If biobank samples can be accessed across generations, sophisticated pedigree analysis techniques may be used to trace inheritance of epigenomic effects.

1.5 Transcriptomics

Transcriptomic studies analyze expression data from genome-wide RNAseq or microarrays to determine which genes are being up- or down-regulated in response to some stimulus, toxin, or disease. Many different microarray platforms exist (by companies such as Illumina, Affymetrix, Agilent, etc.) and have various single-channel and two-channel arrays available for numerous model organisms. There are often multiple probes and probe copies per gene. Array choice, study design, and number of replicates depends largely on the intended goal as well as access to array platforms and cost. For many applications, there has been a shift from DNA microarrays to RNA Sequencing (RNAseq) for assessing gene expression. RNAseq, also known as whole transcriptome shotgun sequencing, sequences cDNA fragments at a particular junction in time via NGS methods. An RNAseq analysis has some advantages over microarrays for differential expression. RNAseq data cover a dynamic range, have a lower background level, and can detect and quantify previously unknown transcripts and isoforms (Soneson & Delorenzi, 2013). *A priori* SNP knowledge is required to create cDNA microarrays but is not necessary for RNAseq, allowing this technology to be used for new organisms whose genomes have not been sequenced (Z. Wang, Gerstein, & Snyder, 2009). However there are some difficulties as well. More reads map to longer genes even when smaller genes have the same expression level. There will also be a different library size for different samples, so direct comparisons are more challenging than with replicates of a microarray (Soneson & Delorenzi, 2013). These two basic platforms for transcriptomics share many similarities, as both offer continuous data on relative transcript abundance. However, initial processing steps vary due to the nature of raw results from microarrays versus RNAseq.

1.5.1 Data Handling and Preprocessing

Many microarray platforms conduct their own quality control filtering, then supply post-processed data. Bioconductor (http://www.bioconductor.org/) R packages exist to help with filtration for these platforms if you prefer to background filter the raw data on your own.

Appropriate quality control techniques will analyze the images to remove or flag low-quality and low-intensity spots and subtract out background intensities. In a microarray experiment there are many sources of non-biological variation that could bias the study results. Batch effects, position of treatments on slides, lab technician error, dye efficiency and heat or light sensitivity, and labeled cDNA hybridization amount differences are all common examples. Normalization procedures minimize this variation so that the signal intensity levels are due to biological differences between experimental conditions. The appropriate normalization approach may depend on the microarray platform used to generate the data. Generating boxplots of the log signal for each channel can help with assessing the need for normalization. One of the simplest approaches is median centering, which shifts the channel distributions so that all are centered at 0. If the spread of each box seems similar, no more normalization may be needed. If necessary, one can apply scale normalization to the median centered data (Yang et al., 2002). More complex approaches involve locally weighted polynomial regression, LOWESS (Cleveland, 1979; Yang, et al., 2002), or quantile normalization (Bolstad, Irizarry, Astrand, & Speed, 2003).

For RNAseq data, processing steps include read mapping, transcriptome reconstruction, and expression quantification. The raw data, consisting of short reads, are aligned to a reference transcriptome or genome. Reconstruction methods can be genome-guided or genome-independent, where the latter is necessary for organisms lacking a reference genome. Normalization procedures are needed to correctly quantify expression levels in the face of technology-specific issues such as differential read counts based upon transcript length and high variability in the number of reads produced in each run/replicate. These issues can be addressed via appropriate computational methods, as have been developed for the older array-based platforms (Garber, Grabherr, Guttman, & Trapnell, 2011).

1.5.2 Analysis Methods

After a gene expression experiment has been conducted and basic pre-processing has been performed, the general analysis procedure for univariate association testing involves comparing samples with a combination of criteria for expression level (e.g. fold-change) and statistical significance (e.g. modified t-test procedures). Multivariate methods are also available and take many forms depending on the unsupervised versus supervised nature of the experiment. The main differences between transcriptomic and GWAS analysis approaches stem from the continuous (versus discrete) readouts and dynamism of gene expression with respect to space and time.

For a typical case-control study, statistical tests can be performed on the expression matrices (dimension: # genes x # experimental scenarios) to compare expression levels for each gene of interest in control individuals to those exposed to the stimulus (or disease status). The same multiple comparison corrections are still necessary when conducting tests on so many genes of interest. Results can be visualized using several techniques, the most common of which are volcano plots (Seifuddin et al., 2013).

As an early 'omics' technology, there are a proliferation of methods for the analysis of transcriptomic data. One such family includes extensions to basic t-tests or ANOVA approaches, where permutation-based cutoffs like the rank product have been applied to expression data. These are less stringent than FDR (have higher statistical power) but require substantial computing time. Significance Analysis of Microarray (SAM) is a permutation-based approach that was developed to combat the high number of false positives in t-test procedures for massive gene sets (Tusher, Tibshirani, & Chu, 2001). The SAM procedure is to (1) run a set of gene-specific t-tests, (2) calculate a score for each gene set based on expression change proportionate to the standard deviation of repeated measures of that gene, (3) mark genes scoring above some threshold as potentially significant, (4) use permutations of the repeated measurements to estimate the percentage of genes identified by chance, the FDR.

Many bioinformatical and machine-learning approaches have been developed on or applied to transcriptomic data. Modeling approaches such as SVMs (M. P. Brown et al., 2000) and other kernel-based methods are adept at building classification models in these data, where the number of predictors often dwarfs the number of samples. Machine-learning approaches can also exploit prior knowledge of gene function to add direction to the search space or recombine with data reduction methods such as principal components analysis (PCA) or multidimensional scaling (MDS).

For hypothesis-generating or unsupervised experiments, clustering is a visual technique that can be used to identify groups of genes with similar expression patterns, all based on the correlation matrix (dimensions: # genes x # genes). Many clustering procedures exist. The following techniques were created with gene expression in mind, and they highlight the progress in the field. Hierarchical clustering (Eisen, Spellman, Brown, & Botstein, 1998) clusters the correlation matrix creating a single dendrogram, or tree diagram, by adding one gene at a time. This creates a relatively strict hierarchy of subsets. Self-organizing maps (Tamayo et al., 1999) try to combat lack of robustness, inversion problems, and the inability to reevaluate clustering along the way in dendrogram approaches. It imposes partial cluster structures, is easy to visualize and interpret, and has good computational properties making it "easy" to run computationally. Simulated annealing (Alon et al., 1999) creates a binary tree, implementing an order to the tree. Then it goes back to visually show this organization in the gene correlation matrix (through rearranging rows and columns to place similar genes together, and through coloring). CLuster Identification via Connectivity Kernels (CLICK) (Sharan, Maron-Katz, & Shamir, 2003) uses a graph-theoretic algorithm and statistical techniques to identify tight groups of highly similar elements, likely to belong to the same true cluster. The approach requires no prior assumptions on the number or structure of clusters, making it less restrictive than some of the earlier methods. It also boasts high accuracy and speed. Modulated Modularity Clustering (Stone & Ayroles, 2009) implements a community structure approach treating transcriptional modules as tight communities in a connected transcriptome graph. It adaptively modulates pairwise

correlations rather than sticking to a particular branch in a dendrogram once it has been placed. The approach also requires no prior cluster number specification, so more clusters are not simply looked at as better.

Clustering can find patterns but is less informative for evidence of statistical significance (Tusher, et al., 2001). It is often used as a dimension-reduction technique. Once the clusters are determined one could apply gene expression techniques to the clusters, massively reducing computing time. The researcher would treat each cluster as a gene and each gene as a probe and follow the typical process outlined above to pinpoint gene clusters with certain expression profiles.

1.5.3 Analytical Challenges and Outlook for Environmental Health and Toxicology

In comparing software for differential expression approaches handling RNAseq data, significant differences were found between methods; however, array-based methods adapted to RNAseq data perform comparably to methods designed for RNAseq. Additionally, increasing the number of replicate samples significantly improves detection power over increased sequencing depth (Rapaport et al., 2013). Therefore, while methods for RNAseq may be able to capitalize on those developed for arrays, basic questions about the correlation between transcriptional abundance and downstream markers of effect must still be addressed (Ghazalpour et al., 2011).

While many environmental stressors exert detectable toxicogenomic effects, the subtlety and complexity (i.e. lack of main effects) of modeling environmental exposures renders transcriptomics more useful for EHS interpretations in combination with other data types (see section 1.8). Nonetheless, gene expression profiling using blood samples may still be informative for detection and prevention of adverse health effects related to chemical exposures (Joseph, Umbright, & Sellamuthu, 2013).

1.6 Proteomics

The addition of the "-omic" suffix to create proteomics is recent, relative to genetics and transcriptomics. Until recently, limitations of biochemical methods and allied technologies, such as two-dimensional gel electrophoresis, have limited proteomic studies to a few proteins, rather than the entire population of expressed proteins in a cell or tissue, the "proteome" (Becker & Bern, 2011). Rapid developments in the scalability of mass spectrometry (MS) has opened up the scale of proteomics, allowing for both targeted and exploratory proteomic investigations (Vidal et al., 2012). Combined with innovative experimental strategies afforded by advances in MS technologies and computational methods, MS-based proteomics now enables 'omic' interrogation, with increasing expectations for quantitation.

As with any technological advancement, numerous analytical challenges have arisen. Chief amongst these is the manual connection of protein data to biological function. The standard with a few or single protein samples is impossible with larger scale data, necessitating bioinformatics as a crucial component of any proteomic analysis. To appreciate the new challenges from high-throughput data, it is useful to understand the workflow for smaller-scale data. In smaller-scale experiments aiming to identify as many proteins as possible in a protein complex, organelle, cell, or tissue lysate, an enzymatic digestion step was usually included. This digestion yielded a large collection of proteolytic peptides that, while not the biological entities of ultimate interest, could be mined for physiochemical and amino acid residue patterns using machine learning approaches, then specifically targeted by specialized mass spectrometric techniques such as multiple reaction monitoring (MRM) (Becker & Bern, 2011; Pan et al., 2009). Newer technologies address many of the obvious limitations of scale inherent in such strategies, although data pre-processing remains an active area of research and crucial step in modern, high-dimensional proteomics.

1.6.1 Data Handling and Preprocessing

For modern proteomic analysis, all proteins from a sample of interest are usually extracted and digested with one or several proteases (typically trypsin alone or in combination with Lys-C) to generate a defined set of peptides (Becker & Bern, 2011). Several enrichment and fractionation steps can be introduced at the protein or peptide level in this general workflow when sample complexity has to be reduced or when a targeted subset of proteins/peptides should be analyzed, as when interested in organelle-specific proteins, post-translationally modified peptides, or well-annotated proteins (Becker & Bern, 2011; Oberg & Mahoney, 2012).

Peptides obtained are subsequently analyzed by liquid chromatography (LC-MS) or gas chromatography coupled to mass spectrometry (GC-MS). The most common approaches used at this stage are designed to achieve deep coverage of the proteome (shotgun MS (Maccarrone, Turck, & Martins-de-Souza, 2010)) or to collect as much quantitative information as possible for a defined set of proteins/peptides (targeted MS (Pan, et al., 2009)). For analysis, peptides eluting from the chromatographic column are selected according to defined rules (discussed below) and further fragmented within the mass spectrometer. The resulting tandem mass spectra (MS2) provide information about the sequence of the peptide, which is key to identification and annotation. For a shotgun approach, no prior knowledge of the peptides present in the sample is required to define peptide selection criteria during the MS analysis. Therefore, the peptides eluting from the chromatographic column are identified in a data-dependent mode, where the most abundant peptides at a given retention time are selected for fragmentation and their masses excluded for further selection during a defined time (Noble & MacCoss, 2012). By using this dynamic exclusion, less abundant peptides are also selected for fragmentation (Hodge, Have, Hutton, & Lamond, 2013).

The data can be displayed as a 3-D map with the mass-to-charge ratios (m/z), retention times (RT) and intensities for the observed peptides as axes, together with fragmentation spectra (MS2) for those peptides that were selected during any of the data

dependent cycles. The intensity of a certain peptide m/z can be plotted against RT to obtain the corresponding chromatographic peak. The area under this curve (AUC) of this peak can be used to quantify the corresponding peptide, with peptide identification typically achieved through its fragmentation spectrum.

The large number of MS2 spectra generated by the current mass spectrometers requires automated search engines capable of identifying and quantifying the analyzed peptides. A review of all the approaches for identification and quantification in proteomics (Link et al., 1999) is beyond the scope of this chapter, but we give an overview of the process here. Briefly, search algorithms aim to explain a recorded MS2 spectrum by a peptide sequence from a pre-defined database, returning a list of peptide sequences that fit the experimental data with a certain probability score or FDR. The databases are normally protein databases translated from genomic data (Mallick et al., 2007), although other strategies like spectral libraries or mRNA databases (Lange, Picotti, Domon, & Aebersold, 2008) have been successfully applied. A final step is then required to assemble the identified peptides into proteins, which can be a challenge, especially when dealing with redundant peptides or alternatively spliced proteins (Deutsch, Lam, & Aebersold, 2008). In any of these cases, several strategies have been described to reduce the false discovery rate of such matching approaches—both for peptide identification and protein assembly (Deutsch, et al., 2008).

This general shotgun/discovery approach leads to the identification of thousands of proteins. Therefore, sensitivity and reproducibility are still major concerns that need to be evaluated in the quality control of proteomics data. Normally, complete coverage of proteins and complexes involved in the same signaling pathway or in the same functional family is not fully achieved. Additionally, reproducibility in protein identification among replicates can vary between 30-60% (Gupta et al., 2008), which is low compared to other ‘omic’ technologies. Newer targeted proteomics approaches attempt to address these issues (Cox & Mann, 2008) via selected reaction monitoring (SRM), where predefined peptides at scheduled RT are selected and fragmented. Due to the increased scan speed and mass

window selectivity of the current mass analyzers, SRM can be simultaneously performed on multiple analytes. This capability led to the multiplexing of SRMs in a method called multiple reaction monitoring (MRM), which has been used to quantify several hundreds of proteins in a broad dynamic range, down to proteins present at very low copy number in the cell (~50 copies/cell) (Kislinger et al., 2006).

The AUC of the monitored fragments is then used for quantification. Spike-ins are becoming an important part of both quality control and the quantification process. By spiking the peptide mixture with isotopically-labelled standard peptides, such targeted approaches can also be used to determine absolute rather than relative quantitation levels of proteins or posttranslational modifications and assess the overall quality and reproducibility of such data (Ishihama et al., 2005; P. Lu, Vogel, Wang, Yao, & Marcotte, 2007). However, since previous knowledge about the proteins is required for targeted approaches, they are typically performed in conjunction with a shotgun approach in truly proteome-wide experiments.

1.6.2 Analysis Methods

After processing and quantification, the output of a proteome experiment is a list of identified and/or unidentified factors that have a probability score and, if applicable, an associated quantitative (or at least semi-quantitative) value. Association analysis with an outcome or experimental condition can be directly performed on the list of protein products provided using traditional tests of hypotheses or nonparametric alternatives. Beyond this first step, additional bioinformatics analysis is conducted to interpret the data and generate testable hypotheses from a systems perspective. To do this, the list has to be further classified and filtered. The first step for a functional analysis of a large protein list is to connect the protein name to a unique identifier. While gene names have been reasonably well standardized, protein names can differ between different databases and even releases of the same database. Although the curation of the most popular databases is constantly improving, this step can still pose quite a computational challenge and lead to a substantial loss of information. Several web-based algorithms exist to connect protein names to their

corresponding gene names (see Table 1.1). After annotating the proteins, many of the common overrepresentation and pathway analysis tools used for transcriptomic and other data (see section 1.8) can be used.

From the beginning of proteomics, appreciation of the importance of protein-protein interaction has been an essential component of analysis. Because a protein can be involved in multiple complexes of varying composition, to completely understand a biological system, it is necessary to analyze the abundant protein complexes as well as the conditions that lead to their formation or dissociation. Databases such as MINT, BioGRID, IntAct, and HRPD (see Table 1.2) specifically include protein-protein interactions, associating those interactions with the biological process in which they function. The interaction information populating these databases derives from empirical data (such as ChIP-ChIP experiments), computational simulation, and/or from literature mining. In fact, there are a number of literature mining tools to screen PubMed abstracts using natural language processing for protein-protein interactions.

Inevitably, working with databases will result in missing data (especially true in recently sequenced organisms). Additionally, there will always be challenges in untargeted proteomics with large numbers of protein products of unidentified function. To learn more about the function of those proteins and how they interact with members of certain pathways, it is helpful to analyze their amino acid sequence for specific folds of protein domains or for motifs for post-translational modifications (see Tables 1.1 and 1.2 for resources mentioned in this section). The simplest analysis represents a BLAST search against the database of known protein sequences to find if proteins with similar amino acid sequences have been described in other organisms. Further, the amino acid sequence can be analyzed by programs such as Pfam, Interpro, SMART, or DAVID to learn if the identified protein shares fold properties with other proteins. These algorithms apply hidden Markov models (HMMs) to classify proteins based upon amino acid sequence and predict the occurrence of a specific protein domain. Knowledge about the abundance of a specific fold could provide evidence for inclusion of unknown proteins into biological networks. Secondly, algorithms such as

MotifX or PhosphoMotif Finder analyze the sequence environment of post-translational modification sites, thereby reporting enrichment of certain amino acid motifs and helping to identify the modifying enzyme.

Table 1.2 Database Resources

Name	Data	URL
1000 Genomes Project	DNA	http://www.1000genomes.org
A Catalog of Published Genome-Wide Association Studies	DNA	http://www.genome.gov/gwastudies/
ArrayExpress	RNA	http://www.ebi.ac.uk/arrayexpress/
BioGRID	RNA, protein	http://thebiogrid.org/
Corum	Protein	http://mips.helmholtz-muenchen.de/genre/proj/corum
dbGaP	DNA	http://www.ncbi.nlm.nih.gov/gap
DrugBank	DNA, RNA, protein, metabolite	http://www.drugbank.ca/
ENCODE Project	Epigenetic (but studies use DNA, RNA, protein data)	http://www.encodeproject.org
Ensembl	DNA, RNA, epigenetic, protein	http://www.ensembl.org/
Entrez Gene	DNA	http://www.ncbi.nlm.nih.gov/gene
Entrez Protein	Protein	http://www.ncbi.nlm.nih.gov/protein
GENSTAT	RNA	http://www.gensat.org/
GEO	RNA	http://www.ncbi.nlm.nih.gov/geo/
GO	Integrated (pathway)	http://www.geneontology.org/
GTEx	Integrated	http://systems.genetics.ucla.edu
HMDB	Metabolite, protein	http://www.hmdb.ca/
HRPD	Protein	http://www.hprd.org/
IntAct	DNA, RNA, protein	http://www.ebi.ac.uk/intact/
International HapMap Project	DNA	http://hapmap.ncbi.nlm.nih.gov/
KEGG	Integrated (pathway)	http://www.genome.jp/kegg/
MeInfoText	Epigenetic (cancer methylation)	http://bws.iis.sinica.edu.tw:8081/MeInfoText2/
MethDB	Epigenetic (methylation)	http://www.methdb.de/
MethPrimerDB	Epigenetic (methylation), DNA or RNA primers	http://medgen.ugent.be/methprimerdb/
MethyLogiX	Epigenetic (methylation)	http://www.methylogix.com/genetics/database.shtml.htm
MINT	Protein	http://mint.bio.uniroma2.it/
MSigDB	Gene set	http://www.broadinstitute.org/gsea/msigdb/index.jsp
NCBI databases	All	http://www.ncbi.nlm.nih.gov/gquery/
NHGRI Histone Database	Epigenetic (histone)	http://research.nhgri.nih.gov/histones/
PeptideAtlas	Protein	http://www.peptideatlas.org/
Personal Genome Project	DNA	http://www.personalgenomes.org
PharmGKB	DNA, RNA	https://www.pharmgkb.org/
PRIDE	Protein	http://www.ebi.ac.uk/pride/archive/
Proteome Exchange Project	Protein	http://www.proteomeexchange.org
PubMeth	Epigenetic (cancer methylation)	http://www.pubmeth.org/
SMPDB	Integrated (pathway)	http://www.smpdb.ca/
Systems Genetics Resource	Integrated	http://systems.genetics.ucla.edu
T3DB	Toxins	http://www.t3db.ca/
UniProtKB	Protein	http://www.uniprot.org/uniprot/

1.6.3 Analytical Challenges and Outlook for Environmental Health and Toxicology

Current challenges faced in systems-level proteomic analysis involve seemingly mundane issues such as the mapping of identified proteins to genomic and microarray identifiers. The many-to-many mapping between proteins and their corresponding genes complicates this problem. Frameworks such as BioMart (See Table 1.1) provide mapping from protein to genomic identifiers but still harbor inconsistencies due to error propagation from legacy issues during automated data integration. Due to historical precedence, most of the ontologies and annotation databases are still 'gene centric', often failing to capture protein-specific characteristics, diversity, and function. On a more abstract level, it is clear that proteomics and other large-scale "post-genomic" technologies will profit tremendously from further investments into accurate and detailed gene and protein ontologies. Indeed, with more comprehensive ontologies, the stronger functional inferences using quantitative proteomics data become.

Although parallels can be drawn between preoteomic and genetic or transcriptomic data with respect to structure and analysis, proteomic data sets are still unique in their constitution and underlying assumptions. Therefore, complete exploitation and optimal harnessing of these data will necessitate development of purpose-built analytical and bioinformatics approaches.

1.7 Metabolomics

The human metabolome is smaller than the genome and the proteome, and because obtaining metabolite samples often involves less-invasive collection methods, the ratio of samples to measured variables in a given study can confer statistical power advantages (van Ravenzwaay et al., 2012). Sampling and preparation of metabolomic data starts with selection of a model organism, type of external stressor, and mode of exposure in the case of environmental metabolomics (Lankadurai, 2013). Typically urine or blood samples are used, but sometimes cerebrospinal fluid, saliva, or erythrocytes are selected. The sampling designs are similar to those discussed earlier regarding the control group, age, gender specifications,

etc. Technologies include liquid chromatography (LC/MS), gas chromatography (GC/MS), and capillary electrophoresis (EC/MS) all coupled to mass spec (Nunes de Paiva, Menezes, & de Lourdes Cardeal, 2014; Sarda & Hannenhalli, 2014). MS-based platforms can detect trace levels of metabolites. Nuclear magnetic resonance (NMR)-based technologies can be used as well. These are nonselective and have easy sample preparation (Lankadurai, 2013).

1.7.1 Data Handling and Preprocessing

Once the data are obtained there are some processing steps. Filtering out chemical noise can be done through a window moving average, median filter in the chromatographic direction, or a Savitzky–Golay type of local polynomial fitting and wavelet transformation. Feature detection is then done to identify signals caused by true ions in order to avoid false positives. Correlation optimized warping (COW) and fast Fourier transform are then often used for alignment of chromatography data (Nunes de Paiva, et al., 2014). Normalization is then performed to remove unwanted systematic bias without tampering with the biological variation in the samples. Normalization techniques in the Transcriptomics section can also apply to this data.

1.7.2 Analysis Methods

The typical data analysis procedure is similar to gene expression and other data already discussed. Data reduction can be accomplished through hierarchical cluster analysis, probabilistic PCA, or discriminant analysis (Nunes de Paiva, et al., 2014). Since the metabolome is relatively small, data reduction is less necessary for this data type, but these techniques can be informative for interactions between metabolites of interest. As with transcriptomic data, identifying differential abundance of metabolites proceeds through data transformation and/or normalization, statistical tests using appropriate multiple comparison corrections or regression-based modeling that includes specific covariates, then clustering and pathway analysis (see Table 1.1 and Table 1.2 for examples of software and pathway databases for metabolomics).

1.7.3 Analytical Challenges and Outlook for Environmental Health and Toxicology

Metabolomics studies are useful because a stress response can be detected more quickly than with other data types. This leads to early warning indicators for potential ecosystem shifts, an increased understanding of the impact of environmental stressors and organisms, and aids in environmental health assessments (Lankadurai, 2013). Through the integration of metabolome data with genome, proteome, and epigenome findings, we can gain more insight into systems biology and EHS.

A major interest of EHS with respect to metabolomics is in their potential use as biomarkers of exposure, detectable remnants of a prior exposure factor. There has been some success in assessing health risks associated with exposure to environmental toxins through metabolite studies recently (C. Lu et al., 2010; Wu et al., 2012) (and genetic, etc. studies dating further back). These studies display the utility in detecting metabolomic biomarkers for environmental toxicity concerns and expression of genes in known pathways. Integration into this pathway framework elucidates a higher biological understanding of 'omic knowledge. To propel metabolome discoveries it will be necessary to gain more standardization in the laboratory and analytical practices for this specific data type. Reporting methods also need a protocol to ensure more consistency among metabolite activity in replicates or similar studies.

1.8 Integration into a Systems Framework

Achieving a true Systems Biology framework requires integration of data collected across levels of biological organization and/or across assays probing specific outputs at each level. Examples of integration across the biological levels covered here include the idea of intermediate phenotypes (Civelek & Lusis, 2014), such as using GWAS data to define quantitative trait loci (QTL), where the quantitative trait of interest is an intermediate phenotype from expression data (eQTL), proteomic data (pQTL), or metabolomics data (mQTL) (Fehrmann et al., 2011; Gieger, et al., 2008; Melzer et al., 2008). Examples of integration across assays include comprehensive analysis of a set of phenotypic readouts

from HTS, such as multiplexed assays or suites of individual assays that have been systematically applied.

1.8.1 Data Handling and Preprocessing

Previous sections have dealt with important pre-processing steps for individual data types, followed by high-level analysis of each type individually. For effective data integration, challenges in reshaping disparate data (or results of high-level analysis) into a common format are exponentiated. To ensure reliability and repeatability of scientific conclusions dependent on data originating from diverse sources, focused investment in software infrastructure is required. Software pipelines for data handling and flexible workflows assure that systematic analyses are implemented (Judson et al., 2012). These pipelines automate the incorporation of new data and standardize outputs for integration with other data sources. Such standardization speeds the development of analysis methods for integration by freeing bioinformaticists and statisticians from excessive "data wrangling" at each stage of analysis (O'Neil & Schutt, 2013).

Mapping and annotation across fields and experimental platforms is another essential step. In studies of environmentally-relevant compounds, common identifier sets should be agreed upon prior to data generation. The level of uniqueness needed will depend on the study, but should be mappable to high-level databases such as PubChem (https://pubchem.ncbi.nlm.nih.gov/) to facilitate meta-analyses. Examples range from common chemical names (which may represent several distinct structures) to more specific CAS numbers (http://cas.org) to structure-based identifiers such as InChIs (http://iupac.org). Ideally, a set of internal identifiers should track compounds all the way from sample origin to allow statistical assessment of data reliability, especially to defend against spurious results from chemical degradation or other procedural artifacts (Judson, et al., 2012).

Outside of test substances, integration also requires mapping and annotation of biological factors such as genes, proteins, and metabolites. On one extreme, the concept of a gene may not be of sufficient granularity if the goal is to identify sequences within a gene

corresponding to differential epigenetic modification or alternatively-spliced mRNA. On the other extreme, when integrating data across biological levels (e.g. transcriptomic probe identifiers and genetic regions) or mapping to biological pathway databases, higher-level identifiers, such as Entrez IDs, may be more useful. There exist several tools to facilitate such mappings, depending on the granularity required (Mohammad, Flight, Harrison, Petruska, & Rouchka, 2012).

1.8.2 Analysis Methods

All analytical methods for integrating systems-level data must account for the combinatoric realities presented by the scale of modern data-generating techniques. While increases in computational processing power will continue, raw speed cannot keep pace with the magnitude of possible multi-way interactions (GxG, GxE, GxPxE, and so forth) that would have to be traversed by simple, brute force algorithms. Accordingly, there has been a proliferation of methods for analyzing systems data, with rapid progress facilitated by the popularity of open-source development tools such as the R language and the Systems Biology Workbench (see Table 1.1). An exhaustive treatment of current versions of particular methods would thus be obsolete almost immediately. Instead, we present illustrative examples of popular methods organized along the two major strategic lines used for systems biology data: (1) direct analysis of associations between data components, and (2) reorganization of data into networks or pathways.

Considering the first strategy, where analysis aims to identify interactions between data components, there has been significant cross-pollination between traditional disciplines of Statistics, Genetics, Computer Science, Mathematics, and related data-intensive subdisciplines. For studies taking an unbiased view of results (often referred to as "hypothesis-generating" or "hypothesis-free"), machine learning approaches are popular as a means to efficiently sift through huge stacks of data in search of the proverbial needle. These methods typically make fewer assumptions than traditional statistics, although both parametric and non-parametric varieties exist. This can be an important advantage for

systems-level data, where distributional properties often differ both within and across data types. In practice, traditional statistics (e.g. t-tests) or information theory measures (e.g. Gini Index) are often used in some aspect of machine learning methods applied to systems data. These hybrid methods may be deployed as a filter (Li et al., 2005), wrapper (Pahikkala, Okser, Airola, Salakoski, & Aittokallio, 2012), or embedded (Rakitsch, Lippert, Stegle, & Borgwardt, 2013) manner.

For applications where a reasonable amount of *a priori* knowledge is available, Bayesian methods are popular (Wilkinson, 2007). In systems biology, this knowledge may be in the form of genetic sequences predicted to be methylation hot spots or truly prior knowledge from a previous experiment whose hypothesis is now being tested (e.g. a candidate eQTL region perturbed by targeted gene editing) (Friedland et al., 2013). In multi-assay HTS experiments, Bayesian approaches can be used to inform estimates of chemical bioactivity by considering patterns of activity across all assays, rather than each assay being treated as a stand-alone piece of information (Wilson, Reif, & Reich, 2014). This family of methods also shows promise for analyzing multiple endpoints collected within single samples, where complex patterns of correlation may exist between related biological processes (Truong et al., 2014).

In the second strategy, data are reorganized into networks or pathways as the functional unit of analysis. These groupings may be constructed based upon curated knowledge, as in biochemical pathways (e.g. the KEGG pathway) or high-level, shared biological functions (e.g. the GO pathway). Groupings may also be derived from empirical results, where data are collected across tissues or time points to derive process models (Jack, Wambaugh, & Shah, 2011). There is intuitive appeal in organizing data in this manner, either as networks for dynamic modeling or pathways/modules to represent shared function.

For modeling networks, data components, such as transcript or protein abundance, are represented as nodes, with relationships between components represented as edges. These edges may be directed or undirected, depending on whether the underlying data specify a directionality (e.g. gene product *X* encodes a transcription factor that modulates the

expression of transcript *Y*). A major advantage of network inference methods is that they can be used to integrate data across multiple levels of organization, from low-level genetic data through apical outcome data (Clark, Dannenfelser, Tan, Komosinski, & Ma'ayan, 2012).

The analysis of data reorganized into pathways (e.g. sets of candidate GWAS associations or differentially-regulated transcripts) presents statistical challenges related to the non-standard distributions of results and irregular correlation structure underlying most ontologies. This correlation structure arises from knowledge bias in pathway construction and the fact that particular entities in a pathway may be present at several levels in a hierarchical ontology such as GO. Simple pathway analysis approaches are based on over-representation analysis (ORA), in which the number of entities (e.g. genes) in a list of interest that are annotated to a pathway is compared to the expected number of entities under the hypergeometric distribution. The logic is that if more entities in a certain pathway are of interest (e.g. significantly up- or down-regulated) than would be expected by chance, then that pathway may be affected by the experimental condition(s). More sophisticated approaches, such as Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) and Significance Analysis of Functional and Expression (SAFE) (Barry, Nobel, & Wright, 2005) address underlying correlation structure of pathway annotations through multi-stage permutation methods.

1.8.3 Analytical Challenges and Outlook for Environmental Health Sciences

As the scale of data and associated results continues to expand, distilling this information into actionable conclusions remains a challenge. This challenge is compounded when presenting integrated results that require knowledge of each component data type as well as heavy doses of information theory. To overcome this challenge, approaches that rely on visualization and user-friendly software will be key (Reif et al., 2013). Visualization is an effective means of communication across disciplines and can act as a bridge between collaborators at the bench and computational scientists analyzing data (Reif et al., 2010).

An essential, as-yet unsolved, requirement for implementing a systems approach for EHS is connecting data to a meaningful exposure context. The "exposome", as the total of all exposures faced by an organism across its lifetime, includes both endogenous and exogenous factors (Nakamura et al., 2014). Measuring the exposome is in its infancy, with concepts such as Environment Wide Association Studies (EWAS) (Patel, Bhattacharya, & Butte, 2010) and Phenome Wide Association Studies (PheWAS) (Denny et al., 2010) emerging to provide analytical frameworks that connect clinical or health data with a range of outcomes. For HTS data, contextualizing *in vitro* concentrations within predicted exposure can be modeled (Wambaugh et al., 2013). However, real-world, personalized exposure measurements remain elusive, due to many technological and privacy issues.

While challenges remain, progress in asthma—a condition with substantial environmental etiology—illustrates the utility of applied computational methods in Systems Biology. Integrating multiple lines of molecular, clinical, and environmental exposure data have transitioned asthma diagnosis from the binary into one of several clinically-relevant subtypes, or "endotypes" (Anderson, 2008). These subtypes, elucidated through analysis of multiscale genetic, transcriptomic, metabolomic, and environmental data, may present more effective, personalized treatment options, as therapies can be tailored to specific etiologies, rather than reliance on clinically-observable symptoms (Williams-DeVane et al., 2013).

1.9 Summary

Environmental health science and toxicology have embraced new genetic, epigenomic, transcriptomic, proteomic, and metabolomic technologies that are able to generate data on a massive scale. Systems Biology provides a theoretical framework that advocates integrating data from these new technologies, but translating this theory into practice will require the development of analytical methods that consider the many contextual layers involved (Krewski et al., 2014). Of particular importance for computational methods development are questions of how data from different biological levels relate to each other, or, restated: How does one robustly characterize the flow of information between

levels? The analysis of biological levels beyond that of static DNA sequence variation is complicated by the dynamism that these measures represent with respect to each other and the apical phenotype (disease state) of interest. Obtaining data that can address such variation in space and time will require experimental systems able to generate repeated samples, such as model organisms (Soste et al., 2014) and *in vitro* systems using cell lines (C. C. Brown et al., 2014) or induced pluripotent stem (iPS) cells obtained from diverse tissues (Sirenko et al., 2013). With such systems-level data, computational methods can be developed that move environmental health and toxicology toward a predictive science.

1.10 References

Adli, M., & Bernstein, B. E. (2011). Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc, 6*(10), 1656-1668.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A, 96*(12), 6745-6750.

Anderson, G. P. (2008). Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet, 372*(9643), 1107-1119.

Bailey, J. A., & Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet, 7*(7), 552-564.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet, 7*(10), 781-791.

Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics, 21*(9), 1943-1949.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell, 129*(4), 823-837.

Becker, C. H., & Bern, M. (2011). Recent developments in quantitative proteomics. *Mutat Res, 722*(2), 171-182.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, 57*(1), 289-300.

Bock, C., & Lengauer, T. (2008). Computational epigenetics. *Bioinformatics, 24*(1), 1-10.

Bock, C., Walter, J., Paulsen, M., & Lengauer, T. (2007). CpG island mapping by epigenome prediction. *PLoS Comput Biol, 3*(6), e110.

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2), 185-193.

Bowers, J. E., Bachlava, E., Brunick, R. L., Rieseberg, L. H., Knapp, S. J., & Burke, J. M. (2012). Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. *G3 (Bethesda), 2*(7), 721-729.

Brown, C. C., Havener, T. M., Medina, M. W., Jack, J. R., Krauss, R. M., McLeod, H. L., et al. (2014). Genome-wide association and pharmacological profiling of 29 anticancer agents using lymphoblastoid cell lines. *Pharmacogenomics, 15*(2), 137-146.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., et al. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A, 97*(1), 262-267.

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet, 81*(5), 1084-1097.

Civelek, M., & Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nat Rev Genet, 15*(1), 34-48.

Clark, N. R., Dannenfelser, R., Tan, C. M., Komosinski, M. E., & Ma'ayan, A. (2012). Sets2Networks: network inference from repeated observations of sets. *BMC Syst Biol, 6*, 89.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scetterplots. *Journal of the American Statistical Association, 74*(368), 829-836.

Collaborative Cross Consortium (2012). The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics, 190*(2), 389-401.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet, 38*(1), 75-81.

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol, 26*(12), 1367-1372.

Daxinger, L., & Whitelaw, E. (2012). Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet, 13*(3), 153-162.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., et al. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics, 26*(9), 1205-1210.

Deutsch, E. W., Lam, H., & Aebersold, R. (2008). PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep, 9*(5), 429-434.

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association, 56*(293), 52-64.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet, 11*(6), 446-450.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A, 95*(25), 14863-14868.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol, 29*(1), 51-63.

Fehrmann, R. S., Jansen, R. C., Veldink, J. H., Westra, H. J., Arends, D., Bonder, M. J., et al. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet, 7*(8), e1002197.

Friedland, A. E., Tzur, Y. B., Esvelt, K. M., Colaiácovo, M. P., Church, G. M., & Calarco, J. A. (2013). Heritable genome editing in C. elegans via a CRISPR-Cas9 system. *Nat Methods, 10*(8), 741-743.

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods, 8*(6), 469-477.

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., et al. (2011). Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet, 7*(6), e1001393.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabé de Angelis, M., Kronenberg, F., Meitinger, T., et al. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet, 4*(11), e1000282.

Greer, E. L., Maures, T. J., Ucar, D., Hauswirth, A. G., Mancini, E., Lim, J. P., et al. (2011). Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. *Nature, 479*(7373), 365-371.

Gupta, N., Benhamida, J., Bhargava, V., Goodman, D., Kain, E., Kerman, I., et al. (2008). Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res, 18*(7), 1133-1142.

Hankowski, K. E., Hamazaki, T., Umezawa, A., & Terada, N. (2011). Induced pluripotent stem cells as a next-generation biomedical interface. *Lab Invest, 91*(7), 972-977.

Hodge, K., Have, S. T., Hutton, L., & Lamond, A. I. (2013). Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS. *J Proteomics, 88*, 92-103.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., et al. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics, 4*(9), 1265-1272.

Jack, J., Rotroff, D., & Motsinger-Reif, A. A. (2014). Lymphoblastoid Cell Lines Models of Drug Response: Successes and Lessons from This Pharmacogenomic Model. *Curr Mol Med*.

Jack, J., Wambaugh, J. F., & Shah, I. (2011). Simulating quantitative cellular responses using asynchronous threshold Boolean network ensembles. *BMC Syst Biol, 5*, 109.

Joseph, P., Umbright, C., & Sellamuthu, R. (2013). Blood transcriptomics: applications in toxicology. *J Appl Toxicol*.

Judson, R. S., Martin, M. T., Egeghy, P., Gangwal, S., Reif, D. M., Kothiya, P., et al. (2012). Aggregating Data for Computational Toxicology Applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System. *Int J Mol Sci, 13*(2), 1805-1831.

Kislinger, T., Cox, B., Kannan, A., Chung, C., Hu, P., Ignatchenko, A., et al. (2006). Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell, 125*(1), 173-186.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*(5720), 385-389.

Krewski, D., Westphal, M., Andersen, M. E., Paoli, G. M., Chiu, W. A., Al-Zoughool, M., et al. (2014). A framework for the next generation of risk science. *Environ Health Perspect, 122*(8), 796-805.

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet, 11*(3), 191-203.

Lange, V., Picotti, P., Domon, B., & Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol, 4*, 222.

Lankadurai, B. P., Nagato, E. G., Simpson, M. J. (2013). Environmental metabolomics: an emerging approach to study organism responses to environmental stressors. *Environmental Reviews, 21*(3), 180-205.

Laubenthal, J., Zlobinskaya, O., Poterlowicz, K., Baumgartner, A., Gdula, M. R., Fthenou, E., et al. (2012). Cigarette smoke-induced transgenerational alterations in genome stability in cord blood of human F1 offspring. *FASEB J, 26*(10), 3946-3956.

Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L., et al. (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics, 85*(1), 16-23.

Lim, S. J., Tan, T. W., & Tong, J. C. (2010). Computational Epigenetics: the new scientific paradigm. *Bioinformation, 4*(7), 331-337.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., et al. (1999). Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol, 17*(7), 676-682.

Liu, E. Y., Li, M., Wang, W., & Li, Y. (2013). MaCH-admix: genotype imputation for admixed populations. *Genet Epidemiol, 37*(1), 25-37.

Lu, C., Wang, Y., Sheng, Z., Liu, G., Fu, Z., Zhao, J., et al. (2010). NMR-based metabonomic analysis of the hepatotoxicity induced by combined exposure to PCBs and TCDD in rats. *Toxicol Appl Pharmacol, 248*(3), 178-184.

Lu, P., Vogel, C., Wang, R., Yao, X., & Marcotte, E. M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol, 25*(1), 117-124.

Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H. E., et al. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet, 82*(2), 453-463.

Maccarrone, G., Turck, C. W., & Martins-de-Souza, D. (2010). Shotgun mass spectrometry workflow combining IEF and LC-MALDI-TOF/TOF. *Protein J, 29*(2), 99-102.

Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., et al. (2007). Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol, 25*(1), 125-131.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511.

Melzer, D., Perry, J. R., Hernandez, D., Corsi, A. M., Stevens, K., Rafferty, I., et al. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet, 4*(5), e1000072.

Mensaert, K., Denil, S., Trooskens, G., Van Criekinge, W., Thas, O., & De Meyer, T. (2014). Next-generation technologies and data analytical approaches for epigenomics. *Environ Mol Mutagen, 55*(3), 155-170.

Mohammad, F., Flight, R. M., Harrison, B. J., Petruska, J. C., & Rouchka, E. C. (2012). AbsIDconvert: an absolute approach for converting genetic identifiers at different granularities. *BMC Bioinformatics, 13*, 229.

Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics, 26*(4), 445-455.

Nakamura, J., Mutlu, E., Sharma, V., Collins, L., Bodnar, W., Yu, R., et al. (2014). The endogenous exposome. *DNA Repair (Amst), 19*, 3-13.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet, 12*(6), 443-451.

Noble, W. S., & MacCoss, M. J. (2012). Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput Biol, 8*(1), e1002296.

Nunes de Paiva, M. J., Menezes, H. C., & de Lourdes Cardeal, Z. (2014). Sampling and analysis of metabolomes in biological fluids. *Analyst, 139*(15), 3683-3694.

O'Neil, C., & Schutt, R. (2013). *Doing Data Science*: O'Reilly Media, Inc.

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M. R., et al. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One, 7*(5), e34861.

Oberg, A. L., & Mahoney, D. W. (2012). Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. *BMC Bioinformatics, 13 Suppl 16*, S7.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., et al. (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet, 32*(4), 650-654.

Pahikkala, T., Okser, S., Airola, A., Salakoski, T., & Aittokallio, T. (2012). Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms Mol Biol, 7*(1), 11.

Pan, S., Aebersold, R., Chen, R., Rush, J., Goodlett, D. R., McIntosh, M. W., et al. (2009). Mass spectrometry based targeted protein quantification: methods and applications. *J Proteome Res, 8*(2), 787-797.

Patel, C. J., Bhattacharya, J., & Butte, A. J. (2010). An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One, 5*(5), e10746.

Rager, J. E., Smeester, L., Jaspers, I., Sexton, K. G., & Fry, R. C. (2011). Epigenetic changes induced by air toxics: formaldehyde exposure alters miRNA expression profiles in human lung cells. *Environ Health Perspect, 119*(4), 494-500.

Rakitsch, B., Lippert, C., Stegle, O., & Borgwardt, K. (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics, 29*(2), 206-214.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., et al. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol, 14*(9), R95.

Reif, D. M., Martin, M. T., Tan, S. W., Houck, K. A., Judson, R. S., Richard, A. M., et al. (2010). Endocrine profiling and prioritization of environmental chemicals using ToxCast data. *Environ Health Perspect, 118*(12), 1714-1720.

Reif, D. M., Sypa, M., Lock, E. F., Wright, F. A., Wilson, A., Cathey, T., et al. (2013). ToxPi GUI: an interactive visualization tool for transparent integration of data from diverse sources of evidence. *Bioinformatics, 29*(3), 402-403.

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biol, 14*(6), 405.

Sarda, S., & Hannenhalli, S. (2014). Next-generation sequencing and epigenomics research: a hammer in search of nails. *Genomics Inform, 12*(1), 2-11.

Schwender, H. (2012). Imputing missing genotypes with weighted k nearest neighbors. *J Toxicol Environ Health A, 75*(8-10), 438-446.

Seifuddin, F., Pirooznia, M., Judy, J. T., Goes, F. S., Potash, J. B., & Zandi, P. P. (2013). Systematic review of genome-wide gene expression studies of bipolar disorder. *BMC Psychiatry, 13*, 213.

Sharan, R., Maron-Katz, A., & Shamir, R. (2003). CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics, 19*(14), 1787-1799.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., & Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet, 12*(7), 771-776.

Sirenko, O., Cromwell, E. F., Crittenden, C., Wignall, J. A., Wright, F. A., & Rusyn, I. (2013). Assessment of beating parameters in human induced pluripotent stem cells

enables quantitative in vitro screening for cardiotoxicity. *Toxicol Appl Pharmacol, 273*(3), 500-507.

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics, 14*, 91.

Soste, M., Hrabakova, R., Wanka, S., Melnik, A., Boersema, P., Maiolica, A., et al. (2014). A sentinel protein assay for simultaneously quantifying cellular processes. *Nat Methods*.

Soubry, A., Schildkraut, J. M., Murtha, A., Wang, F., Huang, Z., Bernal, A., et al. (2013). Paternal obesity is associated with IGF2 hypomethylation in newborns: results from a Newborn Epigenetics Study (NEST) cohort. *BMC Med, 11*, 29.

Stone, E. A., & Ayroles, J. F. (2009). Modulated modularity clustering as an exploratory tool for functional genomic inference. *PLoS Genet, 5*(5), e1000479.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545-15550.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., et al. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A, 96*(6), 2907-2912.

Truong, L., Reif, D. M., St Mary, L., Geier, M. C., Truong, H. D., & Tanguay, R. L. (2014). Multidimensional in vivo hazard assessment using zebrafish. *Toxicol Sci, 137*(1), 212-233.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A, 98*(9), 5116-5121.

Upstill-Goddard, R., Eccles, D., Fliege, J., & Collins, A. (2013). Machine learning approaches for the discovery of gene-gene interactions in disease data. *Brief Bioinform, 14*(2), 251-260.

van Ravenzwaay, B., Herold, M., Kamp, H., Kapp, M. D., Fabian, E., Looser, R., et al. (2012). Metabolomics: a tool for early detection of toxicological effects and an opportunity for biology based grouping of chemicals-from QSAR to QBAR. *Mutat Res, 746*(2), 144-150.

Vidal, M., Chan, D. W., Gerstein, M., Mann, M., Omenn, G. S., Tagle, D., et al. (2012). The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. *Clin Proteomics, 9*(1), 6.

Wallace, C., Chapman, J. M., & Clayton, D. G. (2006). Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *Am J Hum Genet, 78*(3), 498-504.

Wambaugh, J. F., Setzer, R. W., Reif, D. M., Gangwal, S., Mitchell-Blackwood, J., Arnot, J. A., et al. (2013). High-throughput models for exposure-based chemical prioritization in the ExpoCast project. *Environ Sci Technol, 47*(15), 8479-8488.

Wang, J., & Shete, S. (2012). Testing departure from Hardy-Weinberg proportions. *Methods Mol Biol, 850*, 77-102.

Wang, Y., Cai, Z., Stothard, P., Moore, S., Goebel, R., Wang, L., et al. (2012). Fast accurate missing SNP genotype local imputation. *BMC Res Notes, 5*, 404.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63.

Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). from www.genome.gov/sequencingcosts

Wilkinson, D. J. (2007). Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform, 8*(2), 109-116.

Williams-DeVane, C. R., Reif, D. M., Hubal, E. C., Bushel, P. R., Hudgens, E. E., Gallagher, J. E., et al. (2013). Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes. *BMC Syst Biol, 7*, 119.

Wilson, A., Reif, D. M., & Reich, B. J. (2014). Hierarchical dose-response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics, 70*(1), 237-246.

Witte, J. S. (2010). Genome-wide association studies and beyond. *Annu Rev Public Health, 31*, 9-20 24 p following 20.

Woods, I. G., Wilson, C., Friedlander, B., Chang, P., Reyes, D. K., Nix, R., et al. (2005). The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res, 15*(9), 1307-1314.

Wu, B., Liu, S., Guo, X., Zhang, Y., Zhang, X., Li, M., et al. (2012). Responses of mouse liver to dechlorane plus exposure by integrative transcriptomic and metabonomic studies. *Environ Sci Technol, 46*(19), 10758-10764.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res, 30*(4), e15.

CHAPTER 2

Experimental Design Considerations for a Gene-by-Environment (GxE) Association Study in Zebrafish

Chapter 1 discusses computational methods for assessing individual and integrated "omic" data in a Systems Biology framework for environmental and toxicological science. Chapter 2 provides brief background on the specific topic of gene-environment interactions (GxE), as well as the major considerations for mining big data to design a GxE study in a specific population of zebrafish. Chapter 3 discusses the study design, implementation, and results in detail. Chapter 4 looks deeper into population variability that this analysis has brought to the fore. Chapter 5 addresses the impacts of experimental design choices, as well as future directions for the work.

2.1 Hypothesis

Interindividual (i.e. population) genetic variation contributes to differential response to environmental chemical exposure.

2.2 Introduction

New strategies are needed to bridge the data gap between the bioactivity of chemicals in the environment versus existing hazard information. In particular, little is known about why the effects of environmental agents including industrial chemicals, manufacturing byproducts, metals, pesticides, and herbicides differ between individuals and across populations, although there is emerging evidence that GxE play an important role in health outcomes. While there is continued progress in genetic technologies (sequencing, methylation profiling, and fine-scale control of gene expression, etc.) and high-throughput screening (HTS), the salient problems in solving the basic Y=GxE equation remain: inability to contextualize massive genetic data (G) in the face of insufficient characterization of environmental exposures (E) and incomplete representation of phenotypic endpoints (Y).

2.2.1 GxE Analysis Within a GWAS Framework

For more than a decade, researchers have been scouring the human genome for markers associated with disease or other phenotypes in genome-wide association studies (GWAS). Knowledge of many genetic markers has been gained through this endeavor. However, certain phenotypes, especially diseases that are caused by interactions between a person's genetic makeup and environmental factors. A 2005 *Nature Reviews Genetics* article asserts:

> If we estimate only the separate contributions of genes and environment to a disease, and ignore their interactions, we will incorrectly estimate the proportion of the disease (the 'population attributable risk') that is explained by genes, the environment, and their joint effect. Restricting analysis of environmental factors in epidemiological studies to individuals who are genetically susceptible to the exposure should increase the magnitude of relative risks, increasing our confidence that the observed associations are not due to chance (Hunter 2005).

By adding in the environment as a covariate or interaction effect in a statistical model, and in turn providing extra data that (when missing) would have been contributing to the error term, there is more power to detect a genetic effect. GxE interactions also help explain why certain subgroups of a population are more genetically susceptible to certain diseases.

Measuring GxE effects is important to gaining a more complete understanding of the interplay between our genetic makeup, our environment, and our ultimate phenotypes. However, modeling environmental interplay adds a complexity that can be difficult to incorporate into an experimental design and statistical modeling framework. Some of these complications stem from the difficulty in measuring environmental factors and exposures. Especially in human studies, with participants/patients that have not been confined to a controlled environment, it is difficult for researchers to know what environmental factors to include or what timing or duration of exposure to include in an assessment. Obtaining accurate measures of exposure are difficult, because measurements of lifetime exposure tend

to be biased (Thomas et al. 1993), especially if researchers are left relying on self-reporting of habits (like smoking) or micro-environments.

An early approach for assessing GxE effects within a GWAS framework implemented a 2-step design (Murcray et al. 2009). Step 1 involved screening the SNPs using a likelihood ratio test of association between G and E in a logistic model, using the case-only standard GxE approach but on their combined case-control data to determine a potential subgroup of SNPs that could be involved in GxE interactions. This was done to attempt to screen for the most likely candidates of GxE without just (1) using G candidates (since that list may miss important GxE candidates) or (2) testing all SNPs for GxE effects (which did not seem feasible). The second step involved a likelihood ratio test of the null hypothesis that the log of the GxE odds ratio was equivalent to zero for each SNP that passed the screening methodology.

There are additional statistical difficulties with including GxE effects. Hypothesis tests for interaction have lower power than those for main effects, therefore much larger sample sizes are needed to compute interaction effects (Khoury and Wacholder 2009). In fact, the rule of thumb is that the sample size will need to be four times larger to detect an interaction effect of equivalent magnitude to a desired main effect (Smith and Day 1984). Sample size concerns, coupled with biological plausibility that markers within a gene are functionally related has led to the use of gene-set analyses or other more complicated approaches to assess GxE associations in GWAS (Marceau et al. 2015; Tzeng et al. 2011; Zhang et al. 2014; Zhao et al. 2015). Some factors that motivate the assessment of GxE effects at gene level include (1) it decreases necessary sample size since grouping SNPs together decreases the number of multiple tests to correct for, (2) correlation of SNP genotypes within a gene are often higher than across genes (if SNPs are in linkage disequilibrium), and (3) chemicals or other environmental factors may be interacting with proteins, gene products, or pathways.

In summary, GxE interactions play a critical role in understanding susceptibility to diseases and other phenotypic traits (Hunter 2005). It is possible to study an individual

exposure/environment and its relation to a certain gene, but additional insight is gained by testing a multitude of genes (and multiple environments) in a GWAS design. Enhanced statistical methods have been developed to handle his type of data in human GWAS. However, these do not directly address variation within human disease phenotypes and lifetime exposures.

2.2.2 Setting Up GxE Analysis in a Zebrafish GWAS

The zebrafish genome, with over 26,000 protein-coding genes consisting of orthologues for around 70% of human genes, has gained momentum as a model organism in vertebrate genomics (Howe et al. 2013). Toxicity research has benefited from the short generation time, rapid lifecycle, and clear zebrafish embryos. Various morphological endpoints are easy to observe, and effects of multiple chemical exposures on these outcomes have been studied (Asharani et al. 2015, 2008; Bai et al. 2009; Truong et al. 2014; Usenko et al. 2007). Rennekamp and Peterson recount the upward trend in zebrafish chemical screens, especially toward screens of many chemicals in many fish, primarily on 96-well plates (2015).

It was not until recently that toxicity researchers using the zebrafish model have integrated genetics assays to further understand the mechanisms underlying morphologic response to chemical exposures. With zebrafish emerging as a model organism it is important to note the strain/line diversity. There is apparent phenotypic variability from lab to lab and line to line. A group of researchers from France and the UK compared the locomotion (distance, speed, time) of six common zebrafish strains (AB, casper, EK, TU, WIK) including AB lines from two different facilities to determine the stability and repeatability of these behaviors (Lange et al. 2013). Their results demonstrate large variability in locomotion and fast swim events between strains and between laboratories across time.

Lack of concordance in results from similar studies in different laboratories hinders reproducibility efforts. One possible source of discord is genetic heterogeneity between labs, and early work has been done to sequence the genomes of different populations of laboratory

zebrafish. In most of these studies whole genome or exome sequencing has been done at high coverage levels on one or two fish (LaFave et al. 2014; Patowary et al. 2013) or with mid-level coverage on relatively small samples of pooled fish (Butler et al. 2015). Findings on separate lines of zebrafish are in agreement with phenotypic findings; there may be somewhere between 5 and 15 million SNPs segregating in a zebrafish population (Butler et al. 2015; Obholzer et al. 2012), and about half of the variants are population-specific.

Our aim is to tie differences in morphological response to chemical exposure to capture genetic and phenotypic diversity in an established (but previously unsequenced) population of laboratory zebrafish and to link susceptibility to chemical exposure with natural genetic variability. The Tropical 5D (T5D) zebrafish line, an "outbred" population of unknown genetic heterogeneity, was used to screen over a thousand chemicals for adverse biological responses (Truong et al. 2014). By exploiting the large-scale, high-throughput design of previous chemical toxicity work using the same line of zebrafish, we can select a target chemical whose exposure produces repeatable patterns of varied response in exposed zebrafish. We can then test the theory that gene-environment interactions (GxE) play an important role in differing susceptibility across human and non-human populations and can be detected using this zebrafish system that allows comprehensive analysis of phenotypic endpoints, genetic factors, and environmental exposures. By performing a GWAS study in a non-human organism, we have control over the exact exposure and environmental upbringing for each individual, allowing us to compare the genetic backgrounds of susceptible and unsusceptible individuals that are having opposing response to the same controlled exposure. This fine-tune control allows for direct comparisons between individuals in each group without the necessity for complex statistical methods that have been developed for human GWAS studies.

2.3 Power Analysis

2.3.1 Methods

Power calculations were based on the idea of applying an exact test (due to the possibility of small counts in some contingency table cells) to a 2x3 contingency table of genotypes at a bi-allelic locus (alleles denoted “A” and “a”) or 2x2 contingency table of allele counts for each variant of interest. The rows correspond to cases = fish affected by the endpoint(s) of interest; and controls = fish unaffected by the endpoint(s) of interest. These will be referred to as “Affected” and “Unaffected” so as to not confuse the Unaffected with experimental controls (fish never exposed to any concentration of test chemical).

It was proposed that all living fish at the concentration [c*] that has the proportion closest to 50% Affected:50% Unaffected at the endpoint of interest have their genomes sequenced and an exact test carried out for each variant. This plan was developed in an effort to increase power to detect genomic differences between Affected and Unaffected fish while reducing costs by not sequencing individuals in concentrations lacking enough effect to capture the differences. Experimentally, this would entail a first, range-finding study that uses the six [0uM … 64uM] concentrations to confirm the 50:50 ratio concentration, then a second run using plates containing only [0] and [c*], from which DNA would be extracted in a batch manner from entire plates. It should be noted that 50:50 may not be optimal for future studies, depending on the allele frequencies for variants of interest, but in the absence of allele frequency information, 50:50 has the highest expected power.

Given that the character of genetic variation to be encountered (allele frequencies, relative abundance of common polymorphisms, etc.) was not known before developing the study, the first power consideration was how many samples need to be sequenced to observe alleles at a certain frequency. This allows us to make a statement such as: "By sequencing n samples, we have X% confidence that we would observe at least one variant (allele) that occurs at Y% frequency in the population." Figure 2.1 shows an example of such a calculation using relatively low allele frequencies of 1%-10%. By sequencing at least 29 fish,

we have 95% confidence that we would observe at least one variant (allele) that occurs at 10% frequency in the population. For an allele with MAF 5% we would need to sequence at least 59 fish, and for MAF 1% we would need to sequence at least 299 fish to ensure the same confidence. This demonstrates the importance of these data for characterizing genetic variation in this population—it informs us of how many samples are needed to even encounter a variant at a given population frequency.

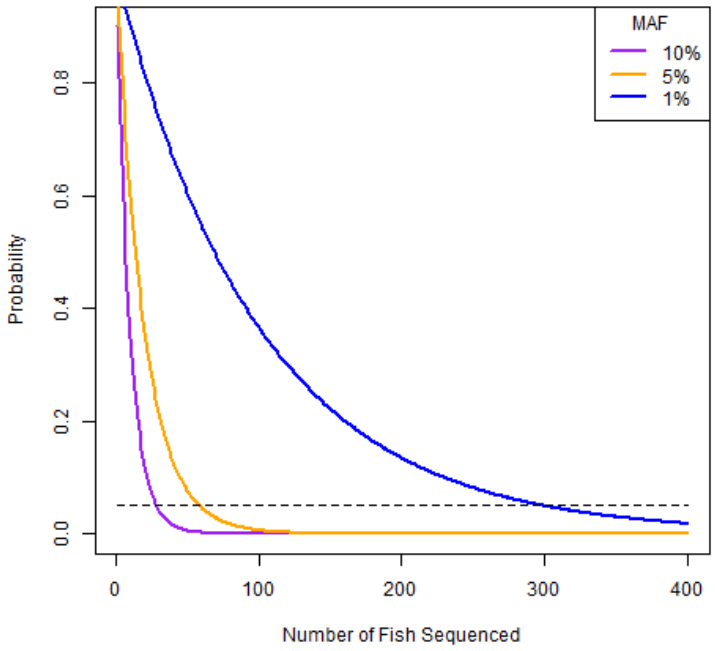

Figure 2.1. Curves displaying the probability of not observing a variant with a minor allele frequency (MAF) of 10%, 5%, and 1%. The dotted line is at 1-0.05 = "95% probability that an allele at a given frequency would be observed".

Next, we proceeded with power calculations for a genetic association study. Two sets of effect scenarios provide examples used for the p-value calculations, used to inform the choice of sample size for the study. For both sets (Table 2.1), the scenarios were created based on the population being in Hardy-Weinberg equilibrium (HWE):

- AA count: $n \times p^2$
- Aa count: $n \times 2pq$
- aa count: $n \times q^2$

where n is the sample size per concentration, p=P(A), and q=1-p=P(a).

Table 2.1. Guide tables used to create contingency tables for sample sizes n = 32, 64, ..., 224. Here, n = the number of fish at a given concentration. The Truong et al. (2014) design of n = 32 was used as a baseline.

P(A) = 0.71 [recessive], Perfect				
	AA	Aa	aa	
Affected	0.5n	0	0	0.5n
Unaffected	0	0.41n	0.09n	0.5n
	$p^2n = 0.5n$	$2pqn = 0.41n$	$q^2n = 0.09n$	n
P(A) = 0.71 [recessive], High Effect				
	AA	Aa	aa	
Affected	0.4n	0.08n	0.02n	0.5n
Unaffected	0.1n	0.33n	0.07n	0.5n
	$p^2n = 0.5n$	$2pqn = 0.41n$	$q^2n = 0.09n$	n
P(A) = 0.71 [recessive], Mid Effect				
	AA	Aa	aa	
Affected	0.35n	0.13n	0.03n	0.51n
Unaffected	0.15n	0.28n	0.06n	0.49n
	$p^2n = 0.5n$	$2pqn = 0.41n$	$q^2n = 0.09n$	n
P(A) = 0.71 [recessive], Low Effect				
	AA	Aa	aa	
Affected	0.3n	0.16n	0.04n	0.5n
Unaffected	0.2n	0.25n	0.05n	0.5n
	$p^2n = 0.5n$	$2pqn = 0.41n$	$q^2n = 0.09n$	n
P(A) = 0.5 [additive], Perfect				
	AA	Aa	aa	
Affected	0.25n	0.25n	0	0.5n
Unaffected	0	0.25n	0.25n	0.5n
	$p^2n = 0.25n$	$2pqn = 0.5n$	$q^2n = 0.25n$	n
P(A) = 0.5 [additive], High Effect				
	AA	Aa	aa	
Affected	0.2n	0.25n	0.05n	0.5n
Unaffected	0.05n	0.25n	0.2n	0.5n
	$p^2n = 0.25n$	$2pqn = 0.5n$	$q^2n = 0.25n$	n
P(A) = 0.5 [additive], Mid Effect				
	AA	Aa	aa	
Affected	0.175n	0.25n	0.075n	0.5n
Unaffected	0.075n	0.25n	0.175n	0.5n
	$p^2n = 0.25n$	$2pqn = 0.5n$	$q^2n = 0.25n$	n
P(A) = 0.5 [additive], Low Effect				
	AA	Aa	aa	
Affected	0.15n	0.25n	0.1n	0.5n
Unaffected	0.1n	0.25n	0.15n	0.5n
	$p^2n = 0.25n$	$2pqn = 0.5n$	$q^2n = 0.25n$	n

The first set of scenarios represents "recessive" effects. The effect sizes range from “Perfect” (where all affected individuals have the AA genotype), to “High Effect” (80% "Perfect"), “Mid Effect” (70% “Perfect”), and “Low Effect” (60% "perfect"). An effect of 50% “perfect” is no longer a recessive effect, since this scenario would have equivalent genotype counts for all Affected and Unaffected individuals. Therefore, Low Effect would probably not be considered a noticeable effect since it does not seem to show much genetic difference in the makeup of Affected and Unaffected individuals. Even Mid Effect may not be statistically significant without a very large sample size.

The second set of scenarios represents completely additive effects, p=0.5, with the same Perfect, High, Mid, and Low distinctions. Dominance does not need to be modeled since this would be equivalent to swapping the Affected and Unaffected row headings and performing a statistical test on an equivalent table.

P-value results were calculated for a two-tailed test using the Freeman-Halton extension of Fisher’s exact test for a 2x3 matrix (http://vassarstats.net/fisher2x3.html). The results are presented in Table 2.2 and Figure 2.2 and are discussed in the following section.

Appendix A contains information and results for 2x2 Fisher’s exact tests of allelic effect (rather than genotypes) based on summarizing these genotypic contingency tables into allelic contingency tables in case noise or sequencing information prohibit the ability to find significant effects for the proposed genotype analysis. Additionally, low coverage would lead to the use of allelic tables rather than reliance on inferred genotypes. Later in this chapter we further discuss genetic v. allelic test decisions.

2.3.2 Results

The design should provide enough power to identify perfect and high recessive effects and perfect additive effects for $n \geq 64$ (Table 2.2, Figure 2.2). Using the Bonferroni cutoff, which is more conservative than an FDR or other correction that we could implement, high additive effects are just missed at n=128.

Table 2.2. P-values achieved by the effect scenarios for n=32, 64, 96, 128, 160, 192, and 224. Red text displays where values would first become significant for each scenario if a strict Bonferroni correction were used and tests were performed for 1 variant per zebrafish gene (~26,000 zebrafish genes), creating a significance threshold of 0.05/26,000 = 1.92e-6. For the actual analysis, we can implement more sophisticated corrections that maintain detection power at nominal p-values higher than those highlighted here.

P(A)	Effect	n=32	64	96	128	160	192	224
0.71	Perfect	3.33e-9	5.46e-19	3.11e-28	4.18e-38	2.17e-47	2.77e-57	6.97e-67
	High	4.20e-4	1.35e-6	4.64e-9	1.93e-11	2.32e-14	3.32e-17	3.17e-20
	Mid	1.10e-1	1.03e-2	5.26e-4	4.52e-5	2.27e-6	1.97e-7	7.83e-8
	Low	4.37e-1	3.11e-1	1.11e-1	1.08e-1	4.12e-2	2.13e-2	1.08e-2
0.5	Perfect	9.16e-5	2.09e-9	3.56e-14	4.02e-19	9.71e-24	1.47e-28	2.63e-33
	High	1.55e-1	2.16e-3	2.17e-4	1.98e-6	2.42e-7	4.38e-8	3.34e-10
	Mid	1.55e-1	1.27e-1	1.36e-2	1.13e-2	1.63e-3	1.94e-4	1.58e-4
	Low	5.97e-1	3.92e-1	5.14e-1	3.56e-1	2.23e-1	1.15e-1	7.71e-2

We recommend using a sample size slightly larger than the one selected as most appropriate in the power analysis. This is to guard against possible inflation of power estimates by (1) not accounting for smaller sample sizes within a concentration if high mortality is encountered, (2) departures from HWE, and (3) Affected/Unaffected ratios departing from 50:50. Using a slightly higher sample size for this first study also augments confidence that we will be able to characterize genetic variability within the population (see Figure 2.1).

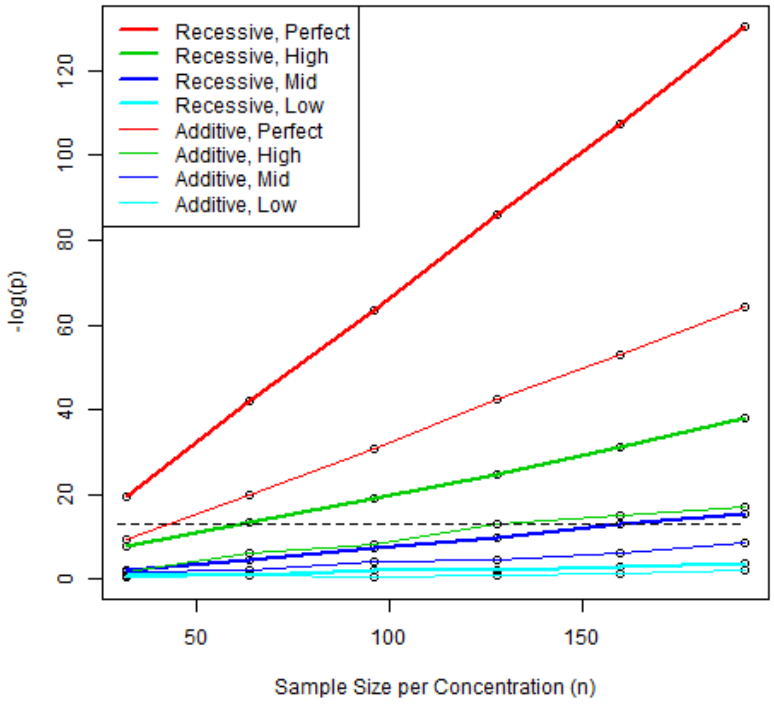

Figure 2.2. Graphical display of the power analysis. –log(p-value) is applied to better visualize the separation. The dotted line depicts the strict Bonferroni correction of a significance cutoff of 1.92e-6 (-log(1.92e-6) = 13.2). Graphical points above this line would pass this significance criterion.

2.4 Additional Considerations for the Final Design Implementation

2.4.1 Pooled v. Individual Sequencing

Most previous work in zebrafish has used pooled sequencing strategies. For the GxE GWAS analysis in the T5D population of zebrafish, there would be a number of benefits associated with pooled sequencing. These include significant cost reduction, analysis and computational simplicity (performing only two alignments, one per susceptibility group), and we would still learn more about the previously unsequenced T5D genetic architecture.

However, there were many concerns with the broad use of the data had a pooled strategy been employed. This would no longer be the first large-scale project to sequence multiple zebrafish separately, losing some of the appeal. We would have less power to detect rare variants since there will be less reads per fish, in many cases no read contributions for many of the fish within a group. This would no longer look at the individuals or allow us to

get genotypes, albeit lower confidence genotype calls (discussed in the following subsection). Without information on heterozygosity we would not be able to comment as solidly on the outbred nature of the model (though we could still show that there is genetic diversity). Even within the framework of the GxE analysis, this would limit follow-up on these individuals. If we wanted to partition the Affected and Unaffected groups based on different criteria we would not be able separate out individual information to re-organize the groupings for continued work based on the initial sample collection.

We ultimately decided that performing whole genome sequencing (WGS) on each individual separately provided the most versatile data for continued study.

2.4.2 Low Coverage

In order to combat high cost of individual sequencing, low coverage sequencing (5X) was ultimately chosen. With high coverage WGS, likelihood-based approaches can do a very good job of inferring an individual's genotype at a particular locus. For example, if all 32 reads at a site show an A allele, then the individual is determined to be homozygous AA. If there are 15 reads with an A allele at that site, and 17 reads with a C, then the individual is most likely heterozygous. The inference becomes more difficult with lower depth. If an individual is sequenced at 5X coverage, then each position should have an average of 5 reads mapped to it. Some positions may have 3, some may have 6, some may have 10, and some may even have 0. Depth of 0 or 1 provides little genotypic information for the individual at that position, but such low depths will be much less probable when higher coverage is used. Additionally, if we look at this same low- or no-depth site in multiple individuals, we can still learn about the dynamics of variability at this site within the population.

Technically we cannot know genotype information for an individual based on one read, we simply assume that an individual is homozygous for that allele (or an algorithm may assign a genotype of heterozygous if the allele is not very rare). When a genotype call is based on 2 reads it is still likely (50% chance) that the two reads came from the same copy,

potentially overestimating homozygosity once again. As the number of reads increases, the likelihood that information is coming from more than one copy increases.

Within an individual, genotype calls from low coverage sequencing will be biased toward a homozygous call. Across individuals, this would provide an overabundance of homozygous genotype calls as well. However, the allele frequencies across individuals should become more accurate. For example, if our sample had 64 AA, 32 Aa, and 4 aa individuals, maybe 20 of the heterozygotes had few reads leading to 10 of them being called AA due to only seeing an 'A' allele and 10 of them being called aa due to only observing an 'a' allele. We may end up with the following inferred distribution of genotypes: 74 AA, 12 Aa, and 14 aa individuals. The genotypes are not accurate, but the allele counts still are. Both the true and the estimated genotypes have a count of 160 'A' alleles (2*64 + 32 or 2*74 + 12) and 40 'a' alleles (2*4 + 32 or 2*14 + 12).

Due to the use of low coverage, we decided to perform allelic Fisher's Exact tests in the final study. However, if miss-called heterozygotes were not evenly distributed across susceptibility groups (Affected and Unaffected), then there could still be residual bias in allele frequencies.

2.4.3 Controls

The calculations in our initial power analysis do not take into account (1) mortality (which we see ~12%) and (2) corrections for multiple testing when more than 26,000 variants are tested. Continued perusal of genomic literature relating to zebrafish and other species leads us to believe that it may be a naive assumption that there would only be one SNP per gene, so we should scale up sample size to protect against an underpowered experiment. Controls will be used simply to inform us of any inherent biases that could come into play. In the absence of control fish that are developmentally normal at 5 dpf, we cannot be sure that there was not some kind of unforeseen batch effect invalidating our analysis of the exposed fish. The inclusion of some control fish protects us from the (slim) possibility of continuing on with biased data. Sequencing of controls may not be necessary. However,

sequencing controls could confirm that any GxE associations we find are actually due to the interaction of the gene with the environmental condition by showing that fish with the same genotype in the absence of the chemical exposure do not display the Affected phenotype.

2.5 Overarching Goals of Individual Sequencing

1. Characterize genomic variability in the outbred, T5D wild-type zebrafish population. Discover the type of variation (common SNPs vs. rare variants, etc.) observable in the population.
2. Establish the validity of the T5D population as a heterogeneous, outbred model. Isogenic models of any species fail to model the influence of genetic diversity on toxicity responses, a critical factor in human responses to toxicants. As in (French et al. 2015): "inadvertent selection of a strain with an idiosyncratic response could result in significant bias and compromise the reliability of safe exposure estimates".
3. Determine variants (SNPs, copy-number variants, etc.) associated with differential chemical responses.
4. As genomic DNA sequence information, these data will "live on".
    a. They can be resurrected for other projects, especially the genotype frequencies.
    b. This database of population genomic information can be expanded in later phases and through other projects.
    c. Changes in genotype frequencies within the population can be tracked.
    d. Align with inbred strains, compare to knockouts, etc.
    e. Unless we apply genotoxic compounds, the sequence information from [0] and any chemical-treated samples are equally useful for characterizing genetic structure.
5. The data can inform later studies aiming to associate GxE with behavior.

6. If sufficient evidence for single-gene association is not found, the collected data would still be valid for combination with data collected in a larger study (e.g. perhaps a study powered to test multigene, or G^n, hypotheses).

2.6 References

Asharani P V., Lianwu Y, Gong Z, Valiyaveettil S. 2015. Comparison of the toxicity of silver, gold and platinum nanoparticles in developing zebrafish embryos. Nanotoxicology; doi:10.3109/17435390.2010.489207.

Asharani P V, Lian Wu Y, Gong Z, Valiyaveettil S. 2008. Toxicity of silver nanoparticles in zebrafish models. Nanotechnology 19:255102; doi:10.1088/0957-4484/19/25/255102.

Bai W, Zhang Z, Tian W, He X, Ma Y, Zhao Y, et al. 2009. Toxicity of zinc oxide nanoparticles to zebrafish embryo: a physicochemical study of toxicity mechanism. J. Nanoparticle Res. 12:1645–1654; doi:10.1007/s11051-009-9740-9.

Butler MG, Iben JR, Marsden KC, Epstein J a., Granato M, Weinstein BM. 2015. SNPfisher: tools for probing genetic variation in laboratory-reared zebrafish. Development 142:1542–1552; doi:10.1242/dev.118786.

French JE, Gatti DM, Morgan DL, Kissling GE, Shockley KR, Knudsen GA, et al. 2015. Diversity outbred mice identify population-based exposure thresholds and genetic factors that infuence benzene-induced genotoxicity. Environ. Health Perspect. 123:237–245; doi:10.1289/ehp.1408202.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503; doi:10.1038/nature12111.

Hunter DJ. 2005. Gene–environment interactions in human diseases. Nat. Rev. Genet. 6:287–298; doi:10.1038/nrg1578.

Khoury MJ, Wacholder S. 2009. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. Am. J. Epidemiol. 169:227-30–5; doi:10.1093/aje/kwn351.

LaFave MC, Varshney GK, Vemulapalli M, Mullikin JC, Burgess SM. 2014. A Defined Zebrafish Line for High-Throughput Genetics and Genomics: NHGRI-1. Genetics 198:167–170; doi:10.1534/genetics.114.166769.

Lange M, Neuzeret F, Fabreges B, Froc C, Bedu S, Bally-Cuif L, et al. 2013. Inter-Individual and Inter-Strain Variations in Zebrafish Locomotor Ontogeny. PLoS One 8; doi:10.1371/journal.pone.0070172.

Marceau R, Lu W, Holloway S, Sale MM, Worrall BB, Williams SR, et al. 2015. A Fast Multiple-Kernel Method With Applications to Detect Gene-Environment Interaction. Genet. Epidemiol. 39:456–68; doi:10.1002/gepi.21909.

Murcray CE, Lewinger JP, Gauderman WJ. 2009. Gene-environment interaction in genome-wide association studies. Am. J. Epidemiol. 169:219–26; doi:10.1093/aje/kwn353.

Obholzer N, Swinburne I a., Schwab E, Nechiporuk a. V., Nicolson T, Megason SG. 2012. Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping

using whole-genome sequencing. Development 139:4280–4290; doi:10.1242/dev.083931.

Patowary A, Purkanti R, Singh M, Chauhan R, Singh AR, Swarnkar M, et al. 2013. A sequence-based variation map of zebrafish. Zebrafish 10:15–20; doi:10.1089/zeb.2012.0848.

Rennekamp AJ, Peterson RT. 2015. 15 years of zebrafish chemical screening. Curr. Opin. Chem. Biol. 24:58–70; doi:10.1016/j.cbpa.2014.10.025.

Smith PG, Day NE. 1984. The design of case-control studies: the influence of confounding and interaction effects. Int. J. Epidemiol. 13:356–65; doi:10.1093/IJE/13.3.356.

Thomas D, Stram D, Dwyer J. 1993. EXPOSURE MEASUREMENT ERROR: Influence on Expo sure-Disease Relationships and Methods of Correction. Annu. Rev. Pub!. Heal. 14: 69–93.

Truong L, Reif DM, Mary LS, Geier MC, Truong HD, Tanguay RL. 2014. Multidimensional in vivo hazard assessment using zebrafish. Toxicol. Sci. 137:212–233; doi:10.1093/toxsci/kft235.

Tzeng J-Y, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, et al. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am. J. Hum. Genet. 89:277–88; doi:10.1016/j.ajhg.2011.07.007.

Usenko CY, Harper SL, Tanguay RL. 2007. In vivo evaluation of carbon fullerene toxicity using embryonic zebrafish. Carbon N. Y. 45:1891–1898; doi:10.1016/j.carbon.2007.04.021.

Zhang R, Chu M, Zhao Y, Wu C, Guo H, Shi Y, et al. 2014. A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. Carcinogenesis 35:1528–35; doi:10.1093/carcin/bgu076.

Zhao G, Marceau R, Zhang D, Tzeng J-Y. 2015. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. Genetics 199:695–710; doi:10.1534/genetics.114.171686.

CHAPTER 3

Elucidating Gene-by-Environment (GxE) Interactions Associated with Differential Susceptibility to Chemical Exposure

Chapter 1 discusses computational methods for assessing individual and integrated "omic" data in a Systems Biology framework for environmental and toxicological science. Chapter 2 provides brief background on the specific topic of gene-environment interactions (GxE), as well as the major considerations for mining big data to design a GxE study in a specific population of zebrafish. Chapter 3 discusses the study design, implementation, and results in detail. Chapter 4 looks deeper into population variability that this analysis has brought to the fore. Chapter 5 addresses the impacts of experimental design choices, as well as future directions for the work.

This chapter contains an article with minor formatting modifications that has been submitted to a peer reviewed journal:

Balik-Meisner M, Truong L, Scholl EH, La Du JK, Tanguay RL, Reif DM. Elucidating Gene-by-Environment (GxE) Interactions Associated with Differential Susceptibility to Chemical Exposure, (*submitted*).

ABSTRACT

BACKGROUND: Modern societies are exposed to vast numbers of potentially hazardous chemicals. Despite demonstrated linkages between chemical exposure and severe health effects, there are limited, often conflicting, data on how adverse health effects of exposure differ across individuals.

OBJECTIVES: We tested the hypothesis that population variability in response to certain chemicals could elucidate a role for gene-environment interactions (GxE) in differential susceptibility.

METHODS: High throughput screening (HTS) data on thousands of chemicals in genetically-heterogeneous zebrafish were leveraged to identify a candidate chemical (Abamectin) with response patterns indicative of population susceptibility differences. We tested the prediction by generating genome-wide sequence data for 276 individual zebrafish displaying susceptible ('Affected') versus resistant ('Unaffected') phenotypes following identical chemical exposure.

RESULTS: We found GxE associated with differential susceptibility in the *sox7* promoter region, then confirmed gene expression differences between phenotypic response classes.

CONCLUSIONS: The results demonstrate that GxE associated with naturally-occurring, population genetic variation play a significant role in mediating individual responses to chemical exposure.

3.1 Introduction

Little is known about why the effects of environmental agents, including industrial chemicals, manufacturing byproducts, metals, pesticides, and herbicides differ between individuals within and across populations (Abdo et al. 2015; Betts and Shelton-Davenport 2016; Blaser et al. 2013). There is strong evidence that gene-environment interactions (GxE) play an important role in health outcomes, and that these interactions are likely a major source of the heterogeneity in chemical response. It is well established that GxE are an important component of the etiology of complex traits and diseases (Hunter 2005). Pharmacogenomic studies have consistently demonstrated that differential susceptibility to chemical exposure is directly related to genetic variation (Johnson 2003; Motsinger-Reif et al. 2013). Thus, for any genetically diverse population, latent genetic variation may contribute to observed differential susceptibility when challenged by chemical exposure. While it is accepted that such GxE interactions are important, there remain significant

challenges—both experimental and statistical—in detecting and characterizing such interactions (Rappaport and Smith 2010; Zeise et al. 2013).

The zebrafish (*D. rerio*), with over 26,000 protein-coding genes consisting of orthologues for over 70% of human genes, has gained momentum as a model organism in vertebrate genomics (Howe et al. 2013; Lieschke and Currie 2007). Toxicological and pharmacological applications in chemical biology have also seized upon the many benefits of zebrafish, including the short generation time, well-characterized development, and early maturation as clear embryos (Kimmel et al. 1995). Various morphological endpoints are easy to observe, and effects of multiple chemical exposures on these outcomes have been broadly studied (Asharani et al. 2015; Bai et al. 2009; Truong et al. 2014; Usenko et al. 2007). These advantages have led to an upward trend in high-throughput zebrafish chemical screens, especially toward screens of many chemicals in many fish, primarily in 96-well plates (Rennekamp and Peterson 2015). Thus, there exists potential for large-scale studies of chemical bioactivity that integrate genetic information to probe mechanisms underlying morphologic response to chemical exposures during development (Baer et al. 2014) or even across multiple generations (Knecht et al. 2017; Kovács et al. 2015).

Zebrafish populations differ from many model organisms in that the standard husbandry practices can be designed to maintain diversity (Nasiadka and Clark 2012), meaning that most laboratory populations contain an unknown level of genetic diversity (Brown et al. 2012). While this diversity is attractive in translating to questions of human and ecological health, it raises critical questions of how unmeasured interindividual genetic variation might contribute to susceptibility differences in response to chemical exposure. Uncharacterized genetic diversity can manifest as apparent “error” effects within and across laboratories (Rennekamp and Peterson 2015).

Comparisons between named strains and inter-lab populations of zebrafish have shown variability in several phenotypes, providing the rationale that constitutive genetic variation may contribute to the variability in exposure response (Lange et al. 2013). Unfortunately, partitioning this variability among genetic, environmental, and phenotypic

factors is hindered by (non)systematic differences in experimentation, statistical analysis, and most importantly, by a lack of available genetic data for the strains evaluated. Despite the small samples (1-2 individual fish or relatively small, pooled samples) used in studies aiming to characterize genetic diversity, results have shown between 5 and 15 million single nucleotide polymorphisms (SNPs) segregating in a zebrafish population, with roughly half of the variants showing evidence of population-specificity (Butler et al. 2015; LaFave et al. 2014; Obholzer et al. 2012; Patowary et al. 2013).

To address these limitations, we directly interrogated GxE interactions by individually sequencing entire genomes from a large sample of zebrafish using refined experimental and statistical methods for characterizing phenotypic responses to chemical exposure (Garcia et al. 2016; Truong et al. 2016; Zhang et al. 2016, 2017). The analytical methods were developed using years of data from high-throughput studies (HTS) of diverse chemicals in the Tanguay laboratory's Tropical 5D zebrafish line (T5D). The T5D line is an "outbred" population of heretofore unknown genetic heterogeneity that has been used to screen thousands of chemicals for adverse biological responses (Reif et al. 2016; Truong et al. 2014). We leveraged these HTS data to identify a chemical (Abamectin) that elicited biological response patterns indicative of population genetic differences, then generated genome-wide sequence data to compare individuals displaying differential susceptibility to chemical exposure.

This is the first genome-wide association study (GWAS) using individually sequenced zebrafish drawn from a diverse population. We show that differential susceptibility can be associated with naturally-occurring, population genetic variation. Our results demonstrate that GxE interactions play a role in mediating response to chemical exposure and have implications for both human health and ecological species. Furthermore, we show that when fine-scale control of experimental factors is exerted, genetic differences can be properly elucidated. Most importantly, our approach effectively shifts the paradigm of typical GxE research, where rather than interrogating the same list of usual chemical

suspects, we only ask questions of particular compounds that have strong evidentiary support for differential population susceptibility.

3.2 Methods

First, we exploited the large-scale design of a high throughput screening (HTS) system that has tested thousands of chemicals in an outbred zebrafish population of unknown genetic heterogeneity to select Abamectin as a chemical that displayed characteristic phenotypic patterns and high variability between individual responses. We characterized the correlation structure across morphological endpoints in order to describe a multivariate phenotype of altered eye, snout, jaw, pericardial edema, yolk sac edema, and axis development in zebrafish exposed to Abamectin. Second, we performed range-finding studies to narrow our estimate of the Abamectin concentration to the critical concentration at which a stable 50:50 Affected:Unaffected proportion was observed. Third, samples were exposed to the critical [0.6 µM] concentration of Abamectin. At 120 hours post-fertilization (hpf), we isolated Affected samples that displayed our multivariate phenotype and Unaffected samples that did not respond to chemical exposure. Fourth, DNA was individually isolated and sequenced from each Affected and Unaffected sample. These data were then used to identify GxE as genetic variants associated with differential response to chemical exposure.

3.2.1 Developmental Screening System and Experimental Population

Adult Tropical 5D (wildtype) zebrafish are housed at Sinnhuber Aquatic Research Laboratory (SARL) at Oregon State University. All experiments in this manuscript utilize this population, for which all generations are propagated with equal proportions of offspring contributed from a minimum of 25 small group crosses, each group containing 3 males and up to 3 females. Adult zebrafish were group spawned to produce embryos that were exposed to varying concentrations of Abamectin (CASRN 71751-41-2). Embryos were dechorionated and placed into individual wells of a 96-well plate. Exposures were initiated at 6 hpf, and

evaluated at 24 and 120 hpf for 22 morbidity and mortality endpoints (Truong et al. 2011). For the initial study, the Zymo Quick-DNA 96-Kit (Cat # D3011) was used for all 8 plates. The protocol was followed according to the manufacturer and DNA was eluted in water. Selected individuals from both Affected and Unaffected groups were submitted for library preparation and sequencing at Oregon State University Center of Genome Research and Biocomputing using the Illumina HiSeq3000. Individuals were multiplexed to result in a 5X coverage (per sample). For the validation studies, a secondary exposure was performed, 28 embryos (per group) from wells of interest were independently snap frozen using liquid nitrogen, then co-purified for RNA and DNA using Zymo ZF-Duet MiniPrep (Cat #D7001). mRNA gene expression analysis was performed as described in Chlebowski et al (2017).

3.2.2 Methods for Chemical Determination

Data from Truong et al. (2014) were evaluated to prioritize chemicals for GWAS mapping. Chemical choice was based on morphological data from a large-scale, concentration-response study of chemically exposed zebrafish embryos assessed for developmental toxicity endpoints at 120 hpf. In searching for patterns indicative of population variability over a broad, multipoint concentration series, our prioritization metric highlights maximal population variability in response across all binary morphological endpoints to identical environmental exposures. Our aim was to identify the chemical with maximal evidence of differential response for optimal GxE power. The full concentration-response data for 1,060 chemicals (n=32 samples tested across each of 5 chemical concentrations) were analyzed. We first produced a sub-list of chemicals that had at least one morphological endpoint at 120 hpf (other than mortality) that had near 50% incidence (32-68%) at a minimum of two concentrations. This measure ensured that there was developmental variability between fish exposed to that chemical (i.e. high proportions of both Affected and Unaffected individuals), that a stable EC_{50} could be estimated, and that the response variability held for more than one concentration. Given our goal of maximizing interindividual response variability, the latter principle ensured a less steep concentration-

response curve and a higher probability that reproducible patterns of differential population response to exposure would be observed in subsequent trials. Figure 3.1A displays a curve indicative of the pattern of variability in red, among an array of concentration-response curves representing common patterns not indicative of the specific inter-individual variability chosen for this approach.

3.2.3 Study Design

The first stage of the study involved pinpointing a critical concentration of Abamectin that induced 50% incidence of effect so that our final association study could draw evenly from Affected and Unaffected individuals to increase power to detect associated variants. The 50% incidence goal was derived from power calculations that estimated effects for an unbiased genetic association model.

Two rounds of range-finding experiments were performed (Figure 3.1B), following the developmental morphology assessment protocols detailed in previous studies (Reif et al. 2016; Truong et al. 2014; Zhang et al. 2016). First, 576 individual zebrafish embryos were distributed into six 96-well plates, with 16 individuals per plate exposed to a concentration (0, 0.03, 0.1, 0.3, 1, 3, 5, or 10 µM) of 20 mM stock of Abamectin digitally dispensed using a HP D300e. Second, a narrower range including 0, 0.5, 0.7, and 0.9 µM was performed using 192 individuals (48 per concentration). From these data, the final critical concentration of 0.6 µM was selected to approximate a 50% effect size while ensuring sufficient numbers of completely “clean” (i.e. absence of adverse developmental endpoints) Unaffected samples.

After the dose was established, individual zebrafish were exposed to the critical concentration of 0.6 µM of Abamectin, to identify individuals responding differently to equivalent environmental concentrations (Figure 3.1C). We also included untreated (negative) controls on each plate (Exposed: Control ratio @ 72:24 per plate) to ensure that we could detect global plate effects and have confirmatory samples to sequence if unexpected genotype distributions had been encountered. Importantly, the authors note that the test compound, Abamectin, is not a genotoxic compound that would have been expected

to alter DNA sequence (Oliveira et al. 2016). The total number of samples exposed was 786 (8 plates with 96 individuals), including 576 Exposed:192 Control. Of the exposed embryos, 155 individuals were Affected, exhibiting a consistent phenotype of altered eye, snout, jaw, pericardial edema, yolk sac edema, and axis development. 200 individuals were Unaffected, exhibiting no morphological defects following exposure. From these samples, we randomly selected a total of 276 (138 Affected, 138 Unaffected) chemical-exposed zebrafish to be sequenced for genome-wide association mapping.

3.2.4 Genotyping by Sequencing

Genomic DNA was extracted from individual larvae, with a subset selected for sequencing. All library preparation and sequencing was performed at Oregon State University's Center for Genome Research and Biocomputing (http://cgrb.oregonstate.edu/core). For these samples, 350 ng of DNA was used in the library preparation. Prior to library prep, the quality and quantity was verified using a fluorometric plate reader and bioanalyzer. Samples were sheared to ~320 bp, and 100 ng was used in the Wafergen robotic DNA library prep. After the library prep, each sample was quantified to verify similar input for sequencing. The samples were sequenced on a Illumina HiSeq3000 with 12 samples per lane (~5X coverage) and 150bp paired end sequencing.

3.2.5 QC and Alignment

FastQC output indicated that reads were 151 base pairs in length. GC content for each sample was ~37%, which is consistent with the zebrafish genome (Han and Zhao 2008). All samples passed a majority of QC tests and were retained. For each sample (DNA from an individual zebrafish), reads were aligned to the Genome Reference Consortium GRCz10 (Howe et al. 2013) reference genome with Bowtie2 (Langmead and Salzberg 2012) using standard settings. The overall alignment rate was ~89% for each sample. Potential PCR duplicates were then removed using Samtools rmdup (Li et al. 2009).

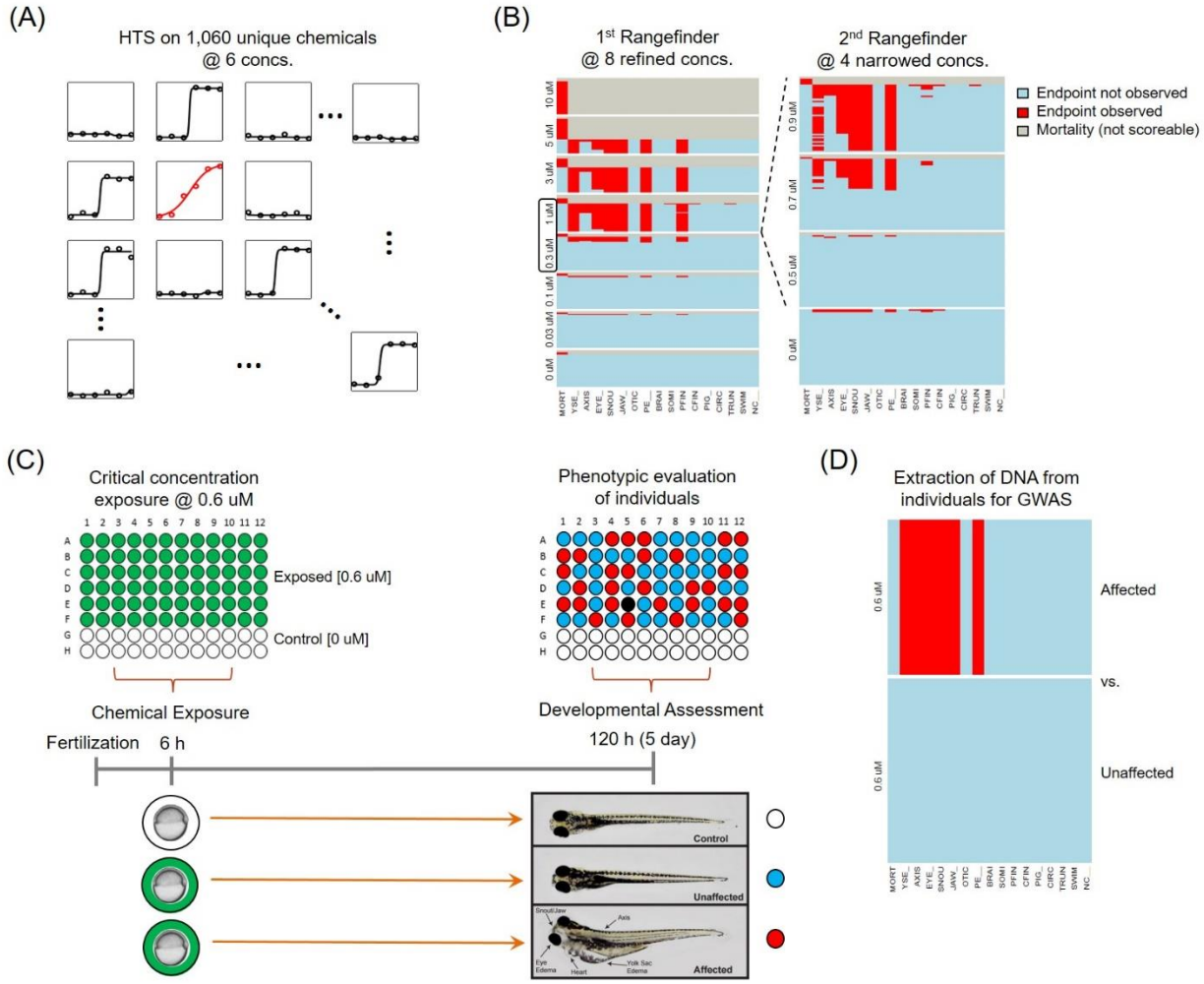

Figure 3.1. Study Design. (A) Chemical selection from HTS data: Example concentration-response curves from 1,060 chemicals interrogated for adverse morphological endpoints. Each panel represents a test chemical, where the proportion of individuals displaying adverse morphological development (vertical axis) is plotted against the tested concentrations (horizontal axis). The curve highlighted in red represents a chemical response suggestive of differential population susceptibility, whereas the black curves depict steeper toxic points-of-departure (i.e. less spread in the range of concentrations eliciting effects across the population). (B) Rangefinders: Successive screens to narrow the critical concentration as the nominal dose where 50% incidence is observed. The heatmaps show horizontal blocks (separated by whitespace) of identical concentrations, whose height corresponds to the number of zebrafish tested. Within each concentration block, each row is the vector of observed morphological endpoints (17 columns) for an individual. Blue represents no endpoint incidence, red represents incidence of an endpoint, and grey represents mortality. (C) Critical concentration exposure: Example of a single exposure plate (eight 96-well plates in total), where 72 individuals (in single wells) were exposed to 0.6 µM Abamectin at 6 hpf, plus 24 individuals exposed to vehicle (DMSO) controls. Developmental morphology screening was performed at 120 hpf to identify 'Affected' individuals (phenotype of altered eye, snout, jaw, pericardial edema, yolk sac edema, and axis development) versus 'Unaffected' individuals (no observed defects). (D) Individual DNA extraction: Individuals classified as Affected and Unaffected were selected for genomic sequencing for genome-wide association analysis.

3.2.6 Variant Calling and Filtering

Variant calls were generated for each individual at every variant site. A variant call was made at any site (across the entire genome, including all chromosomes and mitochondrial DNA, excluding nonchromosomal material or scaffolds not aligned within a chromosome) where there was sufficient evidence (based on reads, quality scores, etc.) of a non-reference base for at least one individual. GATK (Mckenna et al. 2010) HaplotypeCaller was used to call genotypes on all samples simultaneously (joint genotyping). This leverages data across samples to assign genotypes for individuals with low coverage at certain bases using a Bayesian likelihood model for genotyping. Reads with a mapping quality below 20 were not included, and a minimum phred-scaled confidence threshold of 10 was required. Genotypes are reported for every individual at every variant site for which they had any remaining reads.

Before base quality control/filtration there were 44,150,378 variant sites (36,532,474 SNPs) with an average of ~4.2X coverage per site. The GATK VariantFiltration tool was used to implement the GATK Best Practices (Depristo et al. 2011) hard filtering recommendations (filter SNPs with quality by depth (QD) < 2, phred-scaled Fisher's Exact Test p-value (FS) > 60, root mean score mapping quality (MQ) < 35, mapping quality Mann-Whitney Rank Sum < -12.5, or read position Mann-Whitney Rank Sum < -8, strand odds ratio (SOR) > 3). The MQ threshold of 35 was adjusted from GATK's recommendation of 40. This is due to the GATK workflow's use of a different aligner, which outputs a larger range of mapping quality scores for each base, averaging 60 for high confidence reads. The maximum mapping quality outputted from the Bowtie2 aligner is only 42 (for a perfectly aligned read with no mismatches to the reference). After applying the filtering cutoffs 19,973,683 SNPS remained.

Additionally, a final filtering refinement and file conversion for subsequent analysis was performed using VCFtools (Danecek et al. 2011). SNPs with mean depth (across samples) below 2 were excluded. This was done so that SNPs that were based on too few

reads per individual would not come up as potential false positive calls. 19,597,672SNPs remained.

3.2.7 Association Analysis

An allelic association analysis (Fisher's exact tests) was run for each SNP across the genome using PLINK 1.9 (Purcell et al. 2007). For each SNP, the counts of the minor and major allele in the cases (morphologically Affected individuals) and controls (morphologically Unaffected individuals) were used to calculate the exact hypergeometric probability of observing those four counts under the null hypothesis that allele counts in cases and control do not differ. A p-value below α=0.05 provides sufficient evidence that the allele counts for each group do statistically differ. To adjust for multiple testing, a Bonferroni correction was used, leading to a significant call for any SNP with a p-value below the cutoff of 2.55e-9 (0.05/19,597,672).

3.2.8 Validation

The design (Figure 3.1C) was repeated for 4 plates. At 120 hpf, 84 individuals (28 each for control, Affected, and Unaffected) were selected for RT-PCR gene expression analysis at the gene nearest one Bonferroni significant SNP. A list of the forward and reverse primers used for the experiment can be found in Appendix B Table B.3. The protocol used was described earlier, and in more detail in Chlebowski et al (2017).

Expression of target gene *sox7* and housekeeping gene *beta actin* were measured as threshold cycle (C_T) value using relative standard curves to optimize how much input was in the reaction. One control individual was discarded for an abnormally high C_T value ($C_T > 32$, for other samples $27 < C_T < 30$). A 2-sample t-test was first conducted to ensure that *beta actin* was not differentially expressed in control individuals compared to exposed (Affected and Unaffected) individuals ($p = 0.58$). The statistical analysis of *sox7* expression was then based on the $\log_2$(fold change) calculated for each individual using the $\Delta\Delta C_T$ method:

$$\Delta C_T(i) = \ sox7\ C_T(i) - beta\ actin\ C_T(i)$$

$$\Delta\Delta C_T(i) = \ \Delta C_T(i) - \sum_{j=1}^{c} \frac{\Delta C_T(j)}{c}$$

$$Fold\ Change(i) = \ 2^{-\Delta\Delta C_T(i)}$$

Individual *i* is one of the total number of control, Affected, and Unaffected individuals, and *c* is the number of control individuals. The fold change for each individual is with respect to the average control. A 2-sample t-test was then conducted comparing $\log_2$(fold change) for Affected (n=26) versus Unaffected (n=24) individuals.

3.3 Results

The experimental strategy is outlined in Figure 3.1. The following sections detail results of exploiting the large-scale HTS studies to identify a chemical with empirical evidence of population susceptibility differences (Figure 3.1A); performing additional rangefinder experiments to titrate the exact, critical concentration of test chemical that elicited maximal population variability (Figure 3.1B); evaluating developmental consequences of chemical exposure at the critical concentration of Abamectin (Figure 3.1C); selecting exposed individuals displaying an 'Affected' versus 'Unaffected' phenotype for whole-genome sequencing (Figure 3.1D); conducting a GWAS for SNPs associated with GxE (Figure 3.2); and finally validating GWAS results in targeted follow-up experiments (Figure 3.3).

3.3.1 Response Patterns Indicative of Differential Susceptibility

A chemical screen for 1,060 chemicals with full concentration-response data for all 1,060 chemicals (n=32 samples tested across each of 5 chemical concentrations plus a vehicle control) was analyzed with the goal of identifying chemicals with maximal evidence of differential response across the population. We developed a heuristic that required a chemical to show an appreciable incidence rate (32-68%) in morphological endpoints across

a minimum of two concentrations. Given the broad concentration spacing (log_{10}) of the HTS design, this measure highlighted chemicals eliciting adverse responses in an appreciable proportion of the population at concentrations several orders of magnitude lower than Unaffected individuals, despite higher concentrations of that same chemical. There were 19 out of 1,060 chemicals (Appendix B Table B.1) that passed this heuristic. From this empirical short-list, Abamectin had the most robustly-variable response (i.e. highest proportion of responders) across the most concentrations. Abamectin, a standard compound formulation of avermectin B1a and B1b and member of the structurally complex "mectin" class of compounds, is used to control insects in agriculture by acting through glutamate-gated chloride channels (GABA receptor) and as an anthelmintic agent to treat common intestinal worms (William C. Campbell 1989). Abamectin has evidence of population variability in response to pharmaceutical applications (Almehmadi and Aljedani 2016; Churcher et al. 2009; Khaldoun-Oularbi et al. 2013; Slimko et al. 2002).

3.3.2 Rangefinder Experiments to Pinpoint a Critical Concentration

Figure 3.1B illustrates the progression of rangefinder experiments aimed at identifying the critical concentration for morphological effects induced by Abamectin. Power estimates showed that, in the absence of prior knowledge of allele frequencies or population genetic structure, the optimal study design should include a balanced phenotypic ratio of Affected:Unaffected samples. Therefore, rangefinders aimed to find the concentration eliciting an even ratio of our complete phenotype (see Identifying Individuals for Genomic Sequencing). The 1st round narrowed the variable-response concentrations of the original HTS data to a range between 0.3 - 1 µM. The 2nd round tightened the target range between 0.5 - 0.9 µM. From these data, 0.6 µM was chosen as the critical concentration to balance the high incidence at 0.7 µM with the lower incidence at 0.5 µM. Overall, the rangefinder experiments showed that the fine-scale control of chemical delivery afforded by our digital chemical dispensing (Truong et al. 2016) system can reproduce population incidence rates in a consistent, dose-dependent manner.

3.3.3 Identifying Individuals for Genomic Sequencing

Using knowledge of endpoint-endpoint relationships and developmental cascades (Zhang et al. 2016, 2017), we defined a "clean" multivariate phenotype of individuals Affected by exposure to 0.6 µM Abamectin versus individuals Unaffected by the same exposure. For Affected status, an individual had to display all of the following specific endpoints: altered eye, snout, jaw, and axis development, plus pericardial and yolk sac edema (From Figure 3.1D, see columns labeled EYE, SNOU, JAW, AXIS, PE, YSE). The Unaffected phenotypic status was applied to individual embryos having absence of any morphological defect (i.e. normally-developed embryos).

Of the 576 fish exposed to Abamectin at the critical concentration, 155 (28% of surviving exposed fish) displayed the fully-penetrant Affected phenotype, 200 fish (36% of surviving exposed fish) were Unaffected, and the remainder of surviving individuals were scored as having an intermediate phenotype that consisted of a subset of the full Affected endpoint set. This is evidence of population variability in response to chemical exposure. From the fish at the two phenotypic extremes, 276 individuals were randomly chosen for full genome sequencing at 5X coverage (138 Affected and 138 Unaffected). By first focusing on a "clean" Affected group of fish that scored exactly the same on the morphologic measures, we reduced potential variability coming from sources other than genetics and thereby increased the power to detect association. Only 1% of unexposed, control individuals showed any specific morphological deformity (none of whom showed the specific phenotype of interest), versus 65% of exposed individuals, so we concluded that developmental endpoints in the exposed group were due to the exposure to 0.6 µM Abamectin ($p < 10^{-16}$).

3.3.4 Genetic Polymorphisms Associated with GxE

Details regarding sequencing of DNA extracted from individual zebrafish, variant filtering, and single-nucleotide polymorphism (SNP) calling are given in Methods. In summary, 19.6 million SNPs were retained for association analysis**.** An association analysis was conducted for each SNP across the genome using Fisher's Exact Test in PLINK 1.9

(Purcell et al. 2007) to assess allele counts in the Affected versus Unaffected individuals. To adjust for multiple testing, we subjected our nominal p-value (α=0.05) to a Bonferroni correction of (0.05/19,597,672), yielding a genome-wide significance threshold of $p < 2.55 \times 10^{-9}$. The strict statistical significance criteria highlighted 3 SNPs that exceeded the genome-wide significance thresholds (Figure 3.2 and Appendix B Table B.1). The significant SNPs were mapped to genic regions of *sox7*, *erf*, and *cfap74*.

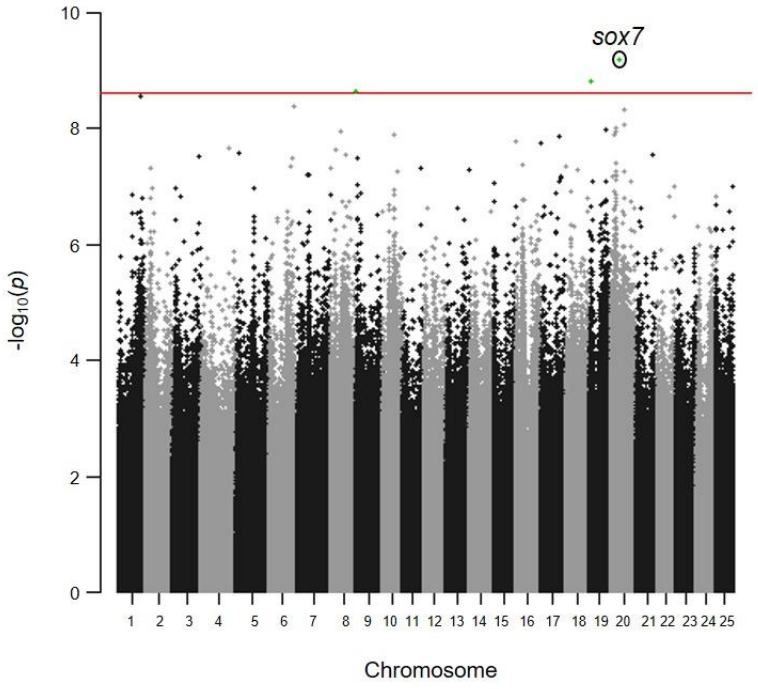

Figure 3.2. Genome Wide Association Study (GWAS) results for Abamectin. The Manhattan plot shows the genomic coordinate for each SNP on the horizontal axis (grouped into chromosomes) versus the strength of its association with phenotypic status on the vertical axis (as the negative logarithm of p-value). The horizontal red line indicates the Bonferroni-adjusted significance threshold. Green dots above this red line indicate candidate SNPs for validation as genetic factors associated with differential susceptibility (i.e. Affected versus Unaffected phenotypes) to Abamectin exposure.

3.3.5 Validation

From the genome-wide data, a G→T variant in the promoter region (569 bp upstream) of *sox7* (the top hit in Figure 3.2) was the most significantly associated with severe developmental endpoints after exposure and observed at high prevalence (> 25%) in the

population tested, highlighting a promising target for functional validation. The SNP region upstream of *sox7* (displaying increased prevalence of the minor base 'T' in the Affected group) and expression primer design for the validation study are highlighted in Figure 3.3A and Appendix B Table B.2. In the validation study, *sox7* showed significantly decreased expression ($p = 0.02$, Figure 3.3B) at 120 hpf for Affected ($n = 26$) versus Unaffected individuals ($n = 24$) after exposure to 0.6 µM Abamectin.

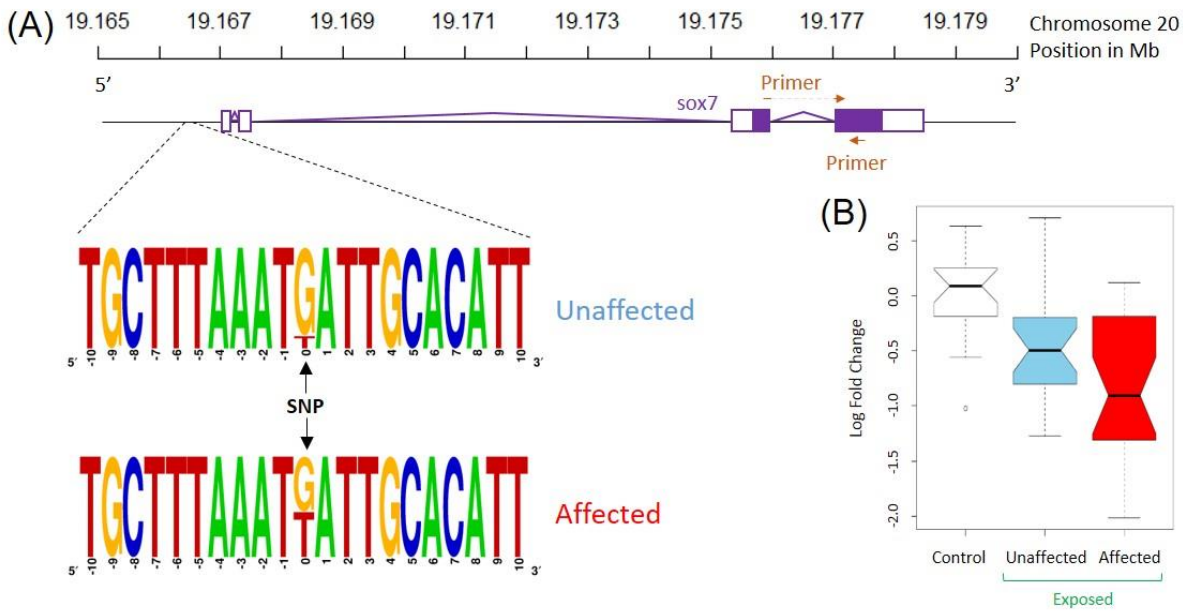

Figure 3.3. Functional Validation of *sox7*. (A) Depiction of *sox7* transcript, gene expression primer locations, and frequency sequence logos for the region surrounding the significant SNP (20:19,166,444) in Affected and Unaffected individuals from the GWAS. Sequence logos are centered at SNP site, denoted 0. Letter size corresponds to frequency of the base at that position. (B) Notched boxplot of $\log_2$(Fold Change) of *sox7* expression by affected status.

3.4 Discussion

Our evidence showed that interindividual (i.e. population) genetic variation contributes to differential response to environmental chemical exposures. To reach this conclusion, we first exploited the large-scale, systematic design of HTS data to select a target chemical (Abamectin) whose exposure produced patterns of maximally-variant response in the exposed zebrafish population. Next, we generated genome-wide sequence data for

individual zebrafish displaying susceptible versus resistant phenotypes following identical chemical exposure. Finally, we identified a genetic region near *sox7* associated with this GxE effect and confirmed gene expression differences between susceptibility groups.

This approach addresses a critical need in the face of an expanding chemical exposome (Wild 2005). Select individuals or entire communities may be especially susceptible to adverse health effects from chemical exposure through common consumer products, occupational hazards, environmental emergencies, or geographic location, such as Superfund sites (Brette et al. 2014; Judson et al. 2010). Models for diverse populations are needed to explore this interindividual susceptibility (French et al. 2015).

In contrast to the pooled samples commonly used for these types of experiments in zebrafish (Butler et al. 2015; Obholzer et al. 2012), we followed individuals (in single-fish wells) immediately post-fertilization through the entire environmental course, all phenotypic assessments, and generation of genetic information. Our results demonstrated reproducible population variability in a multivariate phenotype that showed consistent dose-response to chemical treatment across successive rounds of narrowing concentration. This fine-scale quantification of phenotype and exposure environment elucidated the role of genetic variation in explaining differential susceptibility.

A novel SNP upstream of *sox7* was associated with GxE at a genome-wide significance level. This SNP was highly correlated ($r = 0.97$) with insertion/deletion variation in a highly repetitive region just upstream. There is strong evidence that this gene, a transcription factor, plays a critical role in development related to our Affected phenotype. Ablation of *Sox7* in mice leads to developmental delays, pericardial edema, and yolk sac defects (Wat et al. 2012). In zebrafish, *sox7* mutants have arterial block and pericardial edema after 72 hpf (Hermkens et al. 2015). Our functional validation experiments showed statistically-significant suppression of *sox7* gene expression in Affected individuals versus those Unaffected following chemical exposure.

3.5 Conclusions

Our approach effectively shifts the paradigm of typical GxE research by using observed patterns of differential response as an indicator of possible genetic explanations. Only then do we assay the role of genetics in response to particular compounds with strong evidentiary support for differential population susceptibility. By linking experimentation with informatics predictions, we can make more informed choices on new experimental directions and avoid unnecessary expenditures of time and money chasing effects that are unlikely to reflect constitutive genetic variation.

3.6 References

Abdo N, Xia M, Brown CC, Kosyk O, Huang R, Sakamuru S, et al. 2015. Population-Based in Vitro Hazard and Concentration–Response Assessment of Chemicals: The 1000 Genomes High-Throughput Screening Study. Environ. Health Perspect.; doi:10.1289/ehp.1408775.

Almehmadi RM, Aljedani DM. 2016. Effects of some insecticides on longevity of the foragers honey bee worker of local honey bee race Apis mellifera jemenatica. Electron. Physician 2008–5842; doi:10.19082/1843b.

Asharani P V., Lianwu Y, Gong Z, Valiyaveettil S. 2015. Comparison of the toxicity of silver, gold and platinum nanoparticles in developing zebrafish embryos. Nanotoxicology; doi:10.3109/17435390.2010.489207.

Baer CE, Ippolito DL, Hussainzada N, Lewis J a., Jackson D a., Stallings JD. 2014. Genome-wide gene expression profiling of acute metal exposures in male zebrafish. Genomics Data 2:363–365; doi:10.1016/j.gdata.2014.10.013.

Bai W, Zhang Z, Tian W, He X, Ma Y, Zhao Y, et al. 2009. Toxicity of zinc oxide nanoparticles to zebrafish embryo: a physicochemical study of toxicity mechanism. J. Nanoparticle Res. 12:1645–1654; doi:10.1007/s11051-009-9740-9.

Betts K, Shelton-Davenport M. 2016. Interindividual Variability: New Ways to Study and Implications for Decision Making: Workshop in Brief. Natl. Acad. Press 1–13; doi:10.17226/23413.

Blaser M, Bork P, Fraser C, Knight R, Wang J. 2013. The microbiome explored: recent insights and future challenges. Nat. Rev. Microbiol. 11:213–217; doi:10.1038/nrmicro2973.

Brette F, Machado B, Cros C, Incardona JP, Scholz NL, Block BA. 2014. Crude Oil Impairs Cardiac Excitation-Contraction Coupling in Fish. Science 343:772–776; doi:10.1126/science.1242747.

Brown KH, Dobrinski KP, Lee a. S, Gokcumen O, Mills RE, Shi X, et al. 2012. Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. Proc. Natl. Acad. Sci. 109:529–534; doi:10.1073/pnas.1112163109.

Butler MG, Iben JR, Marsden KC, Epstein J a., Granato M, Weinstein BM. 2015. SNPfisher: tools for probing genetic variation in laboratory-reared zebrafish. Development 142:1542–1552; doi:10.1242/dev.118786.

Chlebowski AC, Garcia GR, Du JK La, Bisson WH, Truong L, Massey Simonich SL, et al. 2017. Mechanistic Investigations Into the Developmental Toxicity of Nitrated and Heterocyclic PAHs. Toxicol. Sci. 157:246–259; doi:10.1093/toxsci/kfx035.

Churcher TS, Pion SDS, Osei-Atweneboana MY, Prichard RK, Awadzi K, Boussinesq M, et al. 2009. Identifying sub-optimal responses to ivermectin in the treatment of River Blindness. Proc. Natl. Acad. Sci. U. S. A. 106:16716–16721; doi:10.1073/pnas.0906176106.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–215810; doi:10.1093/bioinformatics/btr330.

Depristo MA, Banks E, Poplin RE, Garimella K V, Maguire JR, Hartl C, et al. 2011. A framework for variation discovery and genotyping using next- generation DNA sequencing data. Nat Genet. 43:491–498; doi:10.1038/ng.806.

French JE, Gatti DM, Morgan DL, Kissling GE, Shockley KR, Knudsen GA, et al. 2015. Diversity outbred mice identify population-based exposure thresholds and genetic factors that infuence benzene-induced genotoxicity. Environ. Health Perspect. 123:237–245; doi:10.1289/ehp.1408202.

Garcia GR, Noyes PD, Tanguay RL. 2016. Advancements in zebrafish applications for 21st century toxicology. Pharmacol. Ther. 161:11–21; doi:10.1016/j.pharmthera.2016.03.009.

Han L, Zhao Z. 2008. Comparative Analysis of CpG Islands in Four Fish Genomes. Comp. Funct. Genomics; doi:10.1155/2008/565631.

Hermkens DMA, Van Impel A, Urasaki A, Bussmann J, Duckers HJ, Schulte-Merker S. 2015. Sox7 controls arterial specification in conjunction with hey2 and efnb2 function. Development 142:1695–1704; doi:10.1242/dev.117275.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503; doi:10.1038/nature12111.

Hunter DJ. 2005. Gene–environment interactions in human diseases. Nat. Rev. Genet. 6:287–298; doi:10.1038/nrg1578.

Johnson JA. 2003. Pharmacogenetics: potential for individualized drug therapy through genetics. Trends Genet. 19:660–666; doi:10.1016/j.tig.2003.09.008.

Judson RS, Martin MT, Reif DM, Houck KA, Knudsen TB, Rotroff DM, et al. 2010. Analysis of Eight Oil Spill Dispersants Using Rapid, In Vitro Tests for Endocrine and Other Biological Activity. Env. Sci Technol 44:5979–5985; doi:10.1021/es102150z.

Khaldoun-Oularbi H, Richeval C, Djenas N, Lhermitte M, Humbert L, Baz A. 2013. Effect of sub-acute exposure to abamectin (insecticide) on liver rats (Rattus norvegicus). Anal. 25:63–70; doi:10.1051/ata/2013039.

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of embryonic development of the zebrafish. Dev. Dyn. 203:253–310; doi:10.1002/aja.1002030302.

Knecht AL, Truong L, Marvel SW, Reif DM, Garcia A, Lu C, et al. 2017. Transgenerational inheritance of neurobehavioral and physiological deficits from developmental exposure to benzo[a]pyrene in zebrafish.; doi:10.1016/j.taap.2017.05.033.

Kovács R, Csenki Z, Bakos K, Urbányi B, Horváth Á, Garaj-Vrhovac V, et al. 2015. Assessment of toxicity and genotoxicity of low doses of 5-fluorouracil in zebrafish (Danio rerio) two-generation study. Water Res. 77:201–212; doi:10.1016/j.watres.2015.03.025.

LaFave MC, Varshney GK, Vemulapalli M, Mullikin JC, Burgess SM. 2014. A Defined Zebrafish Line for High-Throughput Genetics and Genomics: NHGRI-1. Genetics 198:167–170; doi:10.1534/genetics.114.166769.

Lange M, Neuzeret F, Fabreges B, Froc C, Bedu S, Bally-Cuif L, et al. 2013. Inter-Individual and Inter-Strain Variations in Zebrafish Locomotor Ontogeny. PLoS One 8; doi:10.1371/journal.pone.0070172.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359; doi:10.1038/nmeth.1923.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Appl. NOTE 25:2078–2079; doi:10.1093/bioinformatics/btp352.

Lieschke GJ, Currie PD. 2007. Animal models of human disease: zebrafish swim into view. Nat. Rev. Genet. 8:353–367; doi:10.1038/nrg2091.

Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303; doi:10.1101/gr.107524.110.

Motsinger-Reif AA, Jorgenson E, Relling M V, Kroetz DL, Weinshilboum R, Cox NJ, et al. 2013. Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenet. Genomics 23:383–94; doi:10.1097/FPC.0b013e32833d7b45.

Nasiadka A, Clark MD. 2012. Zebrafish Breeding in the Laboratory Environment. ILAR J. 53:161–168; doi:10.1093/ilar.53.2.161.

Obholzer N, Swinburne I a., Schwab E, Nechiporuk a. V., Nicolson T, Megason SG. 2012. Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing. Development 139:4280–4290; doi:10.1242/dev.083931.

Oliveira R, Grisolia CK, Monteiro MS, Soares AMVM, Domingues I. 2016. Multilevel assessment of ivermectin effects using different zebrafish life stages. Comp. Biochem. Physiol. Part C Toxicol. Pharmacol. 187:50–61; doi:10.1016/j.cbpc.2016.04.004.

Patowary A, Purkanti R, Singh M, Chauhan R, Singh AR, Swarnkar M, et al. 2013. A sequence-based variation map of zebrafish. Zebrafish 10:15–20; doi:10.1089/zeb.2012.0848.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575; doi:10.1086/519795.

Rappaport SM, Smith MT. 2010. Environment and Disease Risks. Science 330:460–461; doi:10.1126/science.1192603.

Reif DM, Truong L, Mandrell D, Marvel S, Zhang G, Tanguay RL. 2016. High-throughput characterization of chemical-associated embryonic behavioral changes predicts teratogenic outcomes. Arch. Toxicol. 90:1459–1470; doi:10.1007/s00204-015-1554-1.

Rennekamp AJ, Peterson RT. 2015. 15 years of zebrafish chemical screening. Curr. Opin. Chem. Biol. 24:58–70; doi:10.1016/j.cbpa.2014.10.025.

Slimko EM, McKinney S, Anderson DJ, Davidson N, Lester H a. 2002. Selective electrical silencing of mammalian neurons in vitro by the use of invertebrate ligand-gated chloride channels. J. Neurosci. 22:7373–7379; doi:20026775.

Truong L, Bugel SM, Chlebowski A, Usenko CY, Simonich MT, Simonich SLM, et al. 2016. Optimizing multi-dimensional high throughput screening using zebrafish. Reprod. Toxicol. 65:139–147; doi:10.1016/j.reprotox.2016.05.015.

Truong L, Harper SL, Tanguay RL. 2011. Evaluation of Embryotoxicity Using the Zebrafish Model. In *Methods in Molecular Biology*, Vol. 691 of, pp. 271–279.

Truong L, Reif DM, Mary LS, Geier MC, Truong HD, Tanguay RL. 2014. Multidimensional in vivo hazard assessment using zebrafish. Toxicol. Sci. 137:212–233; doi:10.1093/toxsci/kft235.

Usenko CY, Harper SL, Tanguay RL. 2007. In vivo evaluation of carbon fullerene toxicity using embryonic zebrafish. Carbon N. Y. 45:1891–1898; doi:10.1016/j.carbon.2007.04.021.

Wat MJ, Beck TF, Hernández-García A, Yu Z, Veenma D, Garcia M, et al. 2012. Mouse model reveals the role of SOX7 in the development of congenital diaphragmatic hernia associated with recurrent deletions of 8p23.1. Hum. Mol. Genet. 21:4115–25; doi:10.1093/hmg/dds241.

Wild CP. 2005. Complementing the Genome with an Exposome'': The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiol Biomarkers Prev 14:1847–1850; doi:10.1158/1055-9965.EPI-05-0456.

William C. Campbell, ed. 1989. *Ivermectin and Abamectin*. Springer-Verlag New York, New York, NY.

Zeise L, Bois FY, Chiu WA, Hattis D, Rusyn I, Guyton KZ. 2013. Addressing Human Variability in Next-Generation Human Health Risk Assessments of Environmental Chemicals. Environ. Health Perspect. 121:23–31; doi:10.1289/ehp.1205687.

Zhang G, Marvel S, Truong L, Tanguay RL, Reif DM. 2016. Aggregate entropy scoring for quantifying activity across endpoints with irregular correlation structure. Reprod. Toxicol. 62:92–99; doi:10.1016/j.reprotox.2016.04.012.

Zhang G, Roell KR, Truong L, Tanguay RL, Reif DM. 2017. A data-driven weighting scheme for multivariate phenotypic endpoints recapitulates zebrafish developmental cascades. Toxicol. Appl. Pharmacol. 314:109–117; doi:10.1016/j.taap.2016.11.010.

CHAPTER 4

Population Genetic Diversity in Zebrafish Lines

Chapter 1 discusses computational methods for assessing individual and integrated "omic" data in a Systems Biology framework for environmental and toxicological science. Chapter 2 provides brief background on the specific topic of gene-environment interactions (GxE), as well as the major considerations for mining big data to design a GxE study in a specific population of zebrafish. Chapter 3 discusses the study design, implementation, and results in detail. Chapter 4 looks deeper into population variability that this analysis has brought to the fore. Chapter 5 addresses the impacts of experimental design choices, as well as future directions for the work.

This chapter contains an article with minor formatting modifications that has been submitted to a peer reviewed journal:

Balik-Meisner M, Truong L, Scholl EH, Tanguay RL, Reif DM. Population Genetic Diversity in Zebrafish Lines, (*submitted*).

ABSTRACT

Toxicological and pharmacological researchers have seized upon the many benefits of zebrafish, including the short generation time, well-characterized development, and early maturation as clear embryos. A major difference from many model organisms is that standard husbandry practices in zebrafish are designed to maintain population diversity. While this diversity is attractive for translational applications in human and ecological health, it raises critical questions on how interindividual genetic variation might contribute to chemical exposure or disease susceptibility differences. Findings from pooled samples of zebrafish support this supposition of diversity yet cannot directly measure allele frequencies for reference versus alternate alleles. Using the Tanguay lab

Tropical 5D zebrafish line (T5D), we performed whole-genome sequencing on a large group (n=276) of individual zebrafish embryos. Paired-end reads were collected on an Illumina 3000HT, then aligned to the most recent zebrafish reference genome (GRCz10). These data were used to compare observed population genetic variation across species (humans, mice, zebrafish), then across lines within zebrafish. We found more single nucleotide polymorphisms (SNPs) in T5D than have been reported in SNP databases for any of the WIK, TU, TL, or AB lines. We theorize that some subset of the novel SNPs may be shared with other zebrafish lines but have not been identified in other studies due to the limitations of capturing population diversity in pooled sequencing strategies. We establish T5D as a model that is representative of diversity levels within laboratory zebrafish lines and demonstrate that experimental design and analysis can exert major effects when characterizing genetic diversity in heterogeneous populations.

4.1 Introduction

Use of the zebrafish (*D. rerio*) as a model organism has gained momentum in vertebrate genomics (Lieschke and Currie 2007). As a vertebrate with one of the largest sets of protein-coding genes, consisting of orthologues for over 70% of human genes, they have been adapted as exposure and human disease models (Howe et al. 2013). There are many benefits to using zebrafish in developmental studies, including early maturation as clear embryos that are amenable to easily observable morphological endpoints, short generation time, and well-characterized development that is conserved across species during the phylotypic period (Kimmel et al. 1995; Irie and Kuratani 2011). These advantages have led to an upward trend in high-throughput zebrafish chemical screens, especially toward screens of many chemicals using large quantities of fish (Usenko et al. 2007; Bai et al. 2009; Truong et al. 2014; Asharani et al. 2015). Thus, this model could be used for large-scale studies of chemical bioactivity that include genetic information on response mechanisms during

development of exposed individuals (Baer et al. 2014) or even across multiple generations (Kovács et al. 2015; Knecht et al. 2017).

Model organisms have long been utilized to study genetic determinants underlying human disease susceptibility, because experiments can exert necessary controls over factors such as diet, lifestyle, and environment that would be impossible in a human setting. The mouse has been extensively used to mechanistically model human disease, but until the inception of a major recombinant inbred line (RIL) panel, the lack of variability within any single inbred strain did not sufficiently model human genetic variability (Churchill et al. 2004). The RIL strategy had been implemented multiple times in mice, but their utility was insufficiently broad due to limited genetic diversity in lines stemming from two inbred strains. In order to create a RIL panel representing the genetic diversity among a more general populace of mice, the Collaborative Cross (CC) (Chesler et al. 2008) was implemented to randomly mix the genomes of 8 founder strains to create hundreds of isogenic RILs (Churchill et al. 2004). The 8 founder strains included 5 classical inbred strains and 3 wild-derived strains that jointly capture 90% of the known allelic diversity in the mouse genome (Roberts et al. 2007). A RIL strategy aiming to capture diversity has also been used in fruit flies (*Drosophila melanogaster*) (Mackay et al. 2012). For these populations, each isogenic line has been sequenced. Individuals within one line are homogeneous, but comparisons of traits or susceptibility between lines has aided in identifying genetic associations (Cirelli et al. 2008; Unckless et al. 2015; Ivanov et al. 2015).

Nonetheless, isogenic models of any species fail to model the influence of genetic diversity on toxicity responses, a critical factor in human responses to toxicants. As noted by French et. al. “inadvertent selection of a strain with an idiosyncratic response could result in significant bias and compromise the reliability of safe exposure estimates” (2015). In order to use the CC mice in an infrastructure more similar to naturally occurring populations with heterozygosity, an outbred population was created. The Diversity Outbred (DO) population was derived from 144 CC lines at various stages (4-12 generations) of inbreeding, allowing recombination events in the early generations to promote recombination and genetic diversity

amongst the DO mice (Svenson et al. 2012). Approximately 45 M single nucleotide polymorphisms (SNPs) segregate in the CC and DO populations, four times more than in any singular laboratory mouse strain (Yang et al. 2011). Each DO individual is unique and cannot be precisely replicated, but haplotypes can be reconstructed based on the determination of recombination events using knowledge of the CC founder strain homozygous genotypes, and CC mice can be used to test hypotheses generated through use of DO mice (Churchill et al. 2012). When employed appropriately, these resources can provide insight on a number of variants that should be more in-line with that found in a wild type (WT) population.

In zebrafish, inbreeding adversely affects fecundity and survival (Mrakovcic and Haley 1979), so endeavors to create isogenic lines have not been fruitful. Zebrafish populations differ from many model organisms in that the standard husbandry practices are often designed to maintain diversity (Nasiadka and Clark 2012). Thus, like human populations, most laboratory zebrafish populations contain an unknown level of genetic diversity (Brown et al. 2012). Comparisons between named strains and inter-lab populations of zebrafish have shown variability in several phenotypes, providing the rationale that constitutive genetic variation may contribute to the variability in exposure response (Lange et al. 2013). Despite the small samples (1-2 individual fish or relatively small, pooled samples) used in studies aiming to characterize genetic diversity, results have shown between 5 and 15 million single nucleotide polymorphisms (SNPs) segregating in a zebrafish population, with roughly half of the variants showing evidence of population-specificity (Obholzer et al. 2012; Patowary et al. 2013; LaFave et al. 2014; Butler et al. 2015). It has been estimated that zebrafish populations have a larger abundance of SNPs per kb of unique sequence than ethnically defined human populations (Butler et al. 2015).

Here, we characterize salient features of population genetic architecture of the Tropical 5D (T5D) line as a representative laboratory population of zebrafish. The T5D line is an “outbred” population of heretofore unknown genetic heterogeneity that has been used to screen thousands of chemicals for adverse biological responses (Truong et al. 2014; Reif et al. 2016). We obtained whole genome sequences of 276 individuals from the T5D

population, aligned reads to the GRCz10 reference genome, called SNPs and indels, and created a T5D-specific reference genome. This was performed with the aims of characterizing genomic variability in the outbred, T5D wild-type zebrafish population, discovering the type of variation (common SNPs vs. rare variants, etc.) observable in the population, and establishing the validity of the T5D population as a heterogeneous model. We then empirically compared genomic characteristics of our zebrafish population with murine and human reference populations, as well as across other zebrafish lines. Finally, we explored whether the higher apparent diversity observed in our T5D line could be due to experimental design factors that tend to underestimate diversity in other published lines.

4.2 Materials & Methods

4.2.1 Developmental Screening System and Experimental Population

The T5D founders were originally imported into the Tanguay lab at Oregon State University from a breeding facility containing thousands of zebrafish in 2007 to generate a Pseudoloma neurophilia (Microsporidia) free line (Stanley et al. 2009). The T5D zebrafish are housed at Sinnhuber Aquatic Research Laboratory (SARL) at Oregon State University and maintained in accordance with their Institutional Animal Care and Use Committee protocols. Fish are raised in a recirculating water system with a temperature of 28° ±1 °C and a 14h light: 10h dark photoperiod. All generations are propagated with equal proportions of offspring contributed from a minimum of 25 small group crosses, each group containing 3 males and up to 3 females.

4.2.2 Genotyping by Sequencing

The sequencing data are described in detail in (Balik-Meisner et al., *submitted*). In brief, genomic DNA was extracted (Zymo Quick-DNA 96-Kit Cat # D3011) from 276 individual larvae exposed to 0.6 µM Abamectin at 120 hours post fertilization. The authors note that Abamectin is non-genotoxic (Oliveira et al. 2016), so exposure would not have

altered constitutive DNA sequence. The extraction protocol was followed according to the manufacturer and DNA was eluted in water. All library preparation and sequencing was performed at Oregon State University's Center for Genome Research and Biocomputing (http://cgrb.oregonstate.edu/core). For these samples, 350 ng of DNA was used in the library preparation. Prior to library prep, the quality and quantity was verified using a fluorometric plate reader and bioanalyzer. Samples were sheared to ~320 bp, and 100 ng was used in the Wafergen robotic DNA library prep. After the library prep, each sample was quantified to verify similar input for sequencing. The samples were sequenced on an Illumina HiSeq3000 with 12 samples per lane (~5X coverage) and 150bp paired end sequencing.

4.2.3 Alignment

FastQC output indicated that reads were 151 base pairs in length. GC content for each sample was ~37%, which is consistent with the zebrafish genome (Han and Zhao 2008). For each sample (DNA from an individual zebrafish), reads were aligned to the Genome Reference Consortium GRCz10 (Howe et al. 2013) reference genome with Bowtie2 (Langmead and Salzberg 2012) using standard settings. The overall alignment rate was ~89% for each sample. Potential PCR duplicates were then removed using Samtools rmdup (Li et al. 2009).

4.2.4 Variant Calling and Filtering

Variant calls were generated for each individual at every variant site. A variant call was made at any site (across the entire genome, including all chromosomes and mitochondrial DNA, excluding nonchromosomal material or scaffolds not aligned within a chromosome), where there was sufficient evidence (based on reads, quality scores, etc.) of a nonreference base for at least one individual. GATK (Mckenna et al. 2010) HaplotypeCaller was used to call genotypes on all samples simultaneously (joint genotyping). This leverages data across samples to assign genotypes for individuals with low coverage at certain bases using a Bayesian likelihood model for genotyping. Reads with a mapping quality below 20

were not included, and a minimum phred-scaled confidence threshold of 10 was required. Genotypes are reported for every individual at every variant site for which they had any remaining reads.

Before base quality control/filtration, there were 36,532,474 SNPs and 7,262,723 indel variants with an average of 4.2X coverage per site. The GATK VariantFiltration tool was used to implement the GATK Best Practices (Depristo et al. 2011) hard filtering recommendations for SNPs and indels (filter SNPs with quality by depth (QD) < 2, phred-scaled Fisher's Exact Test p-value (FS) > 60, root mean score mapping quality (MQ) < 35, mapping quality Mann-Whitney Rank Sum < -12.5, or read position Mann-Whitney Rank Sum < -8, strand odds ratio (SOR) > 3; filter indels with QD < 2, FS > 100, read position Mann-Whitney Rank Sum < -20, SOR > 10). The adjustment of the MQ threshold from GATK's recommendation of 40 to 35 accounted for the difference in quality score reporting between the aligner suggested by GATK (BWA) and Bowtie2. BWA outputs a larger range of mapping quality scores, averaging 60 for high confidence reads, whereas the maximum quality score for Bowtie2 is 42, indicating a perfectly aligned read. After applying the filtering cutoffs 20,385,817 SNPs and 6,304,066 indels remained.

4.2.5 Additional Species and Variant Consequence Predictions

Short genetic variation datasets for human, mouse, and zebrafish from NCBI's dbSNP were downloaded from ftp.ncbi.nih.gov/snp/organisms/. The effect of the variants on genes and transcripts and consequences on protein sequence were annotated for each species using Ensembl's Variant Effect Predictor (VEP) (McLaren et al. 2016) (Figure 4.1).

4.2.6 Variant Set Preparation for Line Comparisons

Consortial variant (CVF) files of SNP and indel variation from 4 other zebrafish lines (AB, TU, TL, WIK), compiled through integration of data from three previous studies (Obholzer et al. 2012; Bowen et al. 2012; Butler et al. 2015), were downloaded from https://snpfisher.nichd.nih.gov/snpfisher/tracks.html. Each of these studies sequenced a pool

of zebrafish between 3X-16X coverage and aligned reads as one sample to the Zv9 reference genome for each line.

To compare T5D variant sites, the positions based on the GRCz10 reference genome needed to be mapped back to equivalent locations in the Zv9 build using Picard's LiftoverVcf with the danRer10ToDanRer7 chain file from hgdownload.cse.ucsc.edu/goldenPath/danRer10/liftOver/. 20,131,988 SNPs and 5,630,544 indels were successfully mapped back to the Zv9 reference.

Additionally, the CVF files had masked variants in non-complex regions of the genome. To filter T5D variants accordingly the repeat masked annotation of Zv9 was downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/danRer7/database/rmsk.txt.gz. Approximately 51% of the genome is masked for having highly-repetitive content. As a reference, over 56% of the human genome is masked (http://www.repeatmasker.org/). Variants located in these noncomplex regions of the genome were removed from the final T5D comparison dataset resulting in 10,301,547 SNPs and 2,375,455 indels. To ensure consistency between datasets, we performed the same masking procedure on the AB, TU, TL, and WIK datasets even though masking had been previously performed. All comparisons with these lines were based on the following approximate counts (T5D: 10.3 M SNPs, 2.4 M indels; AB: 4.3 M SNPs, 0.6 M indels; TU: 3.6 M SNPs, 0.4 M indels; TL: 6.2 M SNPs, 0.8 M indels; WIK: 8.5 M SNPs, 1.1 M indels).

A VCF file for NHGRI-1 (LaFave et al. 2014) was downloaded for use in a separate line comparison due to sequencing strategy differences and alignment to different versions of the reference genome. The file included 17,089,212 variant calls (15,680,057 SNPs) and genotypes for the two founders based on high coverage individual whole genome sequencing and alignment to GRCz10 without masking. The VCF of 20,385,817 SNPs for T5D compared to the GRCz10 reference was used for SNP site comparisons to the NHGRI-1 line only.

4.2.7 Downsampling

To address the impact of sequence design on comparisons between T5D and other lines that used pooled sequencing, a portion of the T5D data was used as a simulated pool. This was performed with the intention to more closely approximate variants that would have been called in T5D had a pooled approach been employed instead of individual sequencing. First, 20 individuals were randomly selected. Next, 20% of each of their reads were randomly selected to create a simulated pooled sample at an average of 20X coverage. Alignment, variant calling, and filtering were all performed with the previous parameters. Before filtering, 18,086,779 SNPs were called. After filtering, 12,179,880 SNPs remained, of which 12,009,411 were successfully mapped to the Zv9 reference genome. For indels, the count decreased from 2,966,260 to 2,608,746 to 2,339,775. After masking variants located in noncomplex regions of the genome, the final pooled approximation T5D comparison dataset resulted in 6,175,287 SNPs and 1,080,749 indels.

4.3 Results

4.3.1 Interspecies Comparisons

The zebrafish genome (1.5 Gb) is roughly half the size of the human (3.3 Gb) or mouse (2.8 Gb) genome. To date, the total number of discovered variants in the zebrafish genome is less than half the number found in human or mouse genomes; consequently, validation is more sparse. The allele frequency distribution of “common” human variants indicates that the majority of common variants are infrequent across the overall human population (minor allele frequency (MAF) < 0.1) (Figure 4.1B). Though these SNPs are private to all save a handful of people, they are only prevalent in specific subpopulations. The majority of common variants in the human genome have already been discovered, but rare variants continue to be discovered via deep whole genome sequencing of cohorts of individuals from geographically/ethnically defined populations (Shen et al. 2013).

The proportion of the types of SNPs found in T5D were similar to those reported by the dbSNP variant sites in both human and mouse. We observed more intron variants in T5D, and synonymous gene transcript variant percentages fell between mouse and human (Figure 4.1A). The larger percentage of intronic variants in zebrafish can be explained by genetic architecture, as the value is proportional to the percent of the genome sequence that is intronic (roughly 43.9% of the zebrafish genome, 39.6% of the human genome, and 26.6% of the mouse genome) (Sakharkar et al. 2005; Moss et al. 2011).

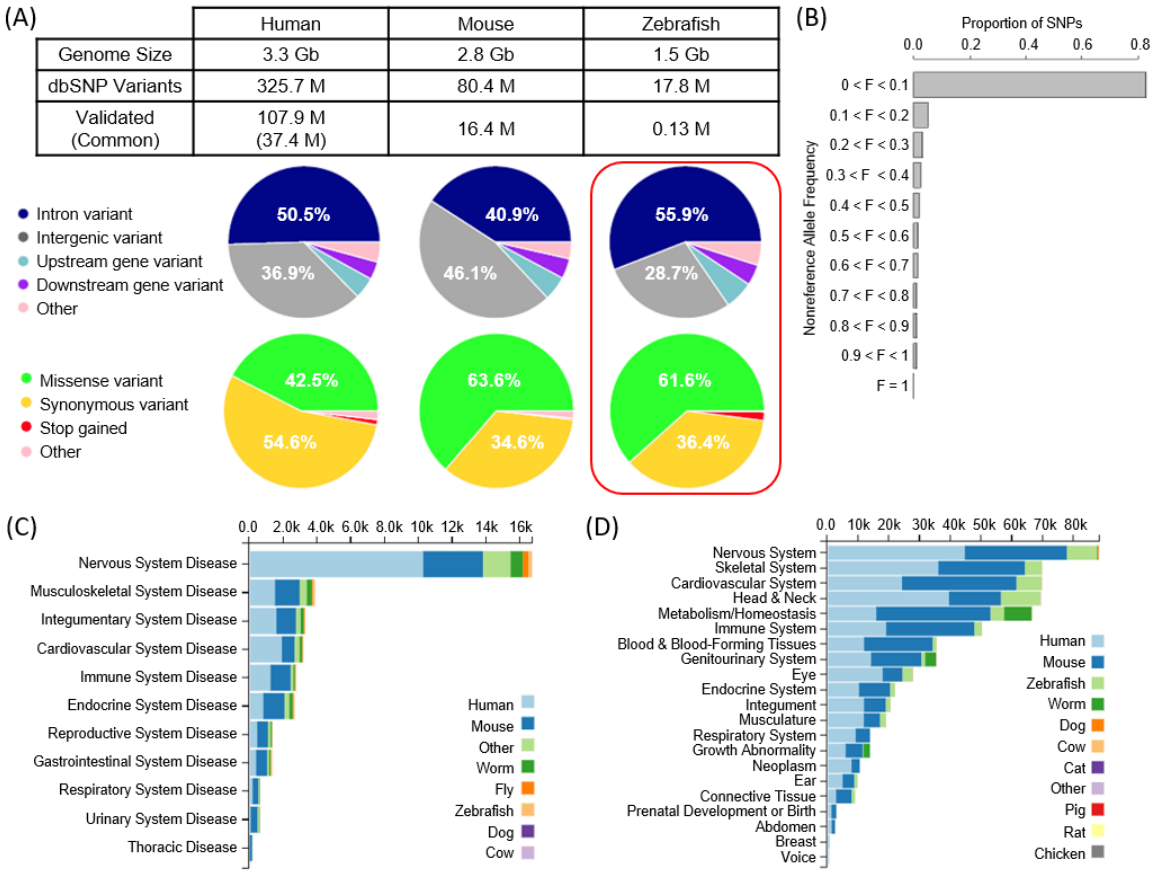

Figure 4.1. Known Variants. (A) Genome size, known variant count in dbSNP, variant effect, and consequences of transcript variants. The red box contains the variant effects for the 20.1 M SNPs found in T5D. (All other zebrafish data refers to the reference genome and publically available data.) (B) Allele frequency spectrum for common human variants. (C) Number of models per disease category stacked by organism (from monarchinitiative.org). (D) Number of phenotype-gene associations per species (from monarchinitiative.org).

The 20.1 M SNPs equate to 13.4 SNPs per 1 kb genomic sequence. Prior studies estimated that certain zebrafish strains contained an average of 7 SNPs per 1 kb of non-repetitive (i.e., non-complex, non-masked) genome sequence per strain, which is still more than in any ethnically defined human population from the 1000 Genomes (Butler et al. 2015). Estimates in other species have been similar (4.9 SNPs per kb in sheep, 5.5 SNPs per kb in chickens, 10.1 SNPs per kb in fly, 13.9 SNPs per kb in mouse), though they have been based on combined line/breed data (Ka-Shu Wong et al. 2004; Kijas et al. 2009; Kang et al. 2016; Srivastava et al. 2017).

On average, an individual in the T5D population was found to carry a nonreference allele (homozygous nonreference or heterozygous) at 6.9 M SNP sites and 1.8 M indel sites (3.7 M SNP sites and 0.84 M indel sites in non-masked genomic regions). This is more than have been identified in individual human genomes. For example, in Caucasians an average of 3.3 M SNPs and 0.49 M indels with nonreference alleles were identified per individual (Shen et al. 2013). In Turkish individuals, an average of 3.3 M SNPs and 0.91 M indels were identified (Alkan et al. 2014). In Chinese individuals, an average of 3.5 M SNPs and 0.63 M indels were identified (Shi et al. 2016). Comparing across broad populations, Cho et. al. found an average of 4.6 M SNPs and 0.68 M indels per African individual, 3.75 M SNPs and 0.60 M indels per Caucasian, and 3.69 M SNPs and 0.54 M indels per Asian. When using a Korean genome as the reference, the number of calls increased for each of the African and Caucasian individuals and decreased for the Asian individuals (Cho et al. 2016).

The abundance of sites with nonreference alleles per T5D zebrafish could imply that within a population, zebrafish are more genetically variable than humans. However, because ethnic/population-level choice of reference may influence the number of variants called (Cho et al. 2016), an individual zebrafish within the T5D population may vary more from the current zebrafish reference genome than individuals from certain human ethnic populations vary compared to the human reference genome. While this could indicate that the human reference genome provides a more representative consensus across human populations, it is

also possible that the absence of admixing between zebrafish laboratory populations may have caused them to diverge more from a historical reference sequence.

4.3.2 T5D Variants and Zebrafish Line Comparisons

The estimate of 20.1 M SNPs segregating in the population (10.3 M in non-repetitive regions of the genome) included nonreference allele frequencies from 0.1% to 99.8%. We posit that the 10.3 - 20.1 M SNPs and 2.8 - 5.6 M indels discovered in T5D are accurate bounds for an estimate of variability in this zebrafish line. With more individuals and higher coverage, we would expect to find even more rare variants segregating in the population.

With the exception of chromosome 4, the number of variants discovered per chromosome was proportional to chromosome length (Appendix C Table C.1). There was a region of chromosome 4 with drastically fewer variants in our study (Appendix C Figure C.1) that was also reported in Butler, et. al. (2015). This low-variability region lies within an area of the genome that has primarily zebrafish-specific genes not homologous to other species (Howe et al. 2013). There is evidence that chromosome 4 is likely the sex chromosome in natural zebrafish populations (Wilson et al. 2014).

T5D was found to have more variants compared to results from studies using pooled sequencing and smaller sample sizes (Figure 4.2A & 4.2C). T5D variants, discovered based on approximately 1380X coverage across individuals (5X for 276 individuals), followed an allele frequency spectrum more similar to known human variants (Figures 4.2B, 4.2D, 4.1B). Variants discovered in the other lines in pooled sequencing experiments were primarily common, because a given site had low coverage (< 20X) across the pool. Additionally, rare variants (those observed at frequencies of < 0.1) would have been missed at small sample sizes. For T5D, the plurality of the variants discovered were rare.

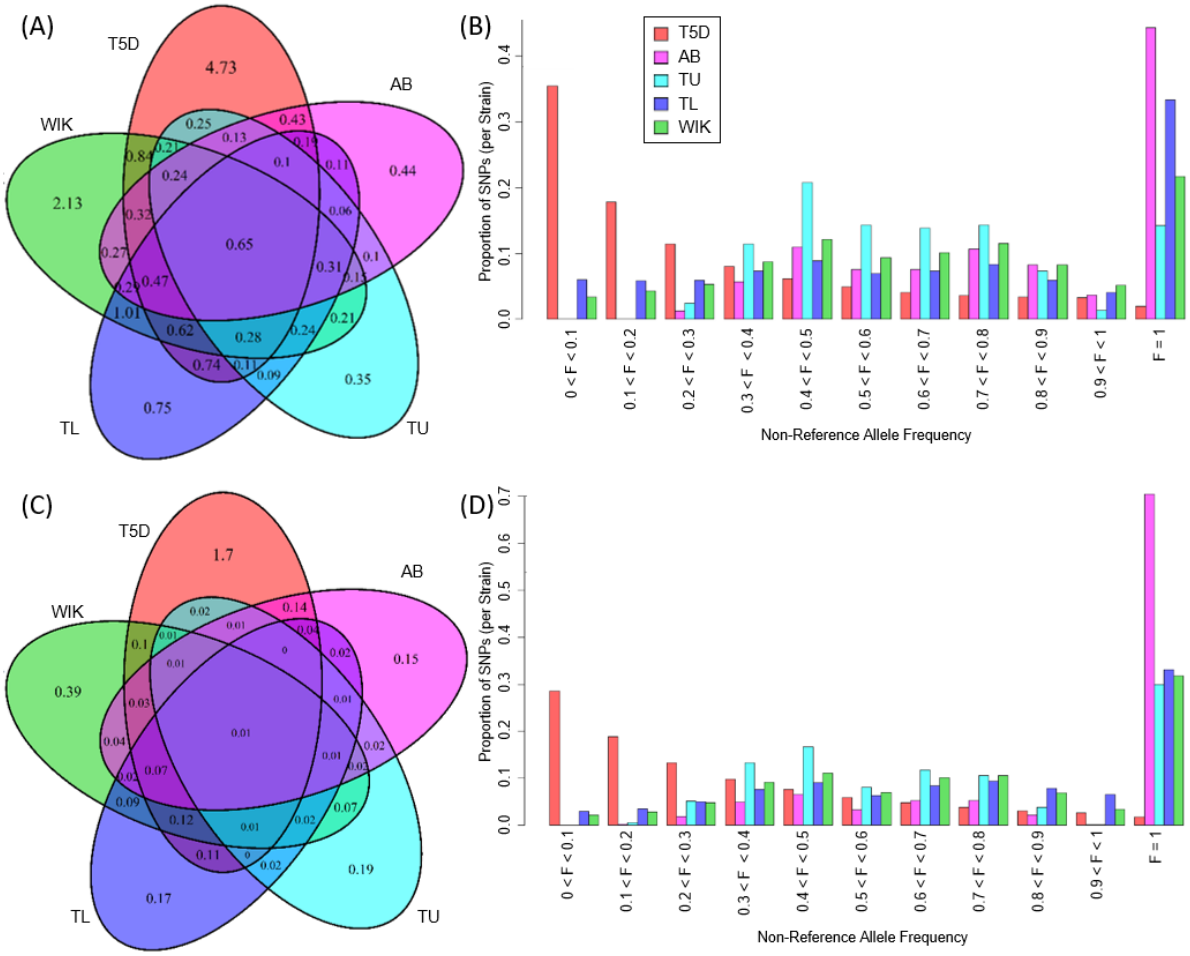

Figure 4.2. Zebrafish Variant Comparisons. (A) Venn diagram of SNP sites (in millions) compared to the Zv9 reference genome. (B) Proportions of SNPs binned by alternate allele frequencies for the 5 lines. The T5D allele frequencies are based on 276 individual whole genome sequences. For all other lines, frequencies were determined based on the proportion of reads with nonreference base calls since no individual genotypes can be determined from pooled sequence alignment. (C) Venn diagram of indel sites (in millions). (D) Proportion of indels for discrete alternate allele frequencies.

The comparator lines displayed an abundance of fixed mutations versus the reference genome that were not observed in T5D. This can also be explained by small sample size and coverage in a pooled sample. Many of these sites may actually be variable in the populations (rather than fixed) yet missed in the sampled subsets.

For the previously discovered variants in AB, TU, TL, and WIK, SNPs in TU followed a slightly different read frequency distribution, with fewer fixed SNPs. This can be

explained in part by the heavy reliance of the reference genome sequence on TU zebrafish. Additionally, AB and TU had even fewer low-frequency SNPs, which can be explained by the lower average read depth per SNP site (median of 8 for AB and 9 for TU compared with 16 for TL and 13 for WIK).

In order to assess the similarity of T5D variation to a hybrid population that has previously employed an individual sequencing approach, SNP sites were compared to NHGRI-1 SNP sites. The NHGRI-1 line was derived from one mating pair of TAB-5 (a TU and AB cross), where the founding male was previously sequenced at 52X coverage and the female at 47X (LaFave et al. 2014). Even with the small sample size of 2, 15.7 M SNPs were discovered, with more than 10 M novel (i.e. not in dbSNP). Of these, 6.85 M overlap with the SNPs discovered in T5D.

Though more SNPs were found in a T5D sample that included more individuals, the two NHGRI-1 founders carried nonreference alleles at more sites (an average of 12.8 M variant sites per individual compared to 6.9 M in T5D). This may be partially-explained by the lower coverage per individual in our design, wherein we sacrificed sequencing depth per individual in order to include a larger sample and better estimate genotype frequencies for rare variants. These rare variants would not be captured without a reasonably large sample of individuals.

4.3.3 Downsampling to Approximate Sequencing Designs in Other Lines

In order to assess whether sequencing design could be a major driver behind observed SNP differences between lines, we used a downsampling strategy to approximate published designs used for other lines. We simulated a pool of 20 T5D individuals with average coverage of 20X across the genome by using a subset of the sequencing reads and analyzing them as one pooled sample. Even before applying filters, 49.8% as many variants were detected in this pooled sample compared to the whole dataset. After the simulated analysis, median read depth per variant site for T5D was 14 (within the range of 8-16 mentioned previously for the other 4 lines).

T5D variant counts and proportions of nonreference reads moved closer to those observed in other lines (Figure 4.3). Low frequency variants were no longer identifiable, and a larger proportion of the nonreference alleles incorrectly displayed themselves as fixed mutations (F=1 in Figure 4.3B & 4.3D). This downsampling approach resulted in a 2-fold reduction in variant calling capability, providing evidence that sequencing design could be a major driver of variability differences among zebrafish lines.

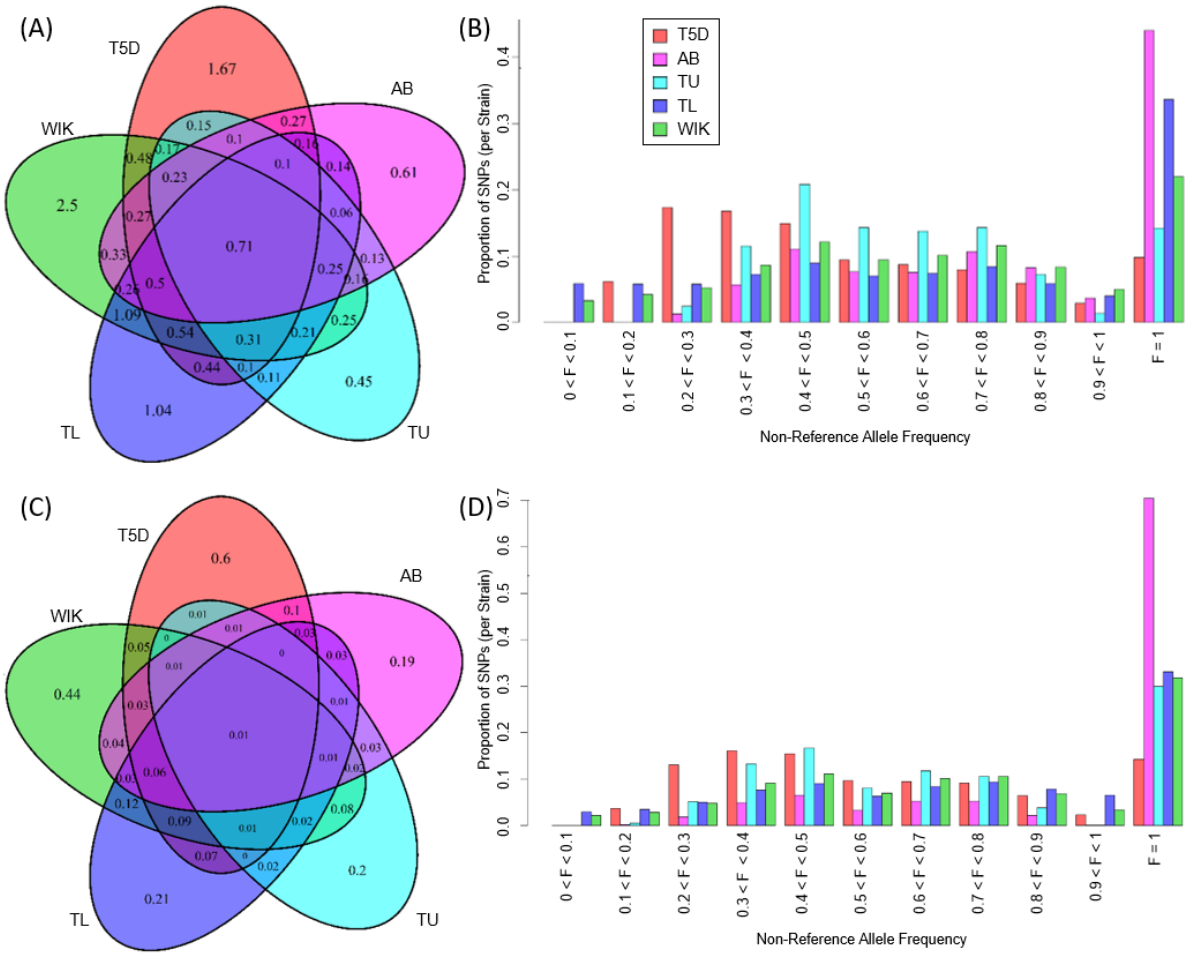

Figure 4.3. Zebrafish Variant Comparisons After Sequencing and Masking a Pooled Subsample. (A) Venn diagram of SNP sites (in millions) compared to the Zv9 reference genome. (B) Proportions of SNPs binned by alternate allele frequencies for the 5 lines. For all lines frequencies were determined based on the proportion of reads with nonreference base calls since no individual genotypes can be determined from pooled sequence alignment. (C) Venn diagram of indel sites (in millions). (D) Proportion of indels for discrete alternate allele frequencies.

4.4 Discussion

We used new data from a genome-wide sequencing project to compare and characterize observed population genetic variation across species (humans, mice, zebrafish). While more variants have been discovered in the human and mouse genomes, the smaller zebrafish genome is on par with—or may even exceed—genetic variability observed between individuals in those species. This diversity is attractive for translational applications in human and ecological health, where natural genetic variability could manifest as susceptibility differences to chemicals, drugs, environmental change, or other stressors. Though there are fewer zebrafish disease models compared to other species (Figure 4.1C), the number of genetic associations for many phenotypes of interest in health and environmental studies in zebrafish follows sequentially after human and mouse (Figure 4.1D). They are also gaining tractability as a model for human disease (Howe et al. 2017).

Variant discovery in the T5D wildtype zebrafish has confirmed the line's status as a heterogenous population. Considerably more SNPs and indels were discovered through individual whole genome sequencing of a large T5D sample than in other zebrafish studies, even exceeding the current build of dbSNP. Pooled sequencing data fundamentally affected the character of genetic variation previously detectable in outbred zebrafish lines, versus the individual-level sequencing data collected for T5D. In addition to discovering more variants, the design allowed us to estimate allele frequencies for a population more accurately than previously possible due to bias when estimating based on read frequencies in a pool (Raineri et al. 2012) or sample size of 2. Subsampling to simulate a pooled sequencing approach showed that T5D variation is in line with the more variable zebrafish laboratory strains (Figure 4.3). This likely means that (1) many of the variants discovered in T5D are present in other lines as well but have not been found due to pooling, low coverage, and sample size restrictions in previous zebrafish experiments, and (2) there are many more rare alleles that are yet to be discovered. This latter trend is very similar to continued improvements in rare allele discovery in humans (Shen et al. 2013). Our observations suggest that interindividual genetic diversity (i.e. natural variation) within laboratory populations may be higher than

currently estimated and may have implications for differential susceptibility observed in toxicological studies.

For exposure research, this means that healthy laboratory zebrafish strains that are sufficiently outbred can be a powerful model for human and other species environmental exposures. Their rapid development allows for high throughput studies that can expand scientific discovery exponentially. Continued work on identifying genetic variation in commonly used zebrafish lines will be important for exploration of gene-environment interactions (GxE), epigenetic modifications, and other genetic effects using the zebrafish model.

There are also long-term benefits associated with creating a database of known SNPs in zebrafish populations. This database of population genomic information can inform future research and can be expanded in later phases and through other projects. Changes in genotype frequencies within the population can be tracked, which can address whether genetic drift or unwanted selection is affecting a laboratory population aiming to maintain an "outbred" strategy that maintains diversity.

Additionally, population genetic information can be used to determine variants (SNPs, copy-number variants, etc.) associated with differential chemical responses (Balik-Meisner et al., *submitted*). Risk assessment can be improved significantly with actual knowledge of subgroup and chemical-specific genetic variability (e.g. confidence bounds or upper/lower limits) (Dankovic et al. 2015; Schulte et al. 2015; Betts and Shelton-Davenport 2016). This is true for applications that range from environmental chemical exposure studies or pharmaceutical trials in human populations to environmental emergencies affecting ecological species, such as the response to the spill of MCHM in West Virginia (http://ntp.niehs.nih.gov/results/areas/wvspill/studies/index.html). Thus, inclusion of knowledge regarding constitutive genetic diversity will benefit all translational applications of the zebrafish model, from the mechanistic to the ecological to the clinical.

4.5 References

Alkan C, Kavak P, Somel M, et al (2014) Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. BMC Genomics. doi: 10.1086/519795

Asharani P V., Lianwu Y, Gong Z, Valiyaveettil S (2015) Comparison of the toxicity of silver, gold and platinum nanoparticles in developing zebrafish embryos. Nanotoxicology. doi: 10.3109/17435390.2010.489207

Baer CE, Ippolito DL, Hussainzada N, et al (2014) Genome-wide gene expression profiling of acute metal exposures in male zebrafish. Genomics Data 2:363–365. doi: 10.1016/j.gdata.2014.10.013

Bai W, Zhang Z, Tian W, et al (2009) Toxicity of zinc oxide nanoparticles to zebrafish embryo: a physicochemical study of toxicity mechanism. J Nanoparticle Res 12:1645–1654. doi: 10.1007/s11051-009-9740-9

Betts K, Shelton-Davenport M (2016) Interindividual Variability: New Ways to Study and Implications for Decision Making: Workshop in Brief. In: National Academies Press (US). pp 1–13

Bowen ME, Henke K, Siegfried KR, et al (2012) Efficient mapping and cloning of mutations in zebrafish by low-coverage whole-genome sequencing. Genetics 190:1017–24. doi: 10.1534/genetics.111.136069

Brown KH, Dobrinski KP, Lee a. S, et al (2012) Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. Proc Natl Acad Sci 109:529–534. doi: 10.1073/pnas.1112163109

Butler MG, Iben JR, Marsden KC, et al (2015) SNPfisher: tools for probing genetic variation in laboratory-reared zebrafish. Development 142:1542–1552. doi: 10.1242/dev.118786

Chesler EJ, Miller DR, Branstetter LR, et al (2008) The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. Mamm Genome 19:382–9. doi: 10.1007/s00335-008-9135-8

Cho YS, Kim H, Kim H-M, et al (2016) An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun 7:13637. doi: 10.1038/ncomms13637

Churchill GA, Airey DC, Allayee H, et al (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137. doi: 10.1038/ng1104-1133

Churchill GA, Gatti DM, Munger SC, Svenson KL (2012) The Diversity Outbred mouse population. Mamm Genome 23:713–8. doi: 10.1007/s00335-012-9414-2

Cirelli C, Tononi G, Mackay TF, et al (2008) Is Sleep Essential? PLoS Biol 6:e216. doi: 10.1371/journal.pbio.0060216

Dankovic DA, Naumann BD, Maier A, et al (2015) The Scientific Basis of Uncertainty Factors Used in Setting Occupational Exposure Limits. J Occup Environ Hyg 12 Suppl 1:S55–68. doi: 10.1080/15459624.2015.1060325

Depristo MA, Banks E, Poplin RE, et al (2011) A framework for variation discovery and genotyping using next- generation DNA sequencing data. Nat Genet 43:491–498. doi: 10.1038/ng.806

French JE, Gatti DM, Morgan DL, et al (2015) Diversity outbred mice identify population-based exposure thresholds and genetic factors that infuence benzene-induced genotoxicity. Environ Health Perspect 123:237–245. doi: 10.1289/ehp.1408202

Han L, Zhao Z (2008) Comparative Analysis of CpG Islands in Four Fish Genomes. Comp Funct Genomics. doi: 10.1155/2008/565631

Howe DG, Bradford YM, Eagle A, et al (2017) The Zebrafish Model Organism Database: new support for human disease models, mutation details, gene expression phenotypes and searching. Nucleic Acids Res 45:D758–D768. doi: 10.1093/nar/gkw1116

Howe K, Clark MD, Torroja CF, et al (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503. doi: 10.1038/nature12111

Irie N, Kuratani S (2011) Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. Nat Commun. doi: 10.1038/ncomms1248

Ivanov DK, Escott-Price V, Ziehm M, et al (2015) Longevity GWAS Using the *Drosophila* Genetic Reference Panel. Journals Gerontol Ser A Biol Sci Med Sci 70:1470–1478. doi: 10.1093/gerona/glv047

Kang L, Aggarwal DD, Rashkovetsky E, et al (2016) Rapid genomic changes in Drosophila melanogaster adapting to desiccation stress in an experimental evolution system. BMC Genomics. doi: 10.1038/351652a0

Ka-Shu Wong G, Liu B, Wang J, et al (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432:717–722. doi: 10.1038/nature03156

Kijas JW, Townley D, Dalrymple BP, et al (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One 4:e4668. doi: 10.1371/journal.pone.0004668

Kimmel CB, Ballard WW, Kimmel SR, et al (1995) Stages of embryonic development of the zebrafish. Dev Dyn 203:253–310. doi: 10.1002/aja.1002030302

Knecht AL, Truong L, Marvel SW, et al (2017) Transgenerational inheritance of neurobehavioral and physiological deficits from developmental exposure to benzo[a]pyrene in zebrafish. doi: 10.1016/j.taap.2017.05.033

Kovács R, Csenki Z, Bakos K, et al (2015) Assessment of toxicity and genotoxicity of low doses of 5-fluorouracil in zebrafish (Danio rerio) two-generation study. Water Res 77:201–212. doi: 10.1016/j.watres.2015.03.025

LaFave MC, Varshney GK, Vemulapalli M, et al (2014) A Defined Zebrafish Line for High-Throughput Genetics and Genomics: NHGRI-1. Genetics 198:167–170. doi: 10.1534/genetics.114.166769

Lange M, Neuzeret F, Fabreges B, et al (2013) Inter-Individual and Inter-Strain Variations in Zebrafish Locomotor Ontogeny. PLoS One. doi: 10.1371/journal.pone.0070172

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi: 10.1038/nmeth.1923

Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinforma Appl NOTE 25:2078–2079. doi: 10.1093/bioinformatics/btp352

Lieschke GJ, Currie PD (2007) Animal models of human disease: zebrafish swim into view. Nat Rev Genet 8:353–367. doi: 10.1038/nrg2091

Mackay TFC, Richards S, Stone EA, et al (2012) The Drosophila melanogaster Genetic Reference Panel. Nature 482:173–178. doi: 10.1038/nature10811

Mckenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. doi: 10.1101/gr.107524.110

McLaren W, Gil L, Hunt SE, et al (2016) The Ensembl Variant Effect Predictor. Genome Biol. doi: 10.1186/s13059-016-0974-4

Moss SP, Joyce DA, Humphries S, et al (2011) Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. Genome Biol Evol 3:1187–96. doi: 10.1093/gbe/evr090

Mrakovcic M, Haley LE (1979) Inbreeding depression in the Zebra fish Brachydanio rerio (Hamilton Buchanan). J Fish Biol 15:323–327.

Nasiadka A, Clark MD (2012) Zebrafish Breeding in the Laboratory Environment. ILAR J 53:161–168. doi: 10.1093/ilar.53.2.161

Obholzer N, Swinburne I a., Schwab E, et al (2012) Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing. Development 139:4280–4290. doi: 10.1242/dev.083931

Oliveira R, Grisolia CK, Monteiro MS, et al (2016) Multilevel assessment of ivermectin effects using different zebrafish life stages. Comp Biochem Physiol Part C Toxicol Pharmacol 187:50–61. doi: 10.1016/j.cbpc.2016.04.004

Patowary A, Purkanti R, Singh M, et al (2013) A sequence-based variation map of zebrafish. Zebrafish 10:15–20. doi: 10.1089/zeb.2012.0848

Raineri E, Ferretti L, Esteve-Codina A, et al (2012) SNP calling by sequencing pooled samples.

Reif DM, Truong L, Mandrell D, et al (2016) High-throughput characterization of chemical-associated embryonic behavioral changes predicts teratogenic outcomes. Arch Toxicol 90:1459–1470. doi: 10.1007/s00204-015-1554-1

Roberts A, Pardo-Manuel de Villena F, Wang W, et al (2007) The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. Mamm Genome 18:473–81. doi: 10.1007/s00335-007-9045-1

Sakharkar MK, Perumal BS, Sakharkar KR, Kangueane P (2005) An analysis on gene architecture in human and mouse genomes. Silico Biol 5:347–365.

Schulte PA, Whittaker C, Curran CP (2015) Considerations for Using Genetic and Epigenetic Information in Occupational Health Risk Assessment and Standard Setting. J Occup Environ Hyg 12 Suppl 1:S69–81. doi: 10.1080/15459624.2015.1060323

Shen H, Li J, Zhang J, et al (2013) Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. PLoS One. doi: 10.1371/journal.pone.0059494

Shi L, Guo Y, Dong C, et al (2016) Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun 7:12065. doi: 10.1038/ncomms12065

Srivastava A, Morgan AP, Najarian ML, et al (2017) Genomes of the Mouse Collaborative Cross.

Stanley KA, Curtis LR, Massey Simonich SL, Tanguay RL (2009) Endosulfan I and endosulfan sulfate disrupts zebrafish embryonic development. Aquat Toxicol 95:355–361. doi: 10.1016/j.aquatox.2009.10.008

Svenson KL, Gatti DM, Valdar W, et al (2012) High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics 190:437–47. doi: 10.1534/genetics.111.132597

Truong L, Reif DM, Mary LS, et al (2014) Multidimensional in vivo hazard assessment using zebrafish. Toxicol Sci 137:212–233. doi: 10.1093/toxsci/kft235

Unckless RL, Rottschaefer SM, Lazzaro BP (2015) A genome-wide association study for nutritional indices in Drosophila. G3 (Bethesda) 5:417–25. doi: 10.1534/g3.114.016477

Usenko CY, Harper SL, Tanguay RL (2007) In vivo evaluation of carbon fullerene toxicity using embryonic zebrafish. Carbon N Y 45:1891–1898. doi: 10.1016/j.carbon.2007.04.021

Wilson CA, High SK, McCluskey BM, et al (2014) Wild Sex in Zebrafish: Loss of the Natural Sex Determinant in Domesticated Strains. Genetics 198:1291–1308.

Yang H, Wang JR, Didion JP, et al (2011) Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet 43:648–55. doi: 10.1038/ng.847

CHAPTER 5

Discussion, Conclusions, and Future Directions

Chapter 1 discusses computational methods for assessing individual and integrated "omic" data in a Systems Biology framework for environmental and toxicological science. Chapter 2 provides brief background on the specific topic of gene-environment interactions (GxE), as well as the major considerations for mining big data to design a GxE study in a specific population of zebrafish. Chapter 3 discusses the study design, implementation, and results in detail. Chapter 4 looks deeper into population variability that this analysis has brought to the fore. Chapter 5 addresses the impacts of experimental design choices, as well as future directions for the work.

5.1 Effects of Experimental Design Decisions

Power calculations were based on a contribution from a specific number of individuals per sample size scenario. In truth, due to the use of low coverage sequencing, fewer than the 276 individuals in the study contributed alleles to the Fisher's Exact test for the association analysis for the majority of the SNPs included in the final analysis. This resulted in an uneven number of total alleles per SNP-wise contingency table test.

The QQ plot for the Fisher's Exact test statistics in the final study showed evidence of putative inflation (Figure 5.1A). If we only included SNPs that were in non-repetitive areas of the genome (unmasked set mentioned in Chapter 4), a similar pattern would still emerge (Figure 5.1B). If we were to correct for this inflation using the genomic control procedure to normalize the test statistics by their median (Devlin and Roeder 1999), the plot would no longer display much obvious signal (Figure 5.2A). Here the dots seem to become elevated, indicating some signal, but they drop back down to and below the red line. However, correcting for a theoretical distribution that may not be accurate for this sample could impose too strict an adjustment. Additionally, if we were to choose a different analysis plan, such as logistic regression (Figure 5.1C), we would be statistically underpowered since that had not been the basis of our original power calculations.

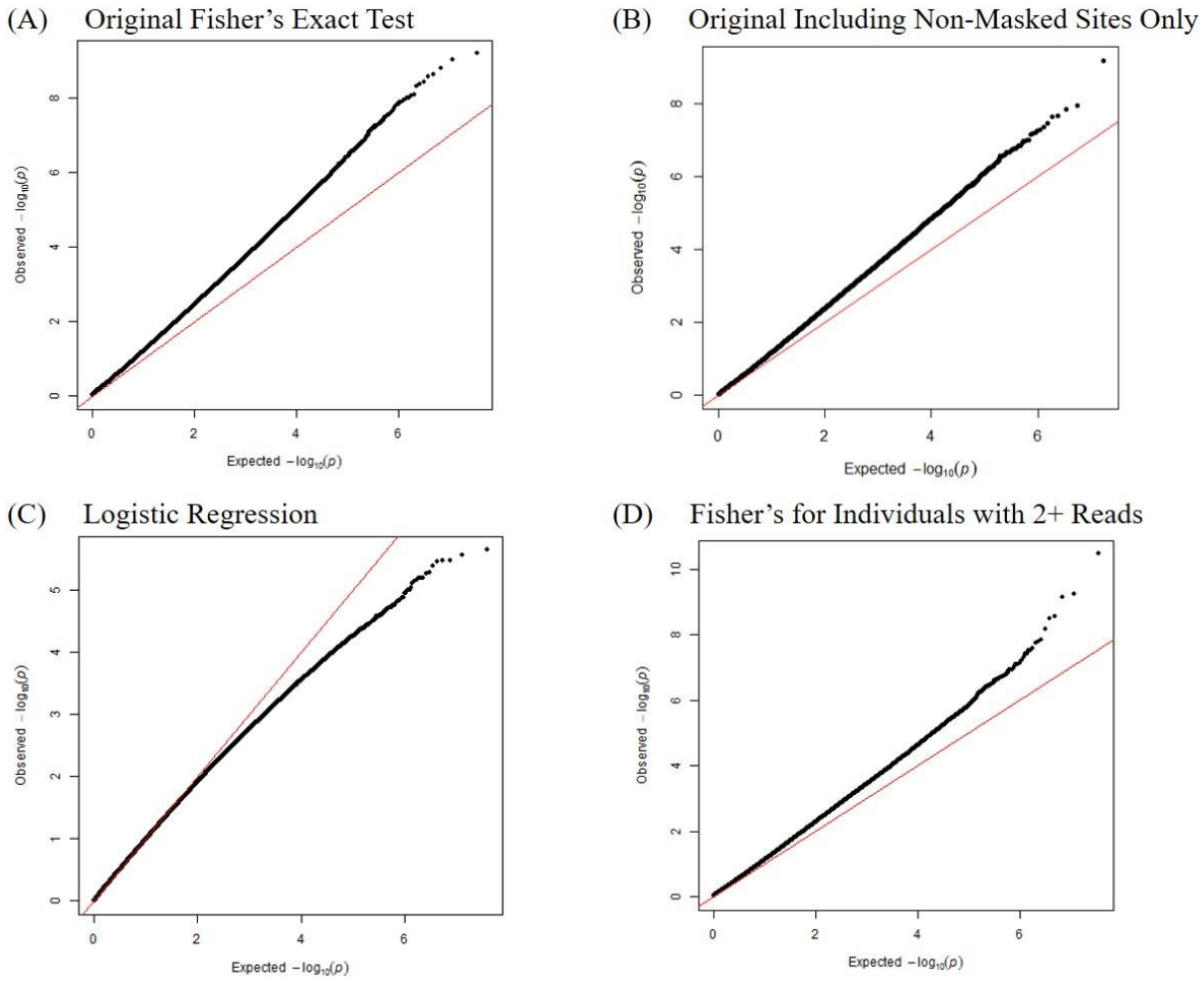

Figure 5.1. QQ Plots for GxE analysis. Red line indicates expectation if the observed log p-values followed the same distribution as the expected log p-values. (A) QQ plot for original Fisher's Exact Test. (B) QQ plot for original Fisher's Exact Test with non-complex regions of the genome masked from the analysis. (C) QQ plot based on logistic regression analysis. (D) QQ plot for Fisher's Exact test including only individuals per SNP with at least 2 reads covering that site.

One potential reason for inflation could be departure from HWE. As discussed in Chapter 2, low coverage can lead to an excess of homozygous calls, therefore skewing the genotypes away from HWE. Another implementation strategy to combat inflation could have been to only include allelic information from individuals that had at least 2 reads covering a given SNP (Figure 5.1D). After using the genomic control to combat inflation in this case, there is still apparent signal in the data (Figure 5.2B). However, this approach throws away information from many individuals.

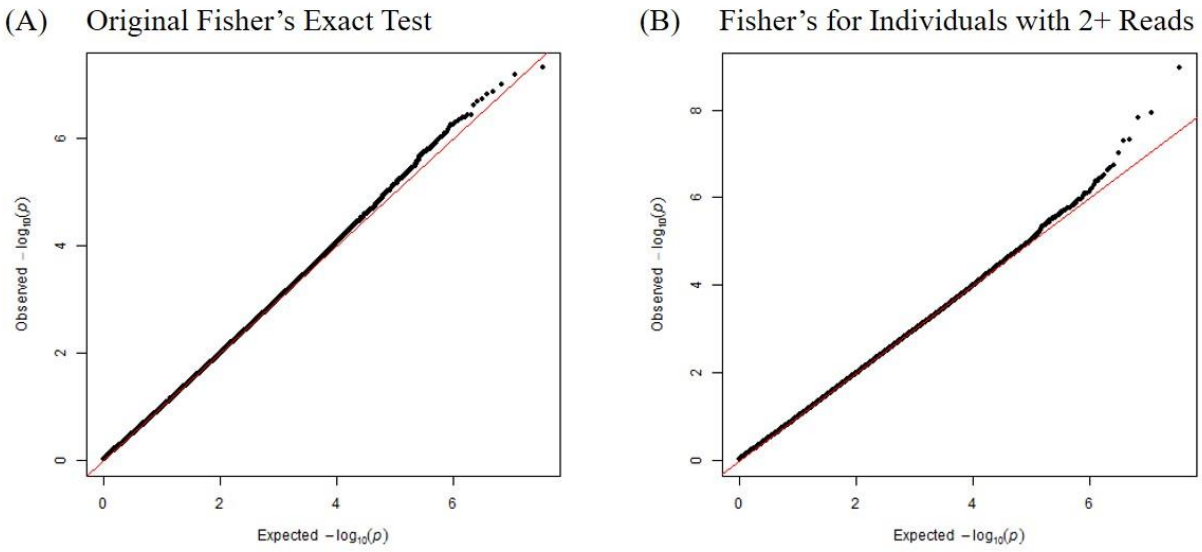

Figure 5.2. QQ Plots for GC-corrected GxE log p-values. Red line indicates expectation if the observed log p-values followed the same distribution as the expected log p-values. (A) QQ plot for GC-corrected original Fisher's Exact Test. (B) QQ plot for GC-corrected Fisher's Exact test including only individuals per SNP with at least 2 reads covering that site.

We ultimately reported the results to the Fisher's Exact test for all individuals with at least one read covering a given SNP site. If the $-\log_{10}$ p-values were, in fact, inflated, certain adjustments could pull our top hit from meeting the multiple-correction, genome-wide significance threshold. However, even in the absence of genome-wide significance, researchers often perform targeted follow-up on the top hits. Based on the biological plausibility of the SNP in the promoter region of *sox7* and the initial functional validation performed in Chapter 3, we are confident that variability in this region has some role in the GxE effect.

True linkage disequilibrium (LD) could also cause a departure from HWE. For example, the top SNP in the GxE association analysis was both a missing site for many individuals and was seen to have large departures from expected frequencies under HWE. We have performed targeted follow-up sequencing in this region and still find that genotype calls in this area are extremely difficult due to high repeat content. We also see that all individuals that have been successfully genotyped in this area are in fact homozygous.

Many different analysis, quality control, filtering, and other design decisions could have been made along the way. These choices could have led to other results. We believe there is more than one causative allele leading to differential susceptibility to 0.6 µM Abamectin exposure in T5D zebrafish, but we have shown that *sox7* differential expression plays an important role.

5.2 Conclusions

We have displayed reproducible population variability for a well-defined multivariate phenotype through consistent concentration-response across successive rounds of concentration narrowing. We have used novel experimental design and fine-tune control to compare the genomic sequences of Affected and Unaffected zebrafish under an identical exposure. The approach led to the identification and confirmation of a genetic region upstream of *sox7*.

Additionally, we have used this individual sequencing data to compare natural genetic diversity among zebrafish lines. The individual sequencing strategy led to the discovery of more SNPs than have been previously discovered in other lines of zebrafish that have utilized pooled sequencing strategies. We showed that similar strategies in T5D would have brought forth comparable results, leading to the conclusion that many of the novel T5D variants are actually shared with other lines but were censored due to previous experimental designs. T5D diversity is on par with the more variable laboratory zebrafish lines.

5.3 Future Directions

5.3.1 Extended Follow-Up

Our design allowed us to scan the entire genome for a handful of markers for increased Abamectin susceptibility. The GWAS significant SNP sites alone were not necessarily expected to be causative. They were viewed as evidence of regions of association, i.e. as markers either within or in LD with the true complete variant allele(s). For

example, we have observed additional insertion and deletion variation upstream of the *sox7* SNP in a highly repetitive region. One such deletion leading to a copy number decrease is highly correlated (r = 0.97) with the GxE significant SNP. It will take a continued effort to determine all of the variants that make up the complete causative allele.

The next most significant association was with a SNP in the first intron of *erf*. Unlike the top SNP, this site would not have remained in the analysis had we eliminated variant calls from noncomplex regions of the genome (52% of the genome masked in ftp://ftp.ensembl.org/pub/release-81/fasta/danio_rerio/dna/Danio_rerio.GRCz10.dna_rm.chromosome*.fa.gz). However, this was a previously discovered SNP (rs504509730) and falls within a gene related to the Affected phenotype. An *Erf* mutation in humans and mice is linked to craniosynostosis and related familial eye, snout, and jaw deformities (Twigg et al. 2013). Mice that are homozygous for an *Erf* mutation have pale yolk sacs (Papadaki et al. 2007). Beyond the top hit, our analysis continued to pick up variation in genomic regions that have biological plausibility and would be good targets for continued functional follow-up.

5.3.2 Rare Allele or Gene-Level SNP Analysis

Chapter 2 discussed more complex statistical designs that have been used in human GWAS studies to assess GxE effects (Marceau et al. 2015; Tzeng et al. 2011; Zhang et al. 2014; Zhao et al. 2015). Through control of all non-genetic factors, it was not necessary to utilize these approaches in our first zebrafish GxE GWAS. However, with upwards of 20 million SNPs in our GxE analysis, these types of approaches could bring to light more associations if SNPs were grouped together in a meaningful way. As annotations of the zebrafish genome continue to improve, grouping together SNPs into gene sets will become more feasible within this species.

5.3.3 Different Chemicals or Endpoints

There were other chemicals on the short-list of compounds eliciting patterns of variability in response to chemical exposure (Appendix B Table B.1). This approach could be redone with different chemicals from that list or for other compounds. For a future attempt it might be feasible to create a SNP chip based on the discovered SNPs rather than incurring the financial burdens of repeated WGS. Depending on the precision and multivariate nature of a chemical-specific Affected phenotype, alternate methods could be employed to determine the most informative grouping of morphological endpoints for subsequent trials (Zhang et al. 2016). Behavioral data could also provide alternative phenotypes to address.

5.3.4 Additional Line Sequencing

Deeper sequencing with larger sample sizes or individual WGS can be done in other zebrafish lines to create a more complete database of known zebrafish variants. Then we can better assess whether many T5D SNPs are truly line-specific, or if much of the variation within the species is common to multiple lines but has not been previously discovered due to pooled sequencing designs.

5.4 References

Devlin B, Roeder K. 1999. Genomic Control for Association Studies. Biometrics 55: 997–1004.

Marceau R, Lu W, Holloway S, Sale MM, Worrall BB, Williams SR, et al. 2015. A Fast Multiple-Kernel Method With Applications to Detect Gene-Environment Interaction. Genet. Epidemiol. 39:456–68; doi:10.1002/gepi.21909.

Papadaki C, Alexiou M, Cecena G, Verykokakis M, Bilitou A, Cross JC, et al. 2007. Transcriptional repressor erf determines extraembryonic ectoderm differentiation. Mol. Cell. Biol. 27:5201–13; doi:10.1128/MCB.02237-06.

Twigg SRF, Vorgia E, Mcgowan SJ, Peraki I, Fenwick AL, Sharma VP, et al. 2013. Reduced dosage of ERF causes complex craniosynostosis in humans and mice, and links ERK1/2 signaling to regulation of osteogenesis. Nat Genet. 43:308–313; doi:10.1038/ng.2539.

Tzeng J-Y, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, et al. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am. J. Hum. Genet. 89:277–88; doi:10.1016/j.ajhg.2011.07.007.

Zhang G, Marvel S, Truong L, Tanguay RL, Reif DM. 2016. Aggregate entropy scoring for quantifying activity across endpoints with irregular correlation structure. Reprod. Toxicol. 62:92–99; doi:10.1016/j.reprotox.2016.04.012.

Zhang R, Chu M, Zhao Y, Wu C, Guo H, Shi Y, et al. 2014. A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. Carcinogenesis 35:1528–35; doi:10.1093/carcin/bgu076.

Zhao G, Marceau R, Zhang D, Tzeng J-Y. 2015. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. Genetics 199:695–710; doi:10.1534/genetics.114.171686.

APPENDICES

Appendix A: Supplemental Materials from Chapter 2

Table A.1. P-values achieved by the effect scenarios for n=32, 64, 96, 128, 160, and 192. Red text displays where Allelic model power analysis using Fisher's exact test (http://vassarstats.net/tab2x2.html) values would first become significant for each scenario if a strict Bonferroni correction were used and tests were performed for 1 variant per zebrafish gene (~26,000 zebrafish genes), creating a significance threshold of 0.05/26,000 = 1.92e-6. For the actual analysis, we can implement more sophisticated corrections that maintain detection power at nominal p-values higher than those highlighted here.

P(A)	Effect	n=32	64	96	128	160	192
0.71	Perfect	7.98e-8	2.42e-15	7.37e-23	2.24e-30	2.33e-37	7.09e-45
	High	5.41e-3	5.06e-6	1.15e-7	1.82e-9	4.73e-12	1.76e-14
	Mid	9.94e-2	1.14e-2	8.25e-4	1.12e-4	4.22e-6	6.10e-7
	Low	2.74e-1	3.33e-1	1.14e-1	9.90e-2	3.67e-2	2.56e-2
0.5	Perfect	1.34e-4	2.23e-8	4.23e-12	8.49e-16	1.76e-19	3.72e-23
	High	7.93e-2	7.20e-4	8.79e-5	9.00e-7	1.21e-7	1.59e-8
	Mid	7.93e-2	5.14e-2	5.96e-3	3.95e-3	5.09e-4	6.51e-5
	Low	4.54e-1	2.16e-1	3.12e-1	1.69e-1	9.34e-2	5.23e-2

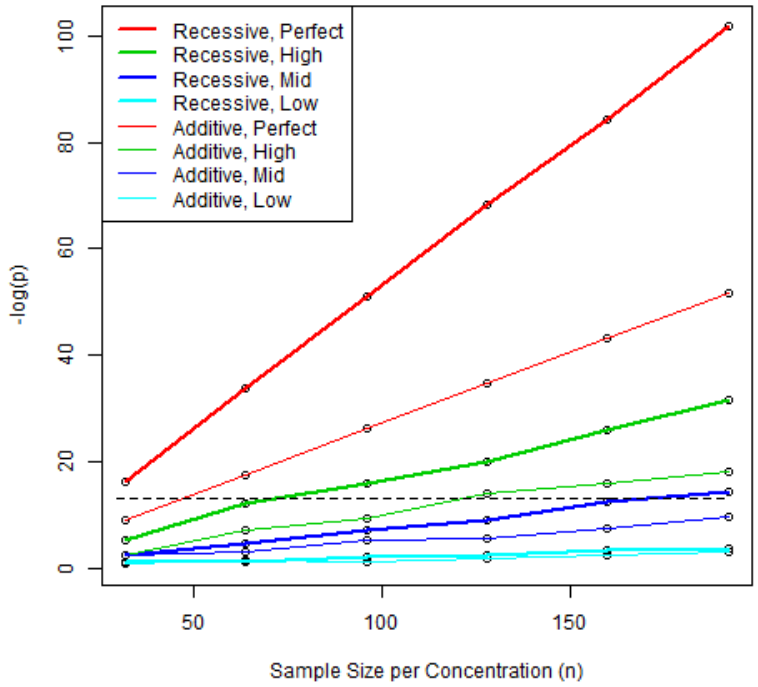

Figure A.1. Graphical display of the allelic power analysis. –log(p-value) is applied to better visualize the separation. The dotted line depicts the strict Bonferroni correction of a significance cutoff of 1.92e-6 (-log(1.92e-6) = 13.2). Graphical points above this line would pass this significance criterion.

Appendix B: Supplemental Materials from Chapter 3

Table B.1. Candidate Chemicals. These 19 chemicals (listed alphabetically) passed the heuristic for chemicals exposures leading to heightened interindividual phenotypic variability.

Chemical Name	CASRN
33’55’-Tetraiodothyroacetic acid	67-30-1
353'-Triiodothyronine	6893-02-3
Abamectin	71751-41-2
Aldicarb	116-06-3
Carfentrazone-ethyl	128639-02-1
Clotrimazole	23593-75-1
Cyclopamine	4449-51-8
Disulfiram	97-77-8
Emamectin benzoate	155569-91-8
Esfenvalerate	66230-04-4
Fipronil	120068-37-3
Pentachlorophenol	87-86-5
Picoxystrobin	117428-22-5
Sodium (2-pyridylthio)-N-oxide	3811-73-2
Tefluthrin	79538-32-2
Thiram	137-26-8
TNP-470	129298-91-5
Trifloxystrobin	141517-21-7
Ziram	137-30-4

Table B.2. Top SNPs (Bonferroni adjusted p < 0.05) associated with adverse outcomes (affected phenotype) in zebrafish exposed to 0.6 uM Abemectin. For each SNP, the bolded allele is the GRCz10 reference allele. Additional information for each SNP includes mean depth, frequency of missing individuals, whether the SNP is in a noncomplex region (52% of the genome) masked in the repeat masked version of GRCz10 downloaded from the Wellcome Trust Sanger Institute website (ftp://ftp.ensembl.org/pub/release-81/fasta/danio_rerio/dna/Danio_rerio.GRCz10.dna_rm.chromosome*.fa.gz), and gene annotation information.

Chromosome	Base	Minor Allele in Sample	Frequency in Affected	Frequency in Unaffected	Major Allele in Sample	P-value	Bonferroni adjusted p-value
20	19166444	T	0.4545	0.1282	G	6.43E-10	0.0126
19	6254036	C	0.2039	0.4821	G	1.60E-09	0.0313
8	53331820	C	0.3444	0.09722	G	2.39E-09	0.0467

Chromosome	Base	Odds Ratio	Mean Depth	Missing Genotype Frequency	rs Number (for previously discovered SNPs)	Masked	Gene
20	19166444	5.67	2.06	0.4783		No	569 bp upstream of sox7
19	6254036	0.28	3.59	0.2210	rs504509730	Yes	intron 1 of erf
8	53331820	4.88	2.23	0.2826		Yes	intron 6 of cfap74 (or 1889 bp upstream) intron 51 of cacna1db

Table B.3. Gene Expression Primers.

Gene	Primer Type	Primer Sequence
sox7	Forward	GCTCAGCAAGATGCTTGGAAA
	Reverse	CTTCCTGCGTGGACGGTATT
beta actin	Forward	AAG CAG GAG TAC GAT GAG TC
	Reverse	TGG AGT CCT CAG ATG CAT TG

Appendix C: Supplemental Materials from Chapter 4

Table C.1. SNP count per chromosome.

Chromosome	SNP Count	Chromosome Length (bp)	SNP Percentage
1	937,216	58,871,917	1.59
2	992,016	59,543,403	1.67
3	903,306	62,385,949	1.45
4	593,111	76,625,712	0.77
5	1,123,780	71,715,914	1.57
6	1,010,933	60,272,633	1.68
7	1,071,615	74,082,188	1.45
8	796,793	54,191,831	1.47
9	928,007	56,892,771	1.63
10	695,209	45,574,255	1.53
11	664,127	45,107,271	1.47
12	708,967	49,229,541	1.44
13	789,917	51,780,250	1.53
14	894,307	51,944,548	1.72
15	756,502	47,771,147	1.58
16	861,924	55,381,981	1.56
17	860,076	53,345,113	1.61
18	817,743	51,008,593	1.60
19	783,706	48,790,377	1.61
20	854,683	55,370,968	1.54
21	726,643	45,895,719	1.58
22	596,605	39,226,288	1.52
23	763,643	46,272,358	1.65
24	677,988	42,251,103	1.60
25	577,000	36,898,761	1.56

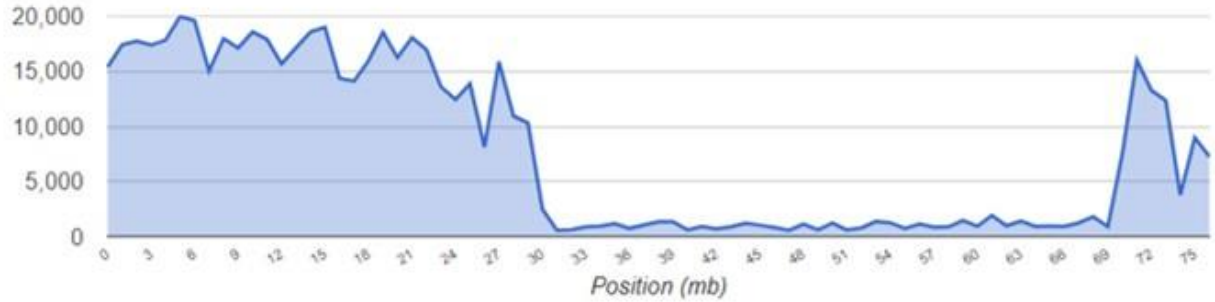

Figure C.1. Distribution of variants on chromosome 4. The y-axis displays the variant count partitioned into 1 mb bins of genomic sequence (x-axis).