ABSTRACT

WILLIAMS, LARISSA MORIARTY. Signatures of Selection in Natural Populations Adapted to Chronic Pollution. (Under the direction of Dr. Damian Shea and Dr. Marjorie F. Oleksiak).

Understanding the genetic basis of ecologically important traits and how such variation translates into functional phenotypes is a central goal of ecological and evolutionary genetics. For this thesis, the genetic basis of resistance to pollution was explored in populations of the estuarine minnow, *Fundulus heteroclitus*, which lives in extremely polluted estuaries along the east coast of the United States of America. Using molecular and genetic techniques, genetic variation within and among several *F. heteroclitus* populations exposed to environmental contamination was determined in order to identify signatures of natural selection.

In the first study, the genetic technique called amplified fragment length polymorphisms (AFLP) was used to generate hundreds of genome-wide markers. Then statistical tests were used to determine whether any of the genome-wide markers were under selection in three polluted *F. heteroclitus* populations compared to each of their reference populations. In total, 24 loci were determined to have outlying F_{ST} values in polluted populations. Several of these loci were outliers in two different, geographically separated populations, indicating a shared evolutionary response to contaminant exposure.

In the second study, high-throughput sequencing and genotyping technologies were used to determine genome-wide single nucleotide markers in order to assess levels of polymorphism along the genome. SNPs were polymorphic and showed latitudinal, clinal variation separating northern and southern *F. heteroclitus* populations. These markers also were used to differentiate *F. heteroclitus* from several other *Fundulus* species: *majalis*,

grandis and *similis*. This study established a technique to quickly and accurately scan the genome for polymorphisms for use in population genetics.

The third study analyzed the SNP data collected from the high-throughput sequencing and genotyping with the goal of identifying SNPs which were under selection in polluted populations. In contrast to the AFLP study, the nucleotide polymorphisms as well as surrounding nucleotide sequences were known which allowed additional statistical tests for neutrality. In total, one to four percent of the SNPs studied showed signatures of selection in any one polluted population. One SNP in the proximal promoter of the xenobiotic metabolizing enzyme, CYP1A, was identified as being under selection in all three polluted populations that were studied.

In order to determine the functional significance of the CYP1A SNPs under selection in polluted populations, the fourth study focused on determining genetic variation in the CYP1A promoter in one polluted and two reference populations and functionally characterizing the inducibility of the promoter *in vitro*. Overall, the nucleotide variability of the CYP1A promoter was high in all populations. There was also no significant, selective signature on the promoter as a whole, or any one portion (functional or non-functional) of the promoter suggesting that it is evolving through neutral processes in all three populations. The CYP1A promoter constructs were shown to be inducible in a dose-dependent manner *in vitro,* and induction was significantly higher in the polluted New Bedford Harbor population as compared to both reference populations. This result is surprising because previous studies have shown that CYP1A is refractory to induction by prototypic inducers in the New Bedford Harbor population.

Overall, this research highlights the role of natural selection due to exposure to anthropogenic contamination in shaping variation in natural populations of *F. heteroclitus*. This research was the first to complete several genome-wide scans for polymorphisms in this species and was also the first to describe signatures of selection in independent *Fundulus* populations exposed to contamination. It also established that the genetic basis for the refractory CYP1A transcriptional phenotype in polluted *F. heteroclitus* populations is not solely due to the CYP1A promoter. Ultimately, this work advanced our knowledge of how natural populations adapt to changing environmental conditions.

Signatures of Selection in Natural Populations Adapted to Chronic Pollution

by
Larissa Moriarty Williams

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

Toxicology

Raleigh, North Carolina

2010

APPROVED BY:

_____	_____
Damian Shea Committee Chair	Marjorie F. Oleksiak Co-Committee Chair
_____	_____
Robert C. Smart	W. Owen McMillan

_____
Patricia McClellan-Green

DEDICATION

I dedicate this work to my parents as well as the many strong women who have supported and guided me through the years.

BIOGRAPHY

Larissa Moriarty Williams was born May 13, 1983 in Ithaca, New York. She was welcomed by her parents David J. Williams and Karen E. Andrésen. Her family later expanded to include her sister Talia A. Williams. Growing up, Larissa was an avid horseback rider, spending endless hours at the barn with her riding instructor Barbara Watts. Through riding, she was introduced to Dr. Jacqueline Sinclair of Dartmouth University, who offered her a research internship in her toxicology laboratory following her first year at Smith College. At Smith College, located in Northampton, Massachusetts, she majored in Biological Sciences with a minor in Marine Science. It was in college that Larissa became an avid fan of marine biology and, at the suggestion of Larissa's favorite Smith professor, Dr. Paulette Peckol, attended a summer program at the Duke University Marine Laboratory (DUML) during her second summer in college where she took classes and completed an independent research project with Dr. Daniel Rittschof. The following summer Larissa worked with Dr. Patricia McClellan-Green at DUML on her summer project, funded by the Bookhout Fellowship Larissa was awarded, which ultimately evolved into her senior honors thesis. Larissa graduated in 2005 from Smith College with Highest Honors in the Biological Sciences. She went on to join the Ph.D. program at North Carolina State University in Environmental Toxicology under the direction of Dr. Margie Oleksiak. Two years into her Ph.D. program she moved to Miami, Florida to join Dr. Oleksiak at her new institution, the University of Miami's Rosenstiel School. Larissa's time in North Carolina and Miami has been fun, fruitful, and memorable.

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Margie Oleksiak, for her invaluable guidance, inspiration, and support through the Ph.D. process. Margie works harder than anyone I know, and I admire her extraordinary talents as a researcher. Margie is also one of the most patient advisors I have ever known, a quality that will benefit her and her future students to come.

I would also like to thank my committee at North Carolina State University, which allowed me to move to Miami and have been extremely supportive during my entire Ph.D. I could not have achieved this degree without them. I would especially like to thank Dr. Damian Shea who made it administratively and financially feasible for me to remain at NCSU and Dr. Patricia McClellan-Green for her endless support.

To Goran: Thanks for all your support, patience, laughs and guidance through this whole process. You are an incredible human being whom I admire, and I am a better person because of you.

Lastly, without the love and support of Michael, my family, and my friends, this thesis would not have been possible. Thanks for always being there for me; I am eternally grateful.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

LIST OF FIGURES

CHAPTER ONE

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

INTRODUCTION

Determining the molecular basis of adaptation is fundamental to understanding evolution. Adaptive evolution proceeds primarily through the process of natural selection whereby individuals with phenotypes that enhance their survival and reproductive rate will pass on heritable genetic characteristics at a higher rate in comparison to other individuals in a population. In successive generations, the genetic basis for traits associated with greater fitness in a particular environment can become more common in a population. It has long been a goal of biologists to identify the basis of adaptation to understand how organisms have adapted to a wide range of environments as well as assess the potential of populations to respond to environmental change. A powerful approach to study adaptation utilizes population genomics to study genome-wide allele frequency distribution and change under the influence of natural selection, genetic drift, mutation and gene flow. Population genomics characterizes molecular variation within and between populations and can be used to identify signatures of selection in populations.

The teleost fish, *Fundulus heteroclitus* has adapted to a diverse assortment of environments. They inhabit estuaries along the east coast of North America where they experience daily fluctuations in temperature, salinity and oxygen. In the last 150 years or so, portions of the east coast of the United States became industrialized and subsequently polluted with a variety of contaminants including persistent organic pollutants, metals, and pesticides, among others. Populations of *F. heteroclitus* have adapted to sites contaminated with these persistent toxicants. While the basis of adaptation for this species has been

studied intensely for over 20 years, the exact mechanism(s) by which *Fundulus heteroclitus* has been able to survive and thrive in such harsh environments is not yet known. In order to start understanding this phenomenon, my thesis has involved the use of population genetics to study the genetic basis of adaptation across three independent, polluted populations of *F. heteroclitus*.

The common mummichug: *Fundulus heteroclitus*

Fundulus heteroclitus is a small minnow, which resides along the east coast of the United States in tidal marshes from the Gulf of St. Lawrence to northeastern Florida. *F. heteroclitus* are mainly carnivorous although algae and detritus are often found in their gut contents (Allen *et al.*, 1994). Due to their significant detrital dietary component and their coastal habitat, they are exposed to a wide variety of environmental contaminants from the water and sediments. With a small home range of only a few hundred yards from its natal creek (Lotrich, 1975), these fish are indicative of local conditions. Nevertheless, migration of *Fundulus* is large enough to minimize genetic drift but small enough that natural selection can still occur (Brown, Chapman, 1991). Local populations of *F. heteroclitus* are calculated at 10^3 to 10^6 individuals, lending to large standing genetic variation (Adams *et al.*, 2006). The species is tolerant to a wide-range of environments ranging from hypo- to hypersalient conditions (Griffith, 1974), lives along the steepest thermal cline in the world (Crawford *et al.*, 1999), and inhabits extremely polluted estuaries (Weis, Weis, 1989). *F. heteroclitus* are easy to collect in the wild and maintain in the laboratory. Their life stages are hardy and individuals can withstand experimental manipulations. For these reasons, *F. heteroclitus*

makes an optimal model species to study the population genetic effects and genetic basis of adaptation to persistent anthropogenic contamination in natural populations.

EPA Superfund Collection Sites

New Bedford Harbor, located in Massachusetts, has been polluted since the growth in population during the mid 1800s where human waste was disposed of into the estuary (Voyer *et al.*, 2000). Following the demise of the textile industry in the early 1900s, the city of New Bedford recruited several electrical component manufacturers, who in turn polluted the Acushnet River estuary and adjoining harbor with polychlorinated biphenyls (PCBs). In addition to PCBs, dyes and chemicals used in rubber processing as well as metals also were dumped into the estuary and harbor following the expansion of the New Bedford industrial scene (Weaver, 1984). The abundance and persistence of the complex mix of contaminants placed New Bedford Harbor on the Superfund National Priority List (NPL) by the Environmental Protection Agency in 1983. The EPA manages the *NPL* as sites where known releases or threatened releases of hazardous substances, pollutants, or contaminants have occurred throughout the United States and its territories. The Superfund program is run through the US federal government to clean up the nation's uncontrolled hazardous waste sites.

In New Jersey, at the Diamond Alkali Company site, pesticides were manufactured including dichlorodiphenyltrichloroethane (DDT) and phenoxy herbicides from the mid 1940s to the 1970s (Trustees, 2007) on the Passaic river upstream of Newark Bay. In addition to the Diamond Alkali company site, several other industrial Superfund sites are

located in or upstream of Newark Bay: the Beazer East Inc. site carried out wood treatment from 1940-1991 leading to the polluting of polycyclic aromatic hydrocarbons (PAHs), Federated Metals Corporation operated from 1943-1984 and manufactured magnesium, aluminum, and zinc cathode protection anodes used on steel structures, and the Tidewater Baling site processed a variety of scrap metals. These industrial activities have contaminated sediments of the lower Passaic River and Newark Bay with dioxins, PCBs, mercury, DDT, pesticides, PAHs and heavy metals (Albrecht *et al.*, 1999; Brown *et al.*, 1994; Finley *et al.*, 1997; Huntley *et al.*, 1997; Huntley *et al.*, 1994; Iannuzzi *et al.*, 2005; Mitra *et al.*, 1999; Stackelberg, 1997; Wenning *et al.*, 1994; Yang, Sanudo-Wilhelmy, 1998). Newark Bay was first placed on the National Priority list in 1984.

From 1926 until 1992 the Atlantic Wood Industry, located on the banks of the Elizabeth River, Virginia, treated wood with the complex PAH mixture, creosote, and pentachlorophenol (PCP). The Navy also leased this land from 1960 to 1977 and disposed of abrasive blast media and sludge from the production of acetylene (Agency, 2009). As a result of these industrial operations, the Elizabeth River is contaminated with high concentrations of PAHs, PCP, dioxins, and several metals including arsenic, chromium, copper, lead and zinc (Conrad, Chisholm-Brause, 2004; Conrad *et al.*, 2007; Padma *et al.*, 1998). Elizabeth River was listed as a Superfund site in 1990.

Resistance to contaminants by *Fundulus heteroclitus*

F. heteroclitus from all three superfund sites have been shown to be resistant to the contaminants mixtures to which they are continuously exposed. In New Bedford Harbor,

fish are approximately 82 times less sensitive to the embryo-larval toxicity of PCB-126 than reference fish and 194% less responsive to cytochrome P4501A induction by prototypical inducers, PAHs and PCBs (Nacci *et al.*, 1999). A more recent study reported that early life stage LC20 values for PCB126, an indicator of relative sensitivity, were 2,000 less among NBH embryos as compared to reference sites (Nacci *et al.*, 2010). In adult fish, certain PCBs are accumulated far less than in reference fish (Gutjahr-Gobell *et al.*, 1999), differential gene expression patterns between polluted and reference sites have been documented (Fisher, Oleksiak, 2007; Oleksiak, 2008), and refractory CYP1A induction to PAHs and PCBs has been shown (Bello, 1999; Bello *et al.*, 2001). NBH fish are also more susceptible to challenges by marine bacterial pathogens in comparison to reference sites (Nacci *et al.*, 2009) but form fewer DNA adducts upon exposure to benzo(a)pyrene (Nacci *et al.*, 2002).

In Newark Bay, embryo and adult *F. heteroclitus* are less sensitive to dioxins and other toxicologically similar compounds (Arzuaga, Elskus, 2002; Bozinovic, Oleksiak, 2010; Elskus *et al.*, 1999; McArdle *et al.*, 2004; Nacci *et al.*, 2010; Prince, Cooper, 1995a). Similar to the New Bedford Harbor *F. heteroclitus*, lack of induction of CYP1A by prototypic inducers (dermal exposure to TCDD) has also been shown (Prince, Cooper, 1995b). Gene expression patterns between Newark and reference site *F. heteroclitus* show significantly different genes (Fisher, Oleksiak, 2007), many of which in the liver are in the oxidative phosphorylation pathway suggesting significant energy metabolism in polluted fish (Oleksiak, 2008). While seemingly resistant to contaminants, Newark Bay fish exhibit molecular and morphological changes indicative of impaired reproductive health and endocrine disruption compared to the reference population (Bugel *et al.*, 2010).

F. heteroclitus from the Elizabeth River site exhibit resistance towards creosote-induced acute toxicity which is heritable (Meyer, Di Giulio, 2002; Meyer, Di Giulio, 2003; Meyer *et al.*, 2002; Meyer *et al.*, 2005; Ownby *et al.*, 2002; Volgelbein WK, 1996). Elizabeth River embryos also exhibit 10,000 times less sensitivity towards PCB-126 (as measured by LC20) as compared to reference sites, a value that is five times greater than LC20 values for NBH and 14 times greater than Newark Bay embryos (Nacci *et al.*, 2010). Elizabeth River fish show elevated levels of proteins involved in phase II and III metabolism, namely glutathione-S-transferases (Van Veld *et al.*, 1991) and P-glycoproteins (Cooper *et al.*, 1999). Like that of the other Superfund sites, microsomal (Van Veld, 1995; Wills *et al.*, 2010b) and mitochondrial (Jung D., 2010) CYP1A is refractory to induction in *F. heteroclitus* from the Elizabeth River. B(a)P-induced damage was less on mitochondrial and nuclear DNA in Elizabeth River fish as compared to reference fish (Jung *et al.*, 2009), a similar finding to that reported in New Bedford Harbor fish (Bello, 1999; Bello *et al.*, 2001; Nacci *et al.*, 1999). Differential gene expression patterns are also found between Elizabeth River and its reference sites (Fisher, Oleksiak, 2007; Oleksiak, 2008), in particular CYP2N2 which is potentially involved in metabolism of xenobiotics (Oleksiak *et al.*, 2000) is up/down regulated in *F. heteroclitus* from the Elizabeth River. The Elizabeth River population also seems to have concomitantly developed resistance to chronic effects, including cancer, as compared to a reference population (Wills *et al.*, 2010a).

While there are many molecular and biochemical measurements of resistance in *F. heteroclitus*, there is still a lack of understanding of the underlying mechanisms driving adaptation. Since a gene candidate approach, such as characterizing CYP1A, has not

provided concrete answers, a global genome approach to identify selectively important loci may be a more fruitful alternative.

Population Genetics and Genome-Wide Scans

Population genomics is the study of numerous loci or genome regions to better understand the roles of evolutionary processes, such as mutation, random genetic drift, gene flow and natural selection, that influence variation across genomes and populations (Luikart *et al.*, 2003). Sampling of the genome across loci is used to determine locus-specific effects like that of selection, mutation, assertive mating and recombination, from genome-wide effects such as drift, bottlenecks, gene flow, and inbreeding. Locus specific effects can be used to identify genes important to fitness and adaptation whereas genome-wide effects provide information about population demography and phylogenetic history. To characterize and measure locus and genome-wide effects genetic or gene-product markers are used.

The earliest markers used to study populations and adaptation were allozymes (Hamrick *et al.*, 1979; Lewontin, 1974; Nevo, 1978; Nevo *et al.*, 1984), or variant forms of an enzyme that are coded by different alleles. Allozymes are readily translated across species (Nevo, 1990), but lack variability and number of loci which limits their power in statistical analyses for population genetics (Oleksiak, 2010). Since the age of DNA sequencing, allozymes have been replaced with more informative DNA markers such as microsatellites, mitochondrial DNA, amplified fragment length polymorphisms (AFLPs), and single nucleotide polymorphisms (SNPs). Microsatellites are repeating sequences of one to six base pairs of DNA and are codominant (homozygote and heterozygotes can be identified)

markers (Bahram, Inoko, 2007) which are highly variable and provide high statistical power for population genetic tests. Amplified fragment length polymorphisms are genomic fragments flanked by restriction sites which are amplified with the polymerase chain reaction (PCR) (Vos *et al.*, 1995) and are dominant (cannot discern homozygotes from heterozygotes) markers. AFLPs are widely used due to their ease of development and scoring in non-model species (Bensch, Akesson, 2005). Single nucleotide polymorphisms are point mutations which are abundant throughout coding and noncoding DNA and are codominant markers (Falque *et al.*, 2005). SNPs are abundant and widespread throughout the genome, can be typed reproducibly by many different methods (sequencing and high-throughput genotyping platforms), and provide high statistical power for population genetic tests.

For a population genomics study, a marker type is selected and tens to hundreds of loci are genotyped in a process called a genome-wide scan (GWS). Genotyping markers is necessary to determine both genome and locus specific effects, and to test for the neutrality of each marker. Neutral loci should not be under selection, are inherited according to Mendelian laws (transmission of heredity characteristics from parent organisms to their offspring), and are in Hardy-Weinberg proportions. Hardy-Weinberg is a law or model in which allele and genotype frequencies will reach equilibrium in one generation and remain constant from generation to generation in large, random-mating populations with no mutation, migration or selection (Luikart *et al.*, 2003). In order to estimate population genetic parameters, such as effective population size (N_e), mutation-drift equilibrium, bottlenecking, population subdivision, and migration between populations, markers should be neutral to avoid biasing results. Selectively important loci, or outlier loci, are generally

identified by comparing a data set against a theoretical (simulated) or empirical (observed) null distribution. The null distribution is founded on Kimura's theory of neutrally evolving mutations that states the vast majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral mutants (Kimura, 1983). The majority of studies use theoretical null distributions because it is difficult to generate robust, empirical null distributions. Thus, robust null distributions are used to model random effects in order to determine loci that are selectively important or linked to selectively important loci (through hitchhiking/linkage) or are outliers due to demography or population dynamics.

Several established tests can be used to detect directional selection in multilocus markers. The most widely used is the F_{ST} or fixation index test (Weir, 1984; Wright, 1951) which calculates the amount of genetic variation (reduction in heterozygosity in subpopulations relative to that expected in the population as a whole with random mating) among groups relative to a panmictic state. When examining multiple populations F_{ST} is often calculated for each locus between a pair of populations (Beaumont, Balding, 2004; Beaumont, Nichols, 1996; Bowcock *et al.*, 1991; Lewontin, Krakauer, 1973; McDonald, Golding, 1994; Porter, 2003; Vitalis *et al.*, 2001). To detect selection at any one locus, simulations are used to generate a null distribution of the F_{ST} summary statistic. Empirical F_{ST} values for each locus are then compared with the null distribution to determine statistical significance. Relative levels of diversity within populations, calculated though Theta (θ)-ratios tests (ratio between lnRV (variance) and lnRH (heterozygosity) statistics), can also be tested to detect selection (Schlotterer, 2002; Vigouroux *et al.*, 2002; Wootton *et al.*, 2002). Theta is defined as $4N_e\mu$, where N_e is the effective population size and μ is the mutation rate.

Theta-ratio tests assume that selective sweeps dramatically reduce genetic diversity in target genes or genomic regions. Theta-ratios are compared for each locus between populations and a p-value is calculated for that ratio based on a null distribution. A third test is known as the Ewens-Watterson test, which tests whether the observed Hardy-Weinberg homozygosity for a sample of size with *n* different alleles is significantly different from homozygosity expected under the neutral theory if the locus is at mutation-drift equilibrium (Ewens, 1972; Vigouroux *et al.*, 2002; Watterson, 1978). In practice, genome scans and their associated statistical tests have been used to identify selection in populations of *Drosophila* (Harr *et al.*, 2002; Kauer, 2003; Kauer *et al.*, 2003; Kingan *et al.*; Orengo, Aguade, 2007; Schofl, Schlotterer, 2004; Tracy *et al.*), parasites (Mu *et al.*; Nair *et al.*, 2003; Wootton *et al.*, 2002), maize (Vigouroux *et al.*, 2002), intertidal snails (Wilding *et al.*, 2001), lake whitefish (Campbell, Bernatchez, 2004), cod (Beaumont, Nichols, 1996; Moen *et al.*, 2008b; Pogson *et al.*, 1995), guppies (Willing *et al.*, 2010), salmon (Vasemagi *et al.*, 2005), three-spined stickleback (Makinen *et al.*, 2008), frogs (Bonin *et al.*, 2006), oysters (Murray, Hare, 2006), seagrass (Oetjen, Reusch, 2007), *Arabidopsis* (Cork, Purugganan, 2005), mosquitos (Bonin *et al.*, 2009), walking-sticks (Nosil *et al.*, 2008), beech trees (Jump *et al.*, 2006), white spruce (Namroud *et al.*, 2008), sunflowers (Kane, Rieseberg, 2007), cattle (Gibbs *et al.*, 2009), horses (Gu *et al.*, 2009), mice (Storz, Dubach, 2004; Storz, Nachman, 2003), rats (Kohn *et al.*, 2003; Tollenaere, 2010), and humans (Amato *et al.*, 2009; Bowcock *et al.*, 1991; Kayser *et al.*, 2003; Ronald, Akey, 2005; Storz *et al.*, 2004). These studies all identified regions of genomes that have been under diverse selective pressures or are linked to areas of the genome under selection.

Technology has enhanced efforts to sequence and genotype hundreds of genome-wide markers in non-model species. One such technology is called 454 pyrosequencing: it is a highly paralleled DNA sequencing approach that can provide over 20 Mbp of sequence in a four-hour run (Margulies *et al.*, 2005). Average read lengths are about 350-400 base pairs which are derived from approximately 300,000 samples (Huse *et al.*, 2007). This method does not require cloning of the DNA and therefore avoids certain biases that can be introduced by enzymatic steps or by instability of sequences in *E. coli* while cloning (Weber *et al.*, 2007). Pyrosequencing technology relies upon enzyme cascades and CCD luminescence detection capabilities to measure the release of inorganic pyrophosphate with every nucleotide incorporation (Ronaghi *et al.*, 1998). Sequencing takes place within an oil and water emulsion containing a capture bead. PCR amplification is performed prior to sequencing whereby one copy of the sample is captured on the bead and amplified to millions of copies. The beads are then distributed on a solid-phase sequencing substrate with 1.6 million wells; each well holds a bead and additional reagents including polymerase, luciferase, and ATP sulfurylase. Four nucleotide triphosphates are cycled over the plate, and when a nucleotide is incorporated, a pyrophosphate is released. This pyrophosphate is the substrate for luciferase, so luminescence occurs and is recorded. The number of reads per run makes it possible to have many different copies of the same gene, creating the opportunity for quality control. Further, studies have shown that 454 pyrosequencing is appropriate for many genomic applications, from SNP analysis ((Alderborn *et al.*, 2000; Moen *et al.*, 2008a; Namroud *et al.*, 2008; Novaes *et al.*, 2008; Ronaghi, Elahi, 2002; Wiedmann *et al.*, 2008)) to transcriptome profiling (Bainbridge *et al.*, 2006; Moustafa *et al.*;

Shin *et al.*, 2008; Vera *et al.*, 2008; Weber *et al.*, 2007). 454 pyrosequencing can be coupled to high-throughput genotyping platforms to expand genotyping to additional individuals or to verify SNPs initially identified with 454.

One platform useful for genotyping SNPs in non-model species due to limited sequence information needs and high multiplexing across SNPs and individuals (which ultimately lowers cost per nucleotide) is MassARRAY. MassARRAY uses an initial locus-specific PCR reaction followed by single base extension using mass-modified dideoxynucleotide terminator of an oligonucleotide primer which anneals immediately upstream of the single nucleotide polymorphic site. Using MALDI-TOF mass spectrometry, the distinct mass of the extended primer identifies the SNP allele (Gabriel *et al.*, 2009). MassARRAY has been used successfully to identify and verify SNPs in a wide array of organisms (Abel *et al.*, 2006; Berard *et al.*, 2009; Buggs *et al.*; Craig *et al.*, 2009; Goddard *et al.*, 2007; Nakai *et al.*, 2002; Rohrer *et al.*, 2007).

Once SNPs have been identified and verified for population genetic studies, the next step is to characterize selectively important loci by definitively linking genotype to phenotype, moving forward from population genomics to functional genomics and defining how much of the variation is biologically important (Oleksiak, 2010). Several studies have proof of concept between selective sweeps on populations and functional genomic outcomes: coat color and shortened limbs in dog breed formation (Pollinger *et al.*, 2005); chloroquine resistance in the malaria-inducing parasite *Plasmodium falciparum* (Wootton *et al.*, 2002); cyclodiene insecticide resistance (Ffrench-Constant, 1994); dichlorodiphenyltrichloroethane (DDT) insecticide resistance (Daborn *et al.*, 2002); organophosphate insecticide resistance

(Raymond *et al.*, 2001); warfarin resistance in rats (Kohn *et al.*, 2003); adaptive color pattern in beach mice (Hoekstra *et al.*, 2006); malarial resistance in humans with sickle cell (Currat *et al.*, 2002; Ohashi *et al.*, 2004); differential susceptibility of humans to cerebral malaria (Aitman *et al.*, 2000; Pain *et al.*, 2001); and variation in zebrafish and human pigmentation (Lamason *et al.*, 2005). In these cases, one locus had a large effect on the phenotype. The paucity of examples linking genotype to phenotype may reflect the fact that selection often acts on multiple loci (Lee *et al.*, 2008), especially under heterogeneous environmental conditions (Weinig *et al.*, 2003). It is this latter point, where multiple selectively important loci are contributing to a phenotype, which makes it difficult to directly link genotype to phenotype to fitness.

Thesis Objectives

For my thesis, I integrated molecular biology and population genetics in order to define regions of the genome subject to selection in *F. heteroclitus* populations adapted to high concentrations of anthropogenic contaminants. This thesis represents the first, comprehensive, genome-wide study of the molecular variation and signatures of selection among polluted and reference populations of *F. heteroclitus*. To this end, I conducted an AFLP analysis of genome-wide variation on both sensitive and resistant populations of *F. heteroclitus* to assess neutral variation and attempt to identify regions of the genome under selection (chapter 1). Because AFLPs are anonymous, I next used high-throughput sequencing and genotyping methods to assess the technology's utility in determining SNPs within AFLP fragments and expressed sequence tags (ESTs). Once identified and verified, I

analyzed SNPs among and between populations and species of the *Fundulus* genus to establish population parameters (chapter 2). To establish SNPs under selection in the same polluted and reference *F. heteroclitus* populations used in the first study, I performed three different statistical tests (chapter 3). In the last data chapter (chapter 4), I sought to characterize the potentially adaptive phenotypic outcome of one SNP, a SNP in the CYP1A promoter, that was under selection in all polluted populations. In total, this thesis describes one of the first genome-wide scans of a single species throughout many, variable, natural populations across geographies and environmental conditions. It elucidates selectively important regions of the genome, which may contribute to the phenotypic resistance of *F. heteroclitus* to anthropogenic pollution.

References

Abel K, Reneland R, Kammerer S, *et al.* (2006) Genome-wide SNP association: Identification of susceptibility alleles for osteoarthritis. *Autoimmunity Reviews* **5**, 258-263.

Adams SM, Lindmeier JB, Duvernell DD (2006) Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. *Molecular Ecology* **15**, 1109-1123.

Agency EP (2009) Current Site Description, Atlantic Wood Industries, Inc.

Aitman TJ, Cooper LD, Norsworthy PJ, *et al.* (2000) Malaria susceptibility and CD36 mutation. *Nature* **405**, 1015-1016.

Albrecht ID, Barkovskii AL, Adriaens P (1999) Production and dechlorination of 2,3,7,8-tetrachlorodibenzo-p-dioxin in historically-contaminated estuarine sediments. *Environmental Science & Technology* **33**, 737-744.

Alderborn A, Kristofferson A, Hammerling U (2000) Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. *Genome Research* **10**, 1249-1258.

Allen EA, Fell PE, Peck MA, *et al.* (1994) Gut contents of common mummichogs, *Fundulus heteroclitus* L., in a restored impounded march and in natural reference marshes *Estuaries* **17**, 462-471.

Amato R, Pinelli M, Monticelli A, *et al.* (2009) Genome-Wide Scan for Signatures of Human Population Differentiation and Their Relationship with Natural Selection, Functional Pathways and Diseases. *PLoS ONE* **4**, e7927.

Arzuaga X, Elskus A (2002) Evidence for resistance to benzo[a]pyrene and 3,4,3 ' 4 '-tetrachlorobiphenyl in a chronically polluted *Fundulus heteroclitus* population. *Marine Environmental Research* **54**, 247-251.

Bahram S, Inoko H (2007) Microsatellite markers for genome-wide association studies. *Nature Reviews Genetics* **8.**

Bainbridge MN, Warren RL, Hirst M, *et al.* (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *Bmc Genomics* **7**.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences* **263**, 1619-1626.

Bello SM (1999) *Characterization of resistance to halogenated aromatic hydrocarbons in a population of Fundulus heteroclitus from a marine superfund site*, Woods Hole Oceanographic Institute/Massachusetts Institute of Technology.

Bello SM, Franks DG, Stegeman JJ, Hahn ME (2001) Acquired resistance to Ah receptor agonists in a population of Atlantic killifish (*Fundulus heteroclitus*) inhabiting a marine superfund site: In vivo and in vitro studies on the inducibility of xenobiotic metabolizing enzymes. *Toxicological Sciences* **60**, 77-91.

Bensch S, Akesson M (2005) Ten years of AFLP in ecology and evolution: why so few animals? *Molecular Ecology* **14**, 2899-2914.

Berard A, Le Paslier MC, Dardevet M, *et al.* (2009) High-throughput single nucleotide polymorphism genotyping in wheat (*Triticum* spp.). *Plant Biotechnology Journal* **7**, 364-374.

Bonin A, Paris M, Tetreau G, David J-P, Despres L (2009) Candidate genes revealed by a genome scan for mosquito resistance to a bacterial insecticide: sequence and gene expression variations. *Bmc Genomics* **10**, 551.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution* **23**, 773-783.

Bowcock AM, Kidd JR, Mountain JL, *et al.* (1991) Drift, Admixture, and Selection in Human-Evolution - a Study with DNA Polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 839-843.

Bozinovic G, Oleksiak MF (2010) Embryonic gene expression among pollutant resistant and sensitive *Fundulus heteroclitus* populations. *Aquatic Toxicology* **98**, 221-229.

Brown BL, Chapman RW (1991) Gene flow and mitochondrial DNA variation in the killifish, *Fundulus heteroclitus. Evolution* **45**, 1147-1161.

Brown RP, Cooper KR, Cristini A, Rappe C, Bergqvist PA (1994) Polychlorinated dibenzo-p-dioxins and dibenzofurans in mya-arenaria in the Newark Raritan Bay estuary *Environmental Toxicology and Chemistry* **13**, 523-528.

Bugel SM, White LA, Cooper KR (2010) Impaired reproductive health of killifish (*Fundulus heteroclitus*) inhabiting Newark Bay, NJ, a chronically contaminated estuary. *Aquatic Toxicology* **96**, 182-193.

Buggs RJA, Chamala S, Wu W, *et al.* Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology* **19**, 132-146.

Campbell D, Bernatchez L (2004) Genomic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology and Evolution* **21**, 1164-1164.

Conrad CF, Chisholm-Brause CJ (2004) Spatial survey of trace metal contaminants in the sediments of the Elizabeth River, Virginia. *Marine Pollution Bulletin* **49**, 319-324.

Conrad CF, Fugate D, Daus J, Chisholm-Brause CJ, Kuehl SA (2007) Assessment of the historical trace metal contamination of sediments in the Elizabeth River, Virginia. *Marine Pollution Bulletin* **54**, 385-395.

Cooper PS, Vogelbein WK, Van Veld PA (1999) Altered expression of the xenobiotic transporter P-glycoprotein in liver and liver tumours of mummichog (*Fundulus heteroclitus*) from a creosote-contaminated environment. *Biomarkers* **4**, 48-58.

Cork JM, Purugganan MD (2005) High-diversity genes in the *Arabidopsis* genome. *Genetics* **170**, 1897-1911.

Craig DW, Millis MP, DiStefano JK (2009) Genome-wide SNP genotyping study using pooled DNA to identify candidate markers mediating susceptibility to end-stage renal disease attributed to Type 1 diabetes. *Diabetic Medicine* **26**, 1090-1098.

Crawford DL, Pierce VA, Segal JA (1999) Evolutionary physiology of closely related taxa: Analyses of enzyme expression. *American Zoologist* **39**, 389-400.

Currat M, Trabuchet G, Rees D, *et al.* (2002) Molecular analysis of the beta-globin gene cluster in the niokholo mandenka population reveals a recent origin of the beta(S) senegal mutation. *American Journal of Human Genetics* **70**, 207-223.

Daborn PJ, Yen JL, Bogwitz MR, *et al.* (2002) A single P450 allele associated with insecticide resistance in *Drosophila*. *Science* **297**, 2253-2256.

Elskus AA, Monosson E, McElroy AE, Stegeman JJ, Woltering DS (1999) Altered CYP1A expression in *Fundulus heteroclitus* adults and larvae: a sign of pollutant resistance? *Aquatic Toxicology* **45**, 99-113.

Ewens WJ (1972) Sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87.

Falque M, Decousset L, Dervins D, *et al.* (2005) Linkage mapping of 1454 new maize candidate gene loci. *Genetics* **170**, 1957-1966.

Ffrench-Constant RH (1994) The molecular and population genetics of cyclodiene insecticide resistance. *Insect Biochemistry and Molecular Biology* **24**, 335-345.

Finley BL, Trowbridge KR, Burton S, *et al.* (1997) Preliminary assessment of PCB risks to human and ecological health in the lower Passaic River. *Journal of Toxicology and Environmental Health* **52**, 95-118.

Fisher MA, Oleksiak MF (2007) Convergence and divergence in gene expression among natural populations exposed to pollution. *Bmc Genomics* **8**.

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr Protoc Hum Genet* **Chapter 2**, Unit 2.12.

Gibbs RA, Taylor JF, Van Tassell CP, *et al.* (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* **324**, 528-532.

Goddard KAB, Tromp G, Romero R, *et al.* (2007) Candidate-gene association study of mothers with pre-eclampsia, and their infants, analyzing 775 SNPs in 190 genes. *Human Heredity* **63**, 1-16.

Griffith RW (1974) Environment and salinity tolerance in the genus *Fundulus Copeia*, 319-331.

Gu J, Orr N, Park SD, *et al.* (2009) A Genome Scan for Positive Selection in Thoroughbred Horses. *PLoS ONE* **4**, e5767.

Gutjahr-Gobell RE, Black DE, Mills LJ, *et al.* (1999) Feeding the mummichog (*Fundulus heteroclitus*) a diet spiked with non-ortho- and mono-ortho-substituted polychlorinated biphenyls: Accumulation and effects. *Environmental Toxicology and Chemistry* **18**, 699-707.

Hamrick JL, Linhart YB, Mitton JB (1979) Relationships between life history characteristics and electrophoretically detectable genetic variation in plants. *Annual Review of Ecology and Systematics* **10**, 173-200.

Harr B, Kauer M, Schlotterer C (2002) Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12949-12954.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313**, 101-104.

Huntley SL, Iannuzzi TJ, Avantaggio JD, *et al.* (1997) Combined sewer overflows (CSOs) as sources of sediment contamination in the lower Passaic River, New Jersey .2. Polychlorinated dibenzo-p-dioxins, polychlorinated dibenzofurans, and polychlorinated biphenyls. *Chemosphere* **34**, 233-250.

Huntley SL, Wenning RJ, Paustenbach DJ, Wong AS, Luksemburg WJ (1994) Potential sources of polychlorinated dibenzothiophenes in the Passaic River, New Jersey *Chemosphere* **29**, 257-272.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**.

Iannuzzi TJ, Armstrong TN, Thelen JB, Ludwig DF, Firstenberg CE (2005) Characterization of chemical contamination in shallow-water estuarine habitats of an industrialized river. Part 1: Organic compounds. *Soil & Sediment Contamination* **14**, 13-33.

Jump AS, Hunt JM, Martinez-Izquierdo JA, Penuelas J (2006) Natural selection and climate change: temperature-linked spatial and temporal trends in gene frequency in *Fagus sylvatica*. *Molecular Ecology* **15**, 3469-3480.

Jung D, Cho Y, Collins LB, Swenberg JA, Di Giulio RT (2009) Effects of benzo[a]pyrene on mitochondrial and nuclear DNA damage in Atlantic killifish (*Fundulus heteroclitus*) from a creosote-contaminated and reference site. *Aquatic Toxicology* **95**, 44-51.

Jung D. DG, R.T. (2010) Identification of mitochondrial cytochrome P450 induced in response to polycyclic aromatic hydrocarbons in the mummichug (*Fundulus heteroclitus*). *Comparative Biochemistry and Physiology C-Toxicology & Pharmacology* **151**, 107-112.

Kane NC, Rieseberg LH (2007) Selective Sweeps Reveal Candidate Genes for Adaptation to Drought and Salt Tolerance in Common Sunflower, Helianthus annuus. *Genetics* **175**, 1823-1834.

Kauer M, Zanger, B., Dieringer, D., Schlotterer, C. (2003) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics* **160**, 247-256.

Kauer MO, Dieringer D, Schlotterer C (2003) A microsatellite variability screen for positive selection associated with the "Out of Africa" habitat expansion of Drosophila melanogaster. *Genetics* **165**, 1137-1148.

Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution* **20**, 893-900.

Kimura M (1983) *The Neutral Theory of Molecular Evolution* Cambridge University Press, Cambridge.

Kingan SB, Garrigan D, Hartl DL Recurrent selection on the Winters sex-ratio genes in Drosophila simulans. *Genetics* **184**, 253-265.

Kohn MH, Pelz HJ, Wayne RK (2003) Locus-specific genetic differentiation at Rw among warfarin-resistant rat (*Rattus norvegicus*) populations. *Genetics* **164**, 1055-1070.

Lamason RL, Mohideen M-APK, Mest JR, *et al.* (2005) SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science* **310**, 1782-1786.

Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. *PLoS Genet* **4**, e1000231.

Lewontin RC (1974) *The genetic basis of evolutionary change* Columbia University Press, New York.

Lewontin RC, Krakauer J (1973) Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms. *Genetics* **74**, 175-195.

Lotrich VA (1975) Summer home range and movements of *Fundulus heteroclitus* (Pisces Cyprinodontidae) in a tidal creek. *Ecology* **56**, 191-198.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.

Makinen HS, Shikano T, Cano JM, Merila J (2008) Hitchhiking mapping reveals a candidate genomic region for natural selection in three-spined stickleback chromosome VIII. *Genetics* **178**, 453-465.

Margulies M, Egholm M, Altman WE, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

McArdle ME, McElroy AE, Elskus AA (2004) Enzymatic and estrogenic responses in fish exposed to organic pollutants in the New York-New Jersey (USA) Harbor Complex. *Environmental Toxicology and Chemistry* **23**, 953-959.

McDonald JH, Golding B (1994) Detecting natural selection by comparing geographic variation in protein and DNA polymorphisms. *Non-neutral evolution: theories and molecular data.*, 88-100.

Meyer J, Di Giulio R (2002) Patterns of heritability of decreased EROD activity and resistance to PCB 126-induced teratogenesis in laboratory-reared offspring of killifish (*Fundulus heteroclitus*) from a creosote-contaminated site in the Elizabeth River, VA, USA. *Marine Environmental Research* **54**, 621-626.

Meyer JN, Di Giulio RT (2003) Heritable adaptation and fitness costs in killifish (*Fundulus heteroclitus*) inhabiting a polluted estuary. *Ecological Applications* **13**, 490-503.

Meyer JN, Nacci DE, Di Giulio RT (2002) Cytochrome P4501A (CYP1A) in killifish (*Fundulus heteroclitus*): Heritability of altered expression and relationship to survival in contaminated sediments. *Toxicological Sciences* **68**, 69-81.

Meyer JN, Volz DC, Freedman JH, Di Giulio RT (2005) Differential display of hepatic mRNA from killifish (*Fundulus heteroclitus*) inhabiting a Superfund estuary. *Aquatic Toxicology* **73**, 327-341.

Mitra S, Dellapenna TM, Dickhut RM (1999) Polycyclic aromatic hydrocarbon distribution within lower Hudson River estuarine sediments: Physical mixing vs sediment geochemistry. *Estuarine Coastal and Shelf Science* **49**, 311-326.

Moen T, Hayes B, Nilsen F, *et al.* (2008a) Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *Bmc Genetics* **9**, 18.

Moustafa A, Evans AN, Kulis DM, *et al.* Transcriptome Profiling of a Toxic Dinoflagellate Reveals a Gene-Rich Protist and a Potential Impact on Gene Expression Due to Bacterial Presence. *PLoS ONE* **5**, e9688.

Mu JB, Myers RA, Jiang HY, *et al. Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature Genetics* **42**, 268-U113.

Murray MC, Hare MP (2006) A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, Crassostrea virginica. *Molecular Ecology* **15**, 4229-4242.

Nacci D, Champlin D, Jayaraman S (2010) Adaptation of the Estuarine Fish *Fundulus heteroclitus* (Atlantic Killifish) to Polychlorinated Biphenyls (PCBs). *Estuaries and Coasts* **33**, 853-864.

Nacci D, Coiro L, Champlin D, *et al.* (1999) Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. *Marine Biology* **134**, 9-17.

Nacci D, Huber M, Champlin D, *et al.* (2009) Evolution of tolerance to PCBs and susceptibility to a bacterial pathogen (*Vibrio harveyi*) in Atlantic killifish (*Fundulus heteroclitus*) from New Bedford (MA, USA) harbor. *Environmental Pollution* **157**, 857-864.

Nacci DE, Kohan M, Pelletier M, George E (2002) Effects of benzo[a]pyrene exposure on a fish population resistant to the toxic effects of dioxin-like compounds. *Aquatic Toxicology* **57**, 203-215.

Nair S, Williams JT, Brockman A, *et al.* (2003) A selective sweep driven by pyrimethamine treatment in southeast Asian malaria parasites. *Molecular Biology and Evolution* **20**, 1526-1536.

Nakai K, Habano W, Fujita T, *et al.* (2002) Highly multiplexed genotyping of coronary artery disease-associated SNPs using MALDI-TOF mass spectrometry. *Human Mutation* **20**, 133-138.

Namroud M, Beaulieu J, Juge N, Laroche J, Bousquet J (2008a) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**, 3599 - 3613.

Nevo E (1978) Genetic variation in natural populations: Patterns and theory. *Theoretical Population Biology* **13**, 121-177.

Nevo E (1990) Molecular Evolutionary Genetics of Isozymes - Pattern, Theory, and Application. *Isozymes* **344**, 701-742

Nevo E, Beiles A, Ben-Shlomo R (1984) The evolutionary significance of genetic diversity: ecological, demographic and life history correlates. *Lecture Notes in Biomathematics* **53**, 13-213.

Nosil P, Egan SP, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: "Isolation by adaptation" and multiple roles for divergent selection. *Evolution* **62**, 316-336.

Novaes E, Drost D, Farmerie W, *et al.* (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *Bmc Genomics* **9**.

Oetjen K, Reusch TBH (2007) Genome scans detect consistent divergent selection among subtidal vs. intertidal populations of the marine angiosperm *Zostera marina*. *Molecular Ecology* **16**, 5156-5167.

Ohashi J, Naka I, Patarapotikul J, *et al.* (2004) Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *American Journal of Human Genetics* **74**, 1198-1208.

Oleksiak MF (2008) Changes in gene expression due to chronic exposure to environmental pollutants. *Aquatic Toxicology* **90**, 161-171.

Oleksiak MF (2010) Genomic approaches with natural fish populations. *Journal of Fish Biology* **76**, 1067-1093.

Oleksiak MF, Wu S, Parker C, *et al.* (2000) Identification, functional characterization, and regulation of a new cytochrome P450 subfamily, the CYP2Ns. *Journal of Biological Chemistry* **275**, 2312-2321.

Orengo DJ, Aguade M (2007) Genome scans of variation and adaptive change: Extended analysis of a candidate locus close to the phantom gene region in Drosophila melanogaster. *Molecular Biology and Evolution* **24**, 1122-1129.

Ownby DR, Newman MC, Mulvey M, *et al.* (2002) Fish (*Fundulus heteroclitus*) populations with different exposure histories differ in tolerance of creosote-contaminated sediments. *Environmental Toxicology and Chemistry* **21**, 1897-1902.

Padma TV, Hale RC, Roberts MH (1998) Toxicity of water-soluble fractions derived from whole creosote and creosote-contaminated sediments. *Environmental Toxicology and Chemistry* **17**, 1606-1610.

Pain A, Urban BC, Kai O, *et al.* (2001) A non-sense mutation in Cd36 gene is associated with protection from severe malaria. *Lancet* **357**, 1502-1503.

Pogson GH, Mesa KA, Boutilier RG (1995) Genetic Population-Structure and Gene Flow in the Atlantic Cod - *Gadus morhua* - a Comparison of Allozyme and Nuclear RFLP Loci. *Genetics* **139**, 375-385.

Pollinger JP, Bustamante CD, Fledel-Alon A, *et al.* (2005) Selective sweep mapping of genes with large phenotypic effects. *Genome Research* **15**, 1809-1819.

Porter AH (2003) A test for deviation from island-model population structure. *Molecular Ecology* **12**, 903-915.

Prince R, Cooper KR (1995a) Comparisons of the Effects of 2,3,7,8-Tetrachlorodibenzo-P-Dioxin on Chemically Impacted and Nonimpacted Subpopulations of *Fundulus-Heteroclitus*.1. TCDD Toxicity. *Environmental Toxicology and Chemistry* **14**, 579-587.

Prince R, Cooper KR (1995b) Comparisons of the Effects of 2,3,7,8-Tetrachlorodibenzo-P-Dioxin on Chemically Impacted and Nonimpacted Subpopulations of *Fundulus heteroclitus*. 2. Metabolic Considerations. *Environmental Toxicology and Chemistry* **14**, 589-595.

Raymond M, Berticat C, Weill M, Pasteur N, Chevillon C (2001) Insecticide resistance in the mosquito *Culex pipiens*: what have we learned about adaptation ? *Genetica* **112**, 287-296.

Rohrer GA, Freking BA, Nonneman D (2007) Single nucleotide polymorphisms for pig identification and parentage exclusion. *Animal Genetics* **38**, 253-258.

Ronaghi M, Elahi E (2002) Discovery of single nucleotide polymorphisms and mutations by pyrosequencing. *Comparative and Functional Genomics* **3**, 51-56.

Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**, 363.

Ronald J, Akey JM (2005) Genome-wide scans for loci under selection in humans. *Human Genomics* **2**, 113-125.

Schlotterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**, 753-763.

Schofl G, Schlotterer C (2004) Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-african D-simulans. *Molecular Biology and Evolution* **21**, 1384-1390.

Shin H, Hirst M, Bainbridge M, *et al.* (2008) Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. *BMC Biology* **6**, 30.

Stackelberg PE (1997) Presence and distribution of chlorinated organic compounds in streambed sediments, New Jersey. *Journal of the American Water Resources Association* **33**, 271-284.

Storz JF, Dubach JM (2004) Natural selection drives altitudinal divergence at the albumin locus in deer mice, *Peromyscus maniculatus*. *Evolution* **58**, 1342-1352.

Storz JF, Nachman MW (2003) Natural selection on protein polymorphism in the rodent genus *Peromyscus*: Evidence from interlocus contrasts. *Evolution* **57**, 2628-2635.

Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Molecular Biology and Evolution* **21**, 1800-1811.

Tollenaere C, Duplantier, J.-M., Rahalison, L., Ranjalahy,M., Brout, C. (2010) AFLP genome scan in the black rat (*Rattus rattus*) from Madagascar: detecting genetic markers undergoing plague-mediated selection. *Molecular Ecology*.

Tracy C, Rio J, Motiwale M, Christensen SM, Betran E Convergently recruited nuclear transport retrogenes are male biased in expression and evolving under positive selection in Drosophila. *Genetics* **184**, 1067-1076.

Trustees FNR (2007) Draft Natural Resource Damage Assessment.

Van Veld PA, Ko UC, Vogelbein WK, Westbrook DJ (1991) Glutathione-s-transferase in intestine, liver, and hepatic lesions of mummichug (*Fundulus heteroclitus*) from a creosote-contaminated environment. *Fish Physiology and Biochemistry* **9**, 369-376.

Van Veld PA, Westbrook, D.J. (1995) Evidence for depression of cytochrome P4501A in a population of chemically resistant mummichug (*Fundulus heteroclitus*). *Environmental Sciences* **3**, 221-234.

Vasemagi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Molecular Biology and Evolution* **22**, 1067-1076.

Vera JC, Wheat CW, Fescemyer HW, *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**, 1636-1647.

Vigouroux Y, McMullen M, Hittinger CT, *et al.* (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 9650-9655.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811-1823.

Volgelbein WK WC, Van Veld PA (1996) Acute toxicity resistance in a fish population with a high prevelance of cancer. Presented at SETAC Annual Meeting, Washington, DC, USA.

Vos P, Hogers R, Bleeker M, *et al.* (1995) AFLP - a New Technique for DNA-Fingerprinting. *Nucleic Acids Research* **23**, 4407-4414.

Voyer RA, Pesch C, Gather J, Copeland J, Comeleo R (2000) New Bedford, Massachusetts - A story of urbanization and ecological connections. *Environmental History* **5**, 352-377.

Watterson GA (1978) Homozygosity test of neutrality. *Genetics* **88**, 405-417.

Weaver G (1984) Pcb Contamination in and around New-Bedford, Mass. *Environmental Science & Technology* **18**, A22-A27.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiology* **144**, 32-42.

Weinig C, Dorn LA, Kane NC, *et al.* (2003) Heterogeneous Selection at Specific Loci in Natural Environments in *Arabidopsis thaliana*. *Genetics* **165**, 321-329.

Weir BS, Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358-1370.

Weis JS, Weis P (1989) Tolerance and stress in a polluted environment *Bioscience* **39**, 89-95.

Wenning RJ, Bonnevie NL, Huntley SL (1994) Accumulation of metals, polychlorinated biphenyls, and polycyclic aromatic hydrocarbons in sediments from the lower Passaic River, New Jersey *Archives of Environmental Contamination and Toxicology* **27**, 64-81.

Wiedmann R, Smith T, Nonneman D (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *Bmc Genetics* **9**, 81.

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology* **14**, 611-619.

Willing EM, Bentzen P, van Oosterhout C, *et al.* (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology* **19**, 968-984.

Wills LP, Jung D, Koehrn K, *et al.* (2010a) Comparative Chronic Liver Toxicity of Benzo[a]pyrene in Two Populations of the Atlantic Killifish (*Fundulus heteroclitus*) with Different Exposure Histories. *Environ Health Perspect*.

Wills LP, Matson CW, Landon CD, Di Giulio RT (2010b) Characterization of the recalcitrant CYP1 phenotype found in Atlantic killifish (*Fundulus heteroclitus*) inhabiting a Superfund site on the Elizabeth River, VA. *Aquatic Toxicology* **In Press, Corrected Proof**.

Wootton JC, Feng XR, Ferdig MT, *et al.* (2002) Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* **418**, 320-323.

Wright S (1951) The genetic structure of populations. *Annals of Eugenics* **15**, 323-354.

Yang M, Sanudo-Wilhelmy SA (1998) Cadmium and manganese distributions in the Hudson River estuary: interannual and seasonal variability. *Earth and Planetary Science Letters* **160**, 403-418.

CHAPTER 1

Signatures of selection in natural populations adapted to chronic pollution

Larissa M. Williams[1] and Marjorie F. Oleksiak[2]

1. Department of Environmental and Molecular Toxicology
Box 7633, North Carolina State University
Raleigh, NC 27695-7633 USA

2. Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149 USA

Corresponding Author:
Marjorie Oleksiak
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149
Fax: 305-421-4600
Email: moleksiak@rsmas.miami.edu

Published in BMC Evolutionary Biology 2008, **8**:282.

Abstract

Background: Populations of the teleost fish *Fundulus heteroclitus* appear to flourish in heavily polluted and geographically separated Superfund sites. Populations from three Superfund sites (New Bedford Harbor, MA, Newark Bay, NJ, and Elizabeth River, VA) have independently evolved adaptive resistance to chemical pollutants. In these polluted populations, natural selection likely has altered allele frequencies of loci that affect fitness or that are linked to these loci. The aim of this study was to identify loci that exhibit non-neutral behavior in the *F. heteroclitus* genome in polluted populations *versus* clean reference populations.

Results: To detect signatures of natural selection and thus identify genetic bases for adaptation to anthropogenic stressors, we examined allele frequencies for many hundreds of amplified fragment length polymorphism markers among populations of *F. heteroclitus.* Specifically, we contrasted populations from three Superfund sites (New Bedford Harbor, MA, Newark Bay, NJ, and Elizabeth River, VA) to clean reference populations flanking the polluted sites. When empirical F_{ST} values were compared to a simulated distribution of F_{ST} values, 24 distinct outlier loci were identified among pairwise comparisons of pollutant impacted *F. heteroclitus* populations and both surrounding reference populations. Upon removal of all outlier loci, there was a strong correlation ($R^2 = 0.79$, $p < 0.0001$) between genetic and geographical distance. This apparently neutral evolutionary pattern was not evident when outlier loci were included ($R^2 = 0.092$, $p = 0.0721$). Two outlier loci were shared between New Bedford Harbor and Elizabeth River populations, and two different loci were shared between Newark Bay and Elizabeth River populations.

Conclusion: In total, 1% to 6% of loci are implicated as being under selection or linked to areas of the genome under selection in three *F. heteroclitus* populations that reside in polluted estuaries. Shared loci among polluted sites indicate that selection may be acting on multiple loci involved in adaptation, and loci shared between polluted sites potentially are involved in a generalized adaptive response.

Background

The genetic basis of adaptation is a fundamental issue in evolutionary biology. Much of the research in this field has been focused on the classic model systems of *Drosophila* (Beaumont, Balding, 2004; Beaumont, Nichols, 1996; Daborn *et al.*, 2002; Feder, 1999; Hoffmann *et al.*, 2003; Kauer *et al.*, 2002; Kopp *et al.*, 2000; Nuzhdin *et al.*, 2004; Posthuma, Vanstraalen, 1993; Presgraves *et al.*, 2003; Riley *et al.*, 2003; Vitalis *et al.*, 2001; Wiehe *et al.*, 2007) and *Arabidopsis* (Fowler, Thomashow, 2002; Maloof *et al.*, 2001; Pigliucci *et al.*, 2003; Tian *et al.*, 2003; Weinig *et al.*, 2002). Recently, insight into adaptation in non-model species has become possible due to advances in molecular biology and statistics (Allendorf, Seeb, 2000; Beaumont, Nichols, 1996; Bradshaw *et al.*, 2000; Frary *et al.*, 2000; Kohn *et al.*, 2003; Mock *et al.*, 2002; Parsons, Shaw, 2001; Peichel *et al.*, 2001; Pogson *et al.*, 1995; Storz, Dubach, 2004; Storz, Nachman, 2003; Whitehead *et al.*, 2003; Yan *et al.*, 1999). This recent expansion into studies of non-model systems allows further development of evolutionary inferences (Luikart *et al.*, 2003), such as the role that selection, mutation, gene flow, and drift play in adaptation (Wang *et al.*, 2003). A powerful approach to understand genome-wide adaptation is to investigate independent natural populations that inhabit environments with strong selective pressures.

One species that has adapted to a wide range of estuarine environments is the teleost fish, *Fundulus heteroclitus* (Griffith, 1974). *F. heteroclitus* is widely distributed along the United States' eastern seaboard from the Gulf of St. Lawrence to northeastern Florida (Duvernell *et al.*, 2008). Subpopulations of *F. heteroclitus* inhabit clean estuaries as well as those heavily impacted by chemical pollutants (reviewed in (Wirgin, Waldman, 2004)).

Three well-known polluted sites where *F. heteroclitus* reside are New Bedford Harbor (Massachusetts), Newark Bay (New Jersey), and Elizabeth River (Norfolk, VA). All three sites have been identified by the Environmental Protection Agency (EPA) as Superfund sites (part of the federal government's program to clean up the nation's uncontrolled hazardous waste sites) and contain high levels of a variety of lipophilic, persistent and toxic contaminants worthy of remediation using Federal funds. All three Superfund sites are highly contaminated with chemical pollutants that are broadly classified as aromatics. New Bedford Harbor is polluted with extremely high levels of polychlorinated biphenyls (Pruell *et al.*, 1990) as well as polychlorinated dibenzo-p-dioxins (PCDD), polychlorinated dibenzofurans (PCD), polycyclic aromatic hydrocarbons (PAH), and several trace metals (Bergen *et al.*, 1998; Pruell *et al.*, 1990). Newark Bay is most notorious for containing 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) as well as other dioxins (Prince, Cooper, 1995; Weis, 2002) and also is contaminated with heavy metals, pesticides, PCBs and PAHs (Iannuzzi *et al.*, 2005). The Elizabeth River is predominantly contaminated with creosote, comprised of a complex mixture of PAHs (Bieri *et al.*, 1986; Huggett *et al.*, 1992; Padma *et al.*, 1998).

F. heteroclitus from these chronically polluted areas are resistant to the aromatic hydrocarbons in their environment as compared to nearby fish from relatively clean environments (Black *et al.*, 1998; Elskus *et al.*, 1999; Meyer, Di Giuliuo, 2003; Meyer *et al.*, 2002; Nacci *et al.*, 1999; Nacci *et al.*, 2002; Ownby *et al.*, 2002; Vogelbein *et al.*, 1990). Resistance in first and second generation embryos from New Bedford Harbor and Elizabeth River and first generation embryos from depurated Newark Bay fish suggests that differential

survival is due to genetic adaptation rather than physiological induction. Investigating and comparing *F. heteroclitus* from these three sites provides the opportunity to study similarities and differences in adaptation to differing chemical pollutant and resistance to general stress conditions among populations.

Previous work to elucidate mechanisms of resistance and the underlying genetic basis in *F. heteroclitus* from these three sites has investigated the refractory phenotype of the xenobiotic metabolizing enzyme cytochrome P4501A (CYP1A) in polluted populations (Bello *et al.*, 2001; Elskus *et al.*, 1999; Nacci *et al.*, 1999; Prince, Cooper, 1995; Van Veld, Westbrook, 1995), epigenetic silencing through CpG methylation of promoter regions of the CYP1A 5' promoter region (Timme-Laragy *et al.*, 2005), and elimination of contaminants through the induction of other phase I, II, and III enzymes ((Armknecht *et al.*, 1998; Bard *et al.*, 2002; Bello *et al.*, 2001; Cooper *et al.*, 1999)), many by way of the aryl hydrocarbon receptor (AHR) pathway (reviewed in (Hahn, 1998)). Yet, none of these research efforts has completely accounted for the differences in the resistance phenotypes between polluted and reference site fish in New Bedford Harbor, Newark Bay and Elizabeth River, nor has the genetic basis for resistance been elucidated.

In contrast to a candidate gene approach, our strategy to begin to understand the genetic mechanisms that enable *F. heteroclitus* populations to inhabit these highly polluted sites was to screen the genome for selectively important loci. The premise is that loci under selection will have patterns of variation statistically different from the majority of neutral loci (Lewontin, Krakauer, 1973). Loci that have a large difference in allele frequencies between populations with respect to what would be expected under the neutral expectation are

outliers. The identification of these outliers provides evidence for which and how many loci may be involved in the evolutionary adaptation to anthropogenic pollution.

Loci can have significantly different frequencies relative to other neutral loci for many reasons. To obviate the detection of outliers due to genetic drift rather than selection, our sampling scheme contrasted each polluted population with two reference populations that were geographically more distant from each other than either was to the polluted population. This provides a control for each Superfund site by identifying which loci are significant outliers relative to two reference sites that are demographically distant from each other. To provide extensive coverage of the genome, we used approximately 300 amplified fragment length polymorphisms (AFLP) (Vos *et al.*, 1995) to genotype 288 individuals from nine *F. heteroclitus* populations and used a modeling approach to reveal significant outliers. Furthermore, we investigated whether outlier loci were shared among polluted populations, suggesting similar patterns of selection on the genome despite differences in pollutant compositions and local conditions.

Methods

F. heteroclitus were collected using minnow traps during the spring of 2005. Fin clip samples from 32 individuals were sampled from each of the nine collection sites along the east coast of the United States (Fig. 1; Table 1). Three of the collection sites were Superfund sites: New Bedford (EPA ID: MAD980731335), Newark (EPA ID: NJD980528996), and Elizabeth River (EPA ID: VAD990710410). Two non-polluted reference sites flanked each

Superfund site, approximately equidistant on either side of each polluted site (Fig. 1; Table 1).

Genomic DNA was extracted from fin clips using a modified version of Aljanabi and Martinez ((Aljanabi, Martinez, 1997)). Fin clips were incubated at 55º C for two hours in 300 µL of 75 mM NaCl, 25 mM EDTA, and 1% SDS with Proteinase K (3 µL of 20 mg/mL). Following incubation, 0.5 volumes of 7.5M ammonium acetate were added and DNA was precipitated on ice with the addition of 0.7 volumes of isopropanol. Subsequently, DNA was pelleted through centrifugation and washed with 70% ethanol. DNA was resuspended overnight at 4º C in 0.1x TE.

The AFLP analysis was performed in replicate following the ligation of the DNA for each individual using a modified version of Vos *et al*. ((Vos *et al.*, 1995)) to generate approximately 300 loci. Genomic DNA (500 ng) was digested with 5U EcoRI (New England Biolabs, MA) and 5U MseI (New England Biolabs, MA) overnight at 37º C in a total volume of 45 µL containing 1X T4 DNA ligase buffer (Epicentre) supplemented with 100 µg/mL BSA. Following incubation, 50 pmol adaptor oligonucleotides (Applied Biosystems) and 1U T4 DNA ligase (Epicentre) were added and incubated overnight at 16º C. Preselective PCRs were performed in a 15 µL volume using 5 µL of diluted (1:10) ligation product with EcoRI + (C/A) primer (Integrated DNA Technologies; 10 pmol), MseI + (C/A) primer (Integrated DNA Technologies; 10 pmol) and 1U *Taq*. PCR conditions were 20 cycles of 94º C for 10 sec, 56º for 30 sec, and 72º C for 2 min. Selective Eco + 3NT primers (Integrated DNA Technologies; 10 pmol) labeled with FAM dye at the 5' end and MseI + 3NT primers (Integrated DNA Technologies; 10 pmol) were added to diluted (1:10)

pre-selective PCR product in a 15 µL volume. PCR conditions in the first cycle were 94º C for 10 sec, 65º C for 30 sec, and 72º C for 2 min with the annealing temperature reduced by 0.4º C for 12 cycles, then 30 cycles of 94º C for 10 sec, 56º C for 30 sec, and 72º C for 2 min. Semi-automated analysis of the selective PCR products was performed on MegaBACE 1000 DNA sequencing system (GE Healthcare). Peak patterns were calculated using MegaBACE Geneprofiler software v. 1.0 (GE Healthcare). The criteria for distinct peaks were a size between 50 and 400 base pairs and an absolute intensity greater than or equal to 1000. Replicated fragments were obtained from all samples (the same template was used for independent PCRs) and replicate fragments were scored as being present or absent using Peakmatcher software (DeHaan *et al.*, 2002). Peakmatcher software automatically creates marker categories and generates a binary table for the presence and absence of markers based on the minimum 75 percent repeatability of markers across replicates.

Statistical Analysis

The frequency of band presence allele was calculated using the formula $P = 1- ((N - C)/N)^{0.5}$ where N equals the sample size and C is the number of individuals with the band (Wilding *et al.*, 2001a). This formula assumes Hardy-Weinberg equilibrium. However, because AFLPs are dominant markers and heterozygotes are not observed, Hardy-Weinberg equilibrium cannot be directly tested. Due to strong selection or increased mutational rates, some of the loci may not be in Hardy-Weinberg equilibrium. Though not directly comparable, microsatellites are in Hardy-Weinberg equilibrium in these *F. heteroclitus* populations (Adams *et al.*, 2006). This calculation also assumes that shared band presence or absence between two individuals is due to common evolutionary origin and not homoplasy.

Pairwise F_{ST} values between populations were calculated for each locus by the method of Nei (Nei, 1977) with the correction of Nei and Chesser (Nei, Chesser, 1983) for finite sample sizes, and a null distribution of F_{ST} values *versus* allele frequency was simulated using the Winkles program ((Wilding *et al.*, 2001b), Fig. 2).

Winkles is based on the model described in Beaumont and Nichols ((Beaumont, Nichols, 1996b)) which employs coalescent simulations using the Island model and an infinite alleles mutational model. Samples of the same size and number as the data are simulated, where each sample is taken from a different island. This simulation uses two populations of size *N* diploid individuals, with a set mutation rate, μ, and a migration rate, *m*, per generation. Parameters for the simulation are estimated through the calculation $F_{ST} = 1/(1 + 16Nm + 16N\square)$. The F_{ST} value is found by calculating the mean F_{ST} from any given pairwise comparison and adjusting that value by -0.0093 to account for the upward bias in the model reported by Wilding et al. ((Wilding *et al.*, 2001b)); this bias is consistent with previous simulations using Nei's methods to calculate pairwise F_{ST} values (Slatkin, Barton, 1989). The *Nm* factor is calculated by solving for that parameter in the above equation. Each simulation used 10^3 and 10^{-4} as estimates of *N* and μ, respectively. Simulated F_{ST} values are relatively unaffected by changing either the sample size of the simulated population or the mutation rate (Beaumont, Nichols, 1996b). Five simulations were run on each pairwise comparison to generate an expected null distribution of 25,000 values. Each simulation started with 500 simulation bi-allelic loci in each of the two populations with uniform random distribution and was allowed to drift for $10N$ generations. The 99^{th} percentile of F_{ST} values within each of the 40 binned mean allele frequency values (each bin

representing a set of 0.025 frequency values from 0 to 1) was calculated after removing monomorphic loci because F_{ST} is strongly dependent on allele frequencies (Beaumont, Nichols, 1996b).

The model we used (Beaumont, Nichols, 1996b) is robust to a wide range of alternative models such as colonization and stepping-stone (Vitalis *et al.*, 2001). It is likely to detect outliers with unusually high F_{ST} values and will identify adaptive selection at one or many loci through pairwise comparisons of populations (Beaumont, Balding, 2004; Vitalis *et al.*, 2001). This model is not able to identify loci under balancing selection and tends to generate discrepancies when numbers of immigrants per generation are unequal, the true population history consists of repeated branching events, or the connectivity of populations is uneven (Vitalis *et al.*, 2001). Isolation, population bottlenecks, and populations which are heterogeneous with respect to their demographic parameters further bias to the model (Beaumont, Nichols, 1996b). There is no evidence for isolation and bottleneck history (Adams *et al.*, 2006) or reduced genetic diversity (McMillan *et al.*, 2006) in our populations. However, if non-homogenous demographic parameters exist (*e.g.*, skewed age structure or sex ratios), this model may be biased. Given the relative robustness of the model to identify loci under adaptive selection, we used theoretical *versus* experimentally derived allele frequencies for loci to determine significant deviations from the neutral expectation.

Results

Total number of loci among populations

Five different primer combinations (Table 2) were used to amplify approximately 300 loci from 288 individuals from nine different *F. heteroclitus* populations. Among New Bedford Harbor and its reference sites, Sandwich and Point Judith, a total of 296 loci were scored. Of those 296 loci, 11 bands were found to be monomorphic (3.7%). Newark and its two reference sites, Tuckerton and Clinton, had a total of 336 loci, of which 7 loci were monomorphic (2.1%). Elizabeth River and its two reference sites, Magotha and Manteo, had a total of 299 loci, with 4 loci found to be monomorphic (1.3%). Among all populations, 450 distinct loci were scored.

Outlier loci among populations

In comparisons of the three Superfund sites and their clean reference sites, twenty-four loci show patterns indicative of selection. The criteria for identifying these selective loci are that they were identified as outliers in pairwise comparisons of each Superfund site population relative to its two reference site populations (polluted *versus* both references, analyzed separately, i.e. the union of polluted *versus* reference 1 and polluted *versus* reference 2) but not in comparisons between the reference site populations. Eighteen of these twenty-four loci were found in the New Bedford Harbor comparisons, four were found in the Newark Bay comparisons, and six were found in the Elizabeth River comparisons (Fig. 3). Four of these loci were shared between two Superfund site populations suggesting conserved mechanisms of adaptation (Fig. 4).

In the northern most Superfund site, New Bedford Harbor, 42 loci representing 14% of total analyzed loci were located above the simulated 0.99 quantile in the polluted *versus* one of the references' comparisons. That is, these 42 loci have F_{ST} values that lie outside the expected neutral distribution of 99% of all loci. This is more than 10 fold greater than the 3 that are expected by chance from the approximately 300 amplified loci. These 42 loci are outliers in the New Bedford Harbor comparison to the Point Judith, RI reference population (36 loci), the Sandwich, MA reference population (23 loci) or relative to both reference sites (18 loci). The 18 outlier loci found in the comparisons of New Bedford Harbor to both of its reference populations were amplified from three different primer combinations, spanning a 100 base pair range (Fig. 3A). The joint probability (<0.01 squared or <0.0001) indicates that less than one locus should be different in both clean sites *versus* the Superfund site. These 18 loci are thus implicated as separate loci under selection or linked to areas of the genome under selection. There are 16 loci that are outliers when comparing the two reference populations to each other. Only one of these 16 outlier loci is specific to the clean reference sites; the other 15 are also found in the comparison to the New Bedford Harbor Superfund site to one of these reference sites. No locus was an outlier in all pairwise comparisons.

Newark Bay, NJ is close to the phylogeographic boundary that separates northern and southern populations of *F. heteroclitus* (Adams *et al.*, 2006; Smith *et al.*, 1998). The Clinton reference population is on the northern side and the Tuckerton reference population is on the southern side. The Newark Bay Superfund site has 26 outlier loci (8% *versus* 1% expected) relative to these two reference sites: 18 (5%) in the comparison with the Clinton reference

population and 13 (4%) in the comparison with the Tuckerton reference site population. Four outlier loci are found in both comparisons between the Newark Bay Superfund site and its two clean reference sites (Fig. 3B) and not among clean sites. These four loci are greater than that predicted from the joint probability of differences in both clean sites *versus* the Superfund site. In pairwise comparisons of the two clean reference sites, 18 loci are outliers. Ten of these 18 loci are common outliers between a northern and two different southern populations *i.e.*, Clinton and Newark Bay populations and Clinton and Tuckerton populations.

Elizabeth River is the most southern Superfund site. The Elizabeth River population, in comparisons to its two reference site populations, had 9 outlier loci (3%). The Elizabeth River and Magotha reference site comparison had 8 outlier loci (2.7% of the total loci) whereas the Elizabeth River and Manteo reference site comparison had 7 (2.4% of the total loci). Six outlier loci were found in both comparisons (Fig. 3C) and not found in the comparison among clean sites. Among the two reference sites (Magotha and Manteo) only three loci were outliers and none of these were unique to the reference-reference comparison. Two loci were in common with outliers from the Elizabeth River-Magotha comparison and one locus was in common with the Elizabeth River-Manteo comparison.

Among the twenty-three loci that were outliers in comparisons only among Superfund sites and both reference sites, four loci are outliers in two of the three Superfund sites (Fig. 4; Table 3). Two of these four outlier loci are shared between New Bedford Harbor and Elizabeth River populations, and two are shared between Newark Bay and Elizabeth River.

None is shared between New Bedford and Newark Bay, nor are any shared among all three Superfund site populations.

F_{ST} values were calculated for comparisons between all sites with and without outlier loci (Table 4). As would be expected, average F_{ST} values were higher in all comparisons before the removal of the outliers. The average F_{ST} value (with outliers) between New Bedford Harbor and its reference sites is 0.038, between Newark and its reference sites it is 0.039, and between Elizabeth River and its reference sites it is 0.018. Upon removal of the outliers, average F_{ST} values fall to 0.010, 0.016, and 0.011 for New Bedford Harbor, Newark Bay, and Elizabeth River, respectively. These values were plotted against log-ten of geographic distance between sites *versus* genetic distance [$F_{ST}/(1- F_{ST})$, (Rousset, 1997)]. There is no apparent pattern in the distribution of pairwise comparisons corresponding to reference-reference, polluted-reference, or polluted-polluted sites. When outliers were included in the calculation of average F_{ST} and plotted against distance, there was no significant linear relationship ($R^2 = 0.092$, $p = .0721$). Upon removal of the outliers, there was a significant and strong linear relationship ($R^2 = 0.79$, $p < 0.0001$) between geographic and genetic distance (Fig. 5). Mantel tests that account for multiple comparisons confirmed the significance of both relationships (data not shown). This relationship indicates that 79% of the variability in the neutral genetic distance (without outlier loci) between sites can be explained by geographic distance.

Discussion

Multiple *F. heteroclitus* populations have independently evolved adaptive resistance to complex suites of pollutants (Black *et al.*, 1998; Elskus *et al.*, 1999; Gutjahr-Gobell *et al.*, 1999; Meyer, Di Giuliuo, 2003; Meyer *et al.*, 2002; Nacci *et al.*, 1999; Nacci *et al.*, 2002; Ownby *et al.*, 2002; Powell *et al.*, 2000; Vogelbein *et al.*, 1990). These different populations provide independent contrasts for identifying loci involved in adaptation. We identified loci suggestive of adaptation for each polluted population by identifying outlier loci in the polluted population relative to two nearby reference populations. These loci are outliers because they are statistically different from the neutral distribution among populations. Only loci exhibiting a non-neutral distribution in comparisons of the polluted population *versus* both a north and south reference population were considered to be adaptive. Through this comparison, we are more likely to identify loci whose non-neutral distribution is due to pollution rather than geography. Similarly, while the model used to identify outlier loci has a false positive rate of approximately 7% (Beaumont, Balding, 2004), it is unlikely that the same loci will be falsely identified in multiple comparisons (*i.e.*, in the polluted population *versus* both a north and south reference populations). In each of the Superfund sites, 1% to 6% (four to 18 loci out of approximately 300) of amplified fragments were identified as being loci under selection or linked to areas of the genome under selection. Four of these loci were outliers in two separate Superfund population comparisons.

We only consider loci exhibiting a non-neutral distribution in comparisons of the polluted population *versus* both a north and south reference population to be adaptive. These populations make up a geographic triangle formed among the northern and southern clean

reference populations and a latitudinally intermediate polluted population (Fig. 1). This double comparison ensures that we are not identifying loci that differ simply due to genetic drift or clinal variation common to this species. This contrast, in addition to the joint comparison among populations, address most of the possible neutral or demographic models. Population isolation can alter allele frequencies among populations. One would expect that a single population that suffered from unique isolation would have significantly greater F_{ST} values among many loci in comparison to similarly geographically distance populations that were not uniquely isolated. This demographic explanation does not fit the data for two reasons: 1) it is the statistically different F_{ST} value for a few loci in comparison to all other loci that we define as being important, and 2) all non-outlier loci follow the more common demographic trend of isolation by distance (Fig. 5). However, differences in F_{ST} values also can result if loci under functional constraints evolve more slowly than loci without functional constraints. Thus, loci with large F_{ST} values would have few, if any constraints, relative to the hundreds of other AFLP loci. However, our comparisons were based on both a significant F_{ST} between both reference sites *versus* a polluted site **and** insignificant differences among reference sites (as well as a difference from the permutation model, see methods). Because we are using three criteria (significant difference *versus* the joint distribution in two reference sites, lack of a difference among reference sites, and a statistical difference from a neutral permutation model), it seems most parsimonious to suggest that these outlier loci are due to natural selection. However, lack of Hardy-Weinberg equilibrium or recent mutations also might cause loci to be outliers. We suggest that the most obvious cause for this evolved difference is chronic exposure to the aromatic hydrocarbons and other

anthropogenic pollutants; yet, we cannot explicitly control every variable in natural environments. Other selective forces also could be different between the three sites. For instance, site complexity differs among the nine sites with the three polluted sites tending to be less complex (have less edges) than the reference sites. Thus, predation or food availability might differ among sites. Similarly, salinity might affect food availability or absorption, and although all populations inhabit brackish waters, the Elizabeth River population is less coastal than the reference populations to which it is compared. Under controlled laboratory conditions, survival differs among fish from clean populations exposed to polluted sediments and fish from polluted populations exposed to clean sediments. This phenomenon points towards adaptation to anthropogenic contaminants rather than differing local conditions for the differences seen between polluted and reference populations. Thus we postulate that outlier loci are due to pollution, especially those loci shared among separate Superfund populations.

Most of the outlier loci are unique to a single polluted population rather than shared across polluted populations (Fig. 4). One explanation for the lack of shared loci is that different loci are involved in the adaptation to a particular pollutant or stress. Alternatively, some of these outliers might be linked to the same locus in the different populations and only appear to be different because the locus under selection dragged different polymorphisms to fixation. This could occur because different polymorphisms existed in the different ancestral populations.

Resistance to pollution is a modern phenotype in *F. heteroclitus* due to recent exposure (approximately within the last 60 years), suggesting that *F. heteroclitus* have rapid

evolutionary responses with respect to their environment. Our data and other data on survival and development indicate that populations of *Fundulus* have adapted to local pollutants and thus selection has favored a few alleles. Resistance phenotypes resulting from rapid evolution have been well documented in plants (Forbes, 1999) and benthic invertebrates (Klerks, Levinton, 1989) in response to metals as well as in insects in response to pesticides (McKenzie, 1996) and depend both on population dynamics as well as the strength of selection. *F. heteroclitus* populations residing in chronically polluted areas provide an advantageous situation whereby strong selective pressures and rapid evolution can be studied. *F. heteroclitus* have high standing genetic variation (Mitton, Koehn, 1975), high reproductive potential (Weis, 2002), limited home ranges (Lotrich, 1975) and large population sizes exceeding 10,000 in a single tidal creek (Adams *et al.*, 2006). These attributes can and have resulted in locally adapted *F. heteroclitus* populations. Adaptation due to positive selection often reduces genetic variation among natural populations because of selective sweeps. For example, reduced genetic variation has occurred in brown rats resistant to the rodenticide, warfarin (Kohn, Pelz, 1999; Kohn *et al.*, 2000; Kohn *et al.*, 2003), tobacco budworm exposed to the pyrethroid insecticide (Taylor *et al.*, 1995), and the human malarial parasite, *Plasmodium falciparum*, exposed to antimalarial agents (Wootton *et al.*, 2002). However, genetic diversity is not reduced in the polluted *F. heteroclitus* populations compared to the reference site populations for either neutral markers (McMillan *et al.*, 2006; Mulvey *et al.*, 2003; Roark *et al.*, 2005) or gene expression (Fisher, Oleksiak, 2007). Maintenance of genetic diversity in these populations subjected to significant selection most likely represents steady influx of alternative alleles by migration. If migration

and resulting gene flow is strong enough to prevent the reduction of genetic diversity at non-selected loci, it suggests that selection at adaptively important loci is equally strong. Importantly, with constant influx of allelic variation at loci without adaptive value, there should be fewer spurious allelic differences among populations. Thus, shared loci between Superfund populations are likely to be affected by selection and therefore biologically important.

Among three *F. heteroclitus* populations inhabiting highly polluted Superfund sites and flanking reference populations, 63 different loci (14% of the collective 450 loci) have F_{ST} values outside the 99% quantile. Using all loci (*i.e.*, including outliers) our F_{ST} values based on AFLP (0.038, 0.039, and 0.018 for New Bedford Harbor, Newark Bay and Elizabeth River, respectively) are approximately one-half of those found for microsatellites (0.077, 0.068, and 0.043, respectively (Adams *et al.*, 2006)) although these genetic measures are difficult to compare due to differences in genomic coverage and mutation rates (Gaudeul *et al.*, 2004). Using AFLPs, McMillan *et al.* ((McMillan *et al.*, 2006) found similar F_{ST} values for the New Bedford Harbor population (0.056). For the Elizabeth River population, Mulvey *et al.* ((Mulvey *et al.*, 2003)) also found similar F_{ST} values (0.014) using allozymes. Notice that these calculated F_{ST} values use all loci and do not distinguish between neutral and non-neutral loci. If selection affects the frequency of alleles among these molecular markers, the perceived genetic distance (F_{ST}) will be exaggerated.

The neutral hypothesis is a powerful tool to explore differences among populations (Kreitman, 1996). However, in order to test evolutionary hypotheses, one needs to distinguish between neutral and non-neutral loci. Among populations for each Superfund

site, the genetic distances among local populations are affected by the outlier loci. New Bedford Harbor and Newark Bay populations are more differentiated in comparison to their reference site populations than the Elizabeth River populations (F_{ST} values of 0.038 and 0.039 *versus* 0.018) because the Elizabeth River population has the fewest outlier loci (2.4% - 2.7%) in comparison to neutral loci. These differences among Superfund sites do not exist upon removal of outliers: F_{ST} values among loci without outlier values are similar for New Bedford Harbor, Newark Bay and Elizabeth River (0.01, 0.016, and 0.011, respectively). With outliers, there is no relationship between F_{ST} values and geographic distance. However, upon removal of outlier loci, there is a strong relationship between genetic and geographical distance indicating an equilibrium model of isolation-by-distance. Similar findings have been shown in other *F. heteroclitus* studies (Adams *et al.*, 2006; Roark *et al.*, 2005), with the intertidal snail (Wilding *et al.*, 2001b), and sea trout (Hansen, Mensberg, 1998). Not surprisingly, these data indicate that loci with unusually large F_{ST} values have a large and potentially misleading effect on the perceived genetic distance among populations. The 63 outliers exhibit this effect; once removed from the data set, the neutral expectation of increasing genetic distance with geographic distance holds true. For twenty-four of these outlier loci, this non-neutral distribution is most likely caused by evolution by natural selection due to pollution or another strong selective force unique to the polluted sites since the geographical effect was taken into account through the comparison of the polluted sites with both a north and south reference population. Ten other loci have a larger than expected distance at the north-south phylogenetic boundary and likely reflect the historic split among northern and southern *F. heteroclitus* populations (Cashon *et al.*, 1981; Gonzalezvillasenor,

Powers, 1990; Ropson *et al.*, 1990). Outlier loci in reference-reference pairwise comparisons likely reflect genetic drift although some may be due to selection. While we can only speculate why these and the remaining 29 loci affect the relationship between genetic and geographic distance, this illustrates the need to distinguish among potentially selected and neutral loci to determine expected differences and posit hypotheses.

Conclusions

Contrasting populations that experience different selective pressures provides insight into evolution by natural selection. Our goal is to understand the genetic basis of adaptive resistance to pollution in chronically contaminated natural populations. Future analyses will address whether polymorphisms between populations are functional and potentially responsible for conferring resistance in populations adapted to chronic exposure to chemical pollutants in the different Superfund sites. We have shown that between 1 to 6% of loci are implicated as being under selection or linked to areas of the genome under selection in three distinct *F. heteroclitus* populations that reside in polluted Superfund estuaries. Shared loci affected by natural selection among polluted sites indicate that there may be a similar mechanism of resistance in these different populations. This study suggests that multiple loci may be involved in adaptation and a few of these loci have a generalized adaptive response.

Acknowledgements

The Authors thank G. Bozinovic for assistance in field collections and D. Crawford for valuable comments on an earlier version of this manuscript. Funding was partially provided by NIEHS Training Grant ES007046 award from the Department of Environmental and Molecular Toxicology at North Carolina State University to LMW and NIH 5 RO1 ES011588 to MFO.

References

Adams SM, Lindmeier JB, Duvernell DD (2006) Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. *Mol Ecol* **15**, 1109-1123.

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res* **25**, 4692-4693.

Allendorf FW, Seeb LW (2000) Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. *Evolution Int J Org Evolution* **54**, 640-651.

Armknecht SL, Kaattari SL, Van Veld PA (1998) An elevated glutathione S-transferase in creosote-resistant mummichog (*Fundulus heteroclitus*). *Aquatic Toxicology* **41**, 1-16.

Bard SM, Bello SM, Hahn ME, Stegeman JJ (2002) Expression of P-glycoprotein in killifish (*Fundulus heteroclitus*) exposed to environmental xenobiotics. *Aquatic Toxicology* **59**, 237-251.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**, 969-980.

Beaumont MA, Nichols RA (1996a) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences* **263**, 1619-1626.

Bello SM, Franks DG, Stegeman JJ, Hahn ME (2001) Acquired resistance to Ah receptor agonists in a population of Atlantic killifish (*Fundulus heteroclitus*) inhabiting a

marine superfund site: In vivo and in vitro studies on the inducibility of xenobiotic metabolizing enzymes. *Toxicological Science* **60**, 77-91.

Bergen BJ, Rahn K, Nelson WG (1998) Remediation at a Marine Superfund Site: Surficial Sediment PCB Congener Concentration, Composition and Redistribution. *Environmental Science and Technology* **32**, 3496-3501.

Bieri R, Hein C, Huggett R, Shou P, Slone H (1986) Polycyclic Aromatic Hydrocarbons in Surface Sediments from the Elizabeth River Subestuary. *International Journal of Environmental Analytical Chemistry* **26**, 97-113.

Black DE, Gutjahr-Gobell R, Pruell RJ, *et al.* (1998) Reproduction and polychlorinated biphenyls in *Fundulus heteroclitus* (Linnaeus) from New Bedford Harbor, Massachusetts, USA. *Environmental Toxicology and Chemistry* **17**, 1405-1414.

Bradshaw HD, Ceulemans R, Davis J, Stettler R (2000) Emerging model systems in plant biology: Poplar (Populus) as a model forest tree. *Journal of Plant Growth Regulation* **19**, 306-313.

Cashon RE, Vanbeneden RJ, Powers DA (1981) Biochemical Genetics of *Fundulus heteroclitus* (L) .4. Spatial Variation in Gene-Frequencies of Idh-a, Idh-B, 6-Pgdh-a, and Est-S. *Biochemical Genetics* **19**, 715-728.

Cooper PS, Vogelbein WK, Van Veld PA (1999) Altered expression of the xenobiotic transporter P-glycoprotein in liver and liver tumours of mummichog (*Fundulus heteroclitus*) from a creosote-contaminated environment. *Biomarkers* **4**, 48-58.

Daborn PJ, Yen JL, Bogwitz MR, *et al.* (2002) A single P450 allele associated with insecticide resistance in Drosophila. *Science* **297**, 2253-2256.

DeHaan LR, Belina RAK, Ehlke NJ (2002) Peakmatcher: Software for semi-automated fluorescence-based AFLP. *Crop Science* **42**, 1361-1364.

Duvernell DD, Lindmeier JB, Faust KE, Whitehead A (2008) Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, *Fundulus heteroclitus*. *Mol Ecol* **17**, 1344-1360.

Elskus AA, Monosson E, McElroy AE, Stegeman JJ, Woltering DS (1999) Altered CYP1A expression in *Fundulus heteroclitus* adults and larvae: a sign of pollutant resistance? *Aquatic Toxicology* **45**, 99-113.

Feder ME (1999) Engineering candidate genes in studies of adaptation: The heat-shock protein Hsp70 in *Drosophila melanogaster*. *American Naturalist* **154**, S55-S66.

Fisher MA, Oleksiak MF (2007) Convergence and divergence in gene expression among natural populations exposed to pollution. *BMC Genomics* **8**, 108.

Forbes VE (1999) *Genetics and ecotoxicology* Taylor & Francis, Philadelphia, PA.

Fowler S, Thomashow MF (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *Plant Cell* **14**, 1675-1690.

Frary A, Nesbitt TC, Grandillo S, *et al.* (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85-88.

Gaudeul M, Till-Bottraud I, Barjon F, Manel S (2004) Genetic diversity and differentiation in *Eryngium alpinum* L. (Apiaceae): comparison of AFLP and microsatellite markers. *Heredity* **92**, 508-518.

Gonzalezvillasenor LI, Powers DA (1990) Mitochondrial-DNA Restriction-Site Polymorphisms in the Teleost *Fundulus heteroclitus* Support Secondary Intergradation. *Evolution* **44**, 27-37.

Griffith RW (1974) Environmental and salinity tolerance in the genus *Fundulus*. *Copeia* **2**, 319-331.

Gutjahr-Gobell RE, Black DE, Mills LJ, *et al*. (1999) Feeding the mummichog (*Fundulus heteroclitus*) a diet spiked with non-ortho- and mono-ortho-substituted polychlorinated biphenyls: Accumulation and effects. *Environmental Toxicology and Chemistry* **18**, 699-707.

Hahn ME (1998) Mechanisms of innate and acquired resistance to dioxin-like compounds. *Rev Toxicol* **2**, 395-443.

Hansen MM, Mensberg KLD (1998) Genetic differentiation and relationship between genetic and geographical distance in Danish sea trout (*Salmo trutta* L.) populations. *Heredity* **81**, 493-504.

Hoffmann AA, Sorensen JG, Loeschcke V (2003) Adaptation of Drosophila to temperature extremes: bringing together quantitative and molecular approaches. *Journal of Thermal Biology* **28**, 175-216.

Huggett R, Van Veld P, Smith C, Hargis W, Vogelbein W (1992) *The Effects of Contaminated Sediments in the Elizabeth River* Lewis Publishers, Boca Raton.

Iannuzzi TJ, Armstrong TN, Thelen JB, Ludwig DF, Firstenberg CE (2005) Characterization of chemical contamination in shallow-water estuarine habitats of an industrialized river. Part 1: Organic compounds. *Soil & Sediment Contamination* **14**, 13-33.

Kauer M, Zangerl B, Dieringer D, Schlotterer C (2002) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics* **160**, 247-256.

Klerks PL, Levinton JS (1989) Rapid Evolution of Metal Resistance in a Benthic Oligochaete Inhabiting a Metal-Polluted Site. *Biological Bulletin* **176**, 135-141.

Kohn MH, Pelz HJ (1999) Genomic assignment of the warfarin resistance locus, Rw, in the rat. *Mamm Genome* **10**, 696-698.

Kohn MH, Pelz HJ, Wayne RK (2000) Natural selection mapping of the warfarin-resistance gene. *Proc Natl Acad Sci U S A* **97**, 7911-7915.

Kohn MH, Pelz HJ, Wayne RK (2003) Locus-specific genetic differentiation at Rw among warfarin-resistant rat (*Rattus norvegicus*) populations. *Genetics* **164**, 1055-1070.

Kopp A, Duncan I, Godt D, Carroll SB (2000) Genetic control and evolution of sexually dimorphic characters in *Drosophila*. *Nature* **408**, 553-559.

Kreitman M (1996) The neutral theory is dead. Long live the neutral theory. *Bioessays* **18**, 678-683; discussion 683.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175-195.

Lotrich VA (1975) Summer Home Range and Movements of *Fundulus heteroclitus* (*Pisces cyprinodontidae*) in a Tidal Creek. *Ecology* **56**, 191-198.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* **4**, 981-994.

Maloof JN, Borevitz JO, Dabi T, *et al.* (2001) Natural variation in light sensitivity of *Arabidopsis*. *Nat Genet* **29**, 441-446.

McKenzie JA (1996) *Ecological and evolutionary aspects of insecticide resistance* R.G. Landes, Austin, Tex.

McMillan AM, Bagley MJ, Jackson SA, Nacci DE (2006) Genetic diversity and structure of an estuarine fish (*Fundulus heteroclitus*) indigenous to sites associated with a highly contaminated urban harbor. *Ecotoxicology* **15**, 539-548.

Meyer JN, Di Giuliuo RT (2003) Heritable adaptation and fitness costs in killifish (*Fundulus heteroclitus*) inhabiting a polluted estuary. *Ecological Applications* **13**, 490-503.

Meyer JN, Nacci DE, Di Giulio RT (2002) Cytochrome P4501A (CYP1A) in killifish (*Fundulus heteroclitus*): Heritability of altered expression and relationship to survival in contaminated sediments. *Toxicological Sciences* **68**, 69-81.

Mitton JB, Koehn RK (1975) Genetic organization and adaptive response of allozymes to ecological variables in *Fundulus heteroclitus*. *Genetics* **79**, 97-111.

Mock KE, Theimer TC, Rhodes OE, Jr., Greenberg DL, Keim P (2002) Genetic variation across the historical range of the wild turkey (*Meleagris gallopavo*). *Mol Ecol* **11**, 643-657.

Mulvey M, Newman MC, Vogelbein WK, Unger MA, Ownby DR (2003) Genetic structure and mtDNA diversity of *Fundulus heteroclitus* populations from polycyclic aromatic hydrocarbon-contaminated sites. *Environ Toxicol Chem* **22**, 671-677.

Nacci D, Coiro L, Champlin D, *et al.* (1999) Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. *Marine Biology* **134**, 9-17.

Nacci DE, Champlin D, Coiro L, McKinney R, Jayaraman S (2002) Predicting the occurrence of genetic adaptation to dioxinlike compounds in populations of the estuarine fish *Fundulus heteroclitus*. *Environ Toxicol Chem* **21**, 1525-1532.

Nei M (1977) F-statistics and analysis of gene diversity in subdivided populations. *Ann Hum Genet* **41**, 225-233.

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. *Ann Hum Genet* **47**, 253-259.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM (2004) Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol Biol Evol* **21**, 1308-1317.

Ownby DR, Newman MC, Mulvey M, *et al.* (2002) Fish (*Fundulus heteroclitus*) populations with different exposure histories differ in tolerance of creosote-contaminated sediments. *Environ Toxicol Chem* **21**, 1897-1902.

Padma T, Hale R, Roberts M (1998) Toxicity of water-soluble fractions derived from whole creosote and creosote-contaminated sediments. *Environmental Toxicology and Chemistry* **17**, 1606-1610.

Parsons YM, Shaw KL (2001) Species boundaries and genetic diversity among Hawaiian crickets of the genus *Laupala* identified using amplified fragment length polymorphism. *Mol Ecol* **10**, 1765-1772.

Peichel CL, Nereng KS, Ohgi KA, *et al.* (2001) The genetic architecture of divergence between threespine stickleback species. *Nature* **414**, 901-905.

Pigliucci M, Pollard H, Cruzan MB (2003) Comparative studies of evolutionary responses to light environments in Arabidopsis. *Am Nat* **161**, 68-82.

Pogson GH, Mesa KA, Boutilier RG (1995) Genetic population structure and gene flow in the Atlantic cod *Gadus morhua*: a comparison of allozyme and nuclear RFLP loci. *Genetics* **139**, 375-385.

Posthuma L, Vanstraalen NM (1993) Heavy-Metal Adaptation in Terrestrial Invertebrates - a Review of Occurrence, Genetics, Physiology and Ecological Consequences. *Comparative Biochemistry and Physiology C-Pharmacology Toxicology & Endocrinology* **106**, 11-38.

Powell WH, Bright R, Bello SM, Hahn ME (2000) Developmental and tissue-specific expression of AHR1, AHR2, and ARNT2 in dioxin-sensitive and -resistant populations of the marine fish *Fundulus heteroclitus*. *Toxicol Sci* **57**, 229-239.

Presgraves DC, Balagopalan L, Abmayr SM, Orr HA (2003) Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* **423**, 715-719.

Prince R, Cooper KR (1995) Comparisons of the Effects of 2,3,7,8-Tetrachlorodibenzo-P-Dioxin on Chemically Impacted and Nonimpacted Subpopulations of *Fundulus heteroclitus*. 1. TCDD Toxicity. *Environmental Toxicology and Chemistry* **14**, 579-587.

Pruell R, Norwood C, Bowen R, *et al.* (1990) Geochemical study of sediment contamination in New Bedford Harbor, Massachusetts. *Marine Environmental Research* **29**, 77-101.

Riley RM, Jin W, Gibson G (2003) Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in Drosophila. *Mol Ecol* **12**, 1315-1323.

Roark SA, Nacci D, Coiro L, Champlin D, Guttman SI (2005) Population genetic structure of a nonmigratory estuarine fish (*Fundulus heteroclitus*) across a strong gradient of polychlorinated biphenyl contamination. *Environ Toxicol Chem* **24**, 717-725.

Ropson IJ, Brown DC, Powers DA (1990) Biochemical Genetics of *Fundulus heteroclitus* (L). Geographical Variation in the Gene-Frequencies of 15 Loci. *Evolution* **44**, 16-26.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219-1228.

Slatkin M, Barton NH (1989) A Comparison of 3 Indirect Methods for Estimating Average Levels of Gene Flow. *Evolution* **43**, 1349-1368.

Smith MW, Chapman RW, Powers DA (1998) Mitochondrial DNA analysis of Atlantic Coast, Chesapeake Bay, and Delaware Bay populations of the teleost *Fundulus heteroclitus* indicates temporally unstable distributions over geologic time. *Molecular Marine Biology and Biotechnology* **7**, 79-87.

Storz JF, Dubach JM (2004) Natural selection drives altitudinal divergence at the albumin locus in deer mice, *Peromyscus maniculatus*. *Evolution Int J Org Evolution* **58**, 1342-1352.

Storz JF, Nachman MW (2003) Natural selection on protein polymorphism in the rodent genus *Peromyscus*: evidence from interlocus contrasts. *Evolution Int J Org Evolution* **57**, 2628-2635.

Taylor MFJ, Shen Y, Kreitman ME (1995) A Population Genetic Test of Selection at the Molecular-Level. *Science* **270**, 1497-1499.

Tian D, Traw MB, Chen JQ, Kreitman M, Bergelson J (2003) Fitness costs of R-gene-mediated resistance in *Arabidopsis thaliana*. *Nature* **423**, 74-77.

Timme-Laragy AR, Meyer JN, Waterland RA, Di Giulio RT (2005) Analysis of CpG methylation in the killifish CYP1A promoter. *Comp Biochem Physiol C Toxicol Pharmacol* **141**, 406-411.

Van Veld PA, Westbrook DJ (1995) Evidence for depression of cytochrome P4501A in a population of chemically resistant mummichog (*Fundulus heteroclitus*). *Environ. Sci.* **3**, 221-234.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811-1823.

Vogelbein WK, Fournie JW, Vanveld PA, Huggett RJ (1990) Hepatic Neoplasms in the Mummichog *Fundulus heteroclitus* from a Creosote-Contaminated Site. *Cancer Research* **50**, 5978-5986.

Vos P, Hogers R, Bleeker M, *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* **23**, 4407-4414.

Wang Z, Baker AJ, Hill GE, Edwards SV (2003) Reconciling actual and inferred population histories in the house finch (*Carpodacus mexicanus*) by AFLP analysis. *Evolution Int J Org Evolution* **57**, 2852-2864.

Weinig C, Ungerer MC, Dorn LA, *et al.* (2002) Novel loci control variation in reproductive timing in Arabidopsis thaliana in natural environments. *Genetics* **162**, 1875-1884.

Weis J (2002) Tolerance to environmental contaminants in the mummichug, *Fundulus heteroclitus*. *Human and Ecological Risk Assessment* **8**, 933-953.

Whitehead A, Anderson SL, Kuivila KM, Roach JL, May B (2003) Genetic variation among interconnected populations of *Catostomus occidentalis*: implications for distinguishing impacts of contaminants from biogeographical structuring. *Mol Ecol* **12**, 2817-2833.

Wiehe T, Nolte V, Zivkovic D, Schlotterer C (2007) Identification of selective sweeps using a dynamically adjusted number of linked microsatellites. *Genetics* **175**, 207-218.

Wilding C, Butlin R, Grahame J (2001a) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology* **14**, 611-619.

Wilding CS, Butlin RK, Grahame J (2001b) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology* **14**, 611-619.

Wirgin I, Waldman JR (2004) Resistance to contaminants in North American fish populations. *Mutat Res* **552**, 73-100.

Wootton JC, Feng XR, Ferdig MT, *et al.* (2002) Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. *Nature* **418**, 320-323.

Yan G, Romero-Severson J, Walton M, Chadee DD, Severson DW (1999) Population genetics of the yellow fever mosquito in Trinidad: comparisons of amplified fragment length polymorphism (AFLP) and restriction fragment length polymorphism (RFLP) markers. *Mol Ecol* **8**, 951-963.

Table 1. Sample locations. Locations along the east coast of the United States where *F. heteroclitus* was collected.

Reference/Superfund	Abbreviation	Geographical location	Latitude (N)	Longitude (W)
Reference	SAND	Sandwich, MA	41°44.0’	70°23.0’
Superfund	NBH	New Bedford, MA	41°34.0’	70º54.9’
Reference	PTJ	Point Judith, RI	41°21.7’	71°28.9’
Reference	CLI	Clinton, CT	41°15.3’	72°32.8’
Superfund	NEW	Newark, NJ	40°41.2’	74°06.7’
Reference	TUCK	Tuckerton, NJ	39°32.2’	74°19.4’
Reference	MAG	Magotha, VA	37°10.6’	75°56.5’
Superfund	ER	Elizabeth River, VA	36°48.5’	76°17.7’
Reference	MAN	Manteo, NC	35°53.8’	75°36.9’

Table 2. Primer sequences used in AFLP analyses.

Primers	**Sequence (5'-3')**
Eco +1	
Eco +A	GACTGCGTACCAATTCA
Eco +C	GACTGCGTACCAATTCC
Mse +1	
Mse +A	GATGAGTCCTGAGTAAA
Mse +C	GATGAGTCCTGAGTAAC
Eco +3	
Eco +ACT	GACTGCGTACCAATTCACT
Eco +ACC	GACTGCGTACCAATTCACC
Eco +AAG	GACTGCGTACCAATTCAAG
Mse +3	
Mse +AGT	GATGAGTCCTGAGTAAAGT
Mse +ATC	GATGAGTCCTGAGTAAATC
Mse +CAA	GATGAGTCCTGAGTAACAA
Mse +CGA	GATGAGTCCTGAGTAACGA
Combinations	
A	Eco+ACT and Mse+AGT
B	Eco+ACC and Mse+ATC
C	Eco+AAG and Mse+CAA
D	Eco+ACT and Mse+CGA
E	Eco+ACC and Mse+CAA

Table 3. Outlier loci shared among the Superfund site *Fundulus* populations.

Population 1 and locus number	**Population 2 and locus number**	**Primer Set**
New Bedford Harbor, 19	Elizabeth River, 120	A
New Bedford Harbor, 98	Elizabeth River, 190	B
Newark Bay, 78	Elizabeth River, 147	A
Newark Bay, 335	Elizabeth River, 194	C

Table 4. Pairwise F_{ST} values with and without outlier loci. Mean F_{ST} between populations of *Fundulus heteroclitus* with and without outlier loci. Below diagonal: mean F_{ST} including outlier loci. Above diagonal: mean F_{ST} without outlier loci. Average of F_{ST} values below diagonal is 0.034 and 0.023 after the removal of outlier loci.

	SAND	NBH	PTJ	CLI	NEW	TUCK	MAG	ER	MAN
SAND		0.0090	0.0157	0.0258	0.0238	0.0213	0.0220	0.0315	0.0298
NBH	0.0361		0.0112	0.0168	0.0249	0.0299	0.0318	0.0318	0.0277
PTJ	0.0399	0.0399		0.0149	0.0228	0.0219	0.0250	0.0338	0.0308
CLI	0.0313	0.0245	0.0190		0.0211	0.0278	0.0309	0.0288	0.0370
NEW	0.0286	0.0294	0.0283	0.0584		0.0112	0.0239	0.0338	0.0328
TUCK	0.0270	0.0309	0.0264	0.0584	0.0197		0.0230	0.0218	0.0247
MAG	0.0365	0.0388	0.0303	0.0355	0.0310	0.0322		0.008	0.0159
ER	0.0387	0.0393	0.0460	0.0294	0.0722	0.0320	0.0140		0.0144
MAN	0.0349	0.0330	0.0350	0.0464	0.0607	0.0318	0.0243	0.0217	

Figure 1. Sample locations. Sampling locations for *Fundulus heteroclitus* populations. Circles are reference sites and stars are Superfund sites.

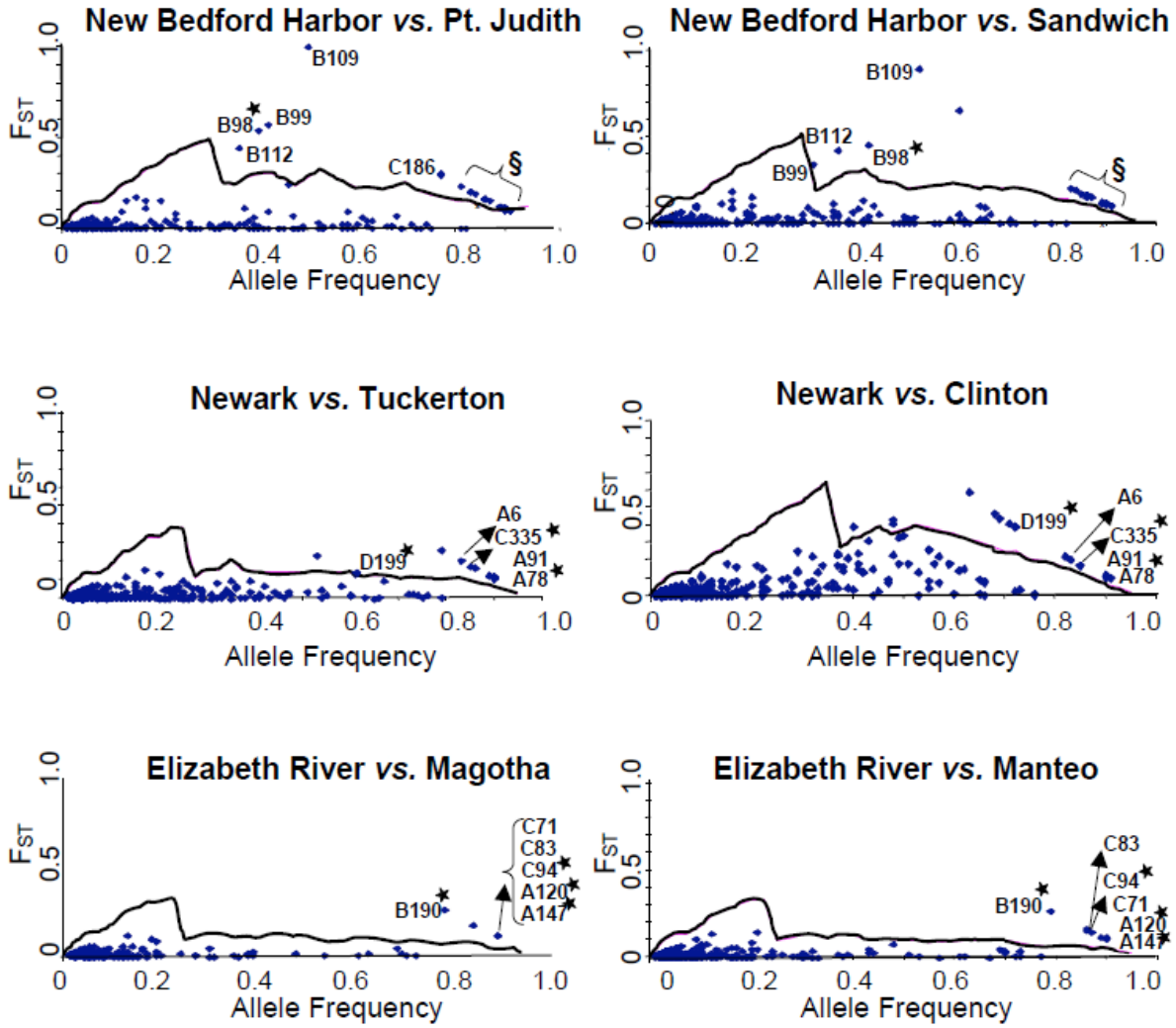


Figure 2. F_{ST} *versus* allele frequency values. F_{ST} values estimated from approximately 300 variable AFLP loci plotted against mean allele frequency. The solid line represents the 0.99 quantile estimated from a simulation model for each comparison. Loci shared among the same Superfund site are labeled with their primer set (letter) and number. Loci shared between Superfund sites are starred. § Shared loci included in these points are: A2, A19, A34, A56, D87, E118, E127, E137, E150, E156, C186, C194, C205, and C252. E118 also is shared between New Bedford Harbor and Elizabeth River populations.

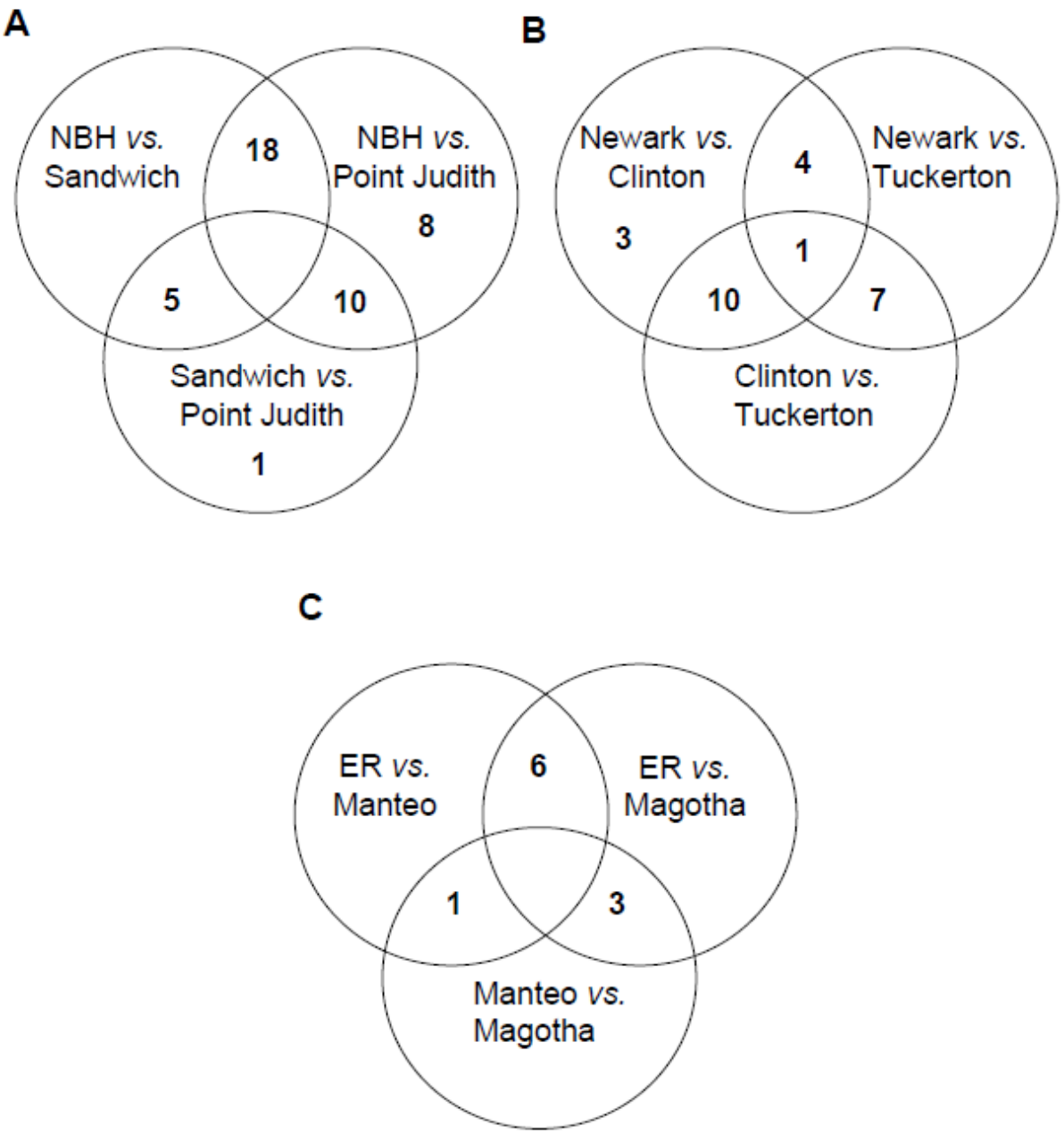


Figure 3. Venn diagrams of shared outlier loci in each Superfund comparison. Outlier loci in comparisons of each Superfund populations to both its clean reference sites; numbers in the unions of circles represent outlier loci shared among populations. A) New Bedford Harbor, MA Sandwich, MA and Pt. Judith, RI comparison. B) Newark Bay, NJ, Clinton, CT, and Tuckerton, NJ comparison. C) Elizabeth River, VA, Magotha, VA and Manteo, NC comparison.

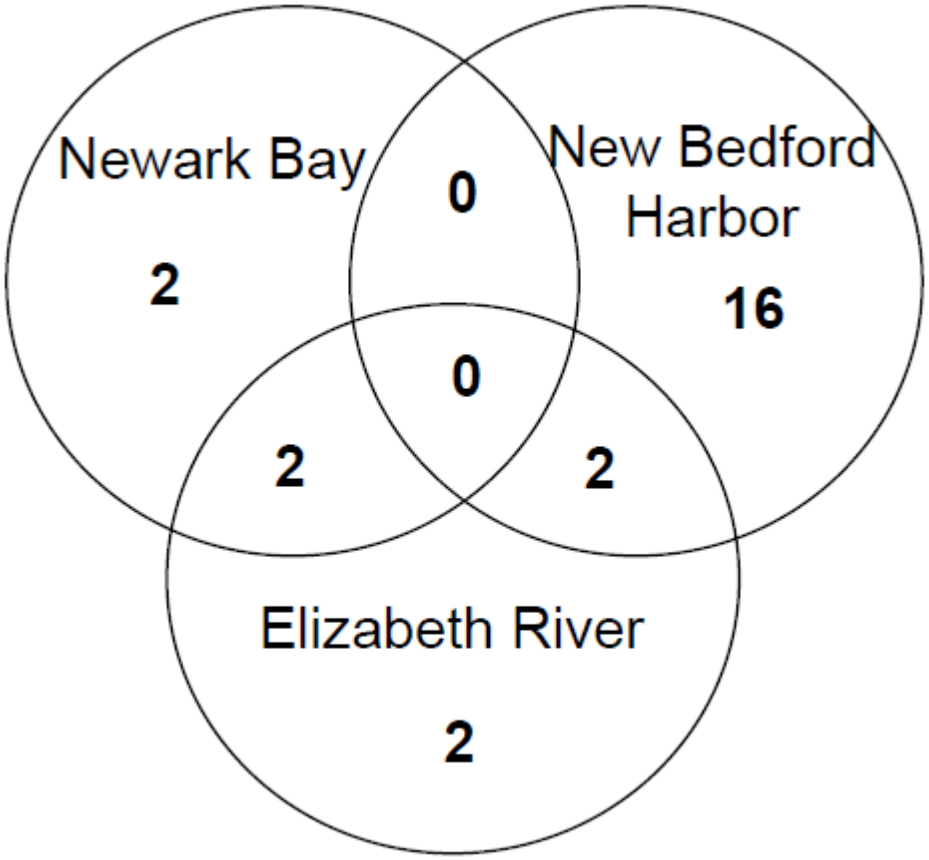


Figure 4. Venn diagram of shared outlier loci among Superfund populations. Shared outlier loci among Superfund population comparisons to both clean reference sites; numbers in the unions of circles represent outlier loci shared between two Superfund populations.

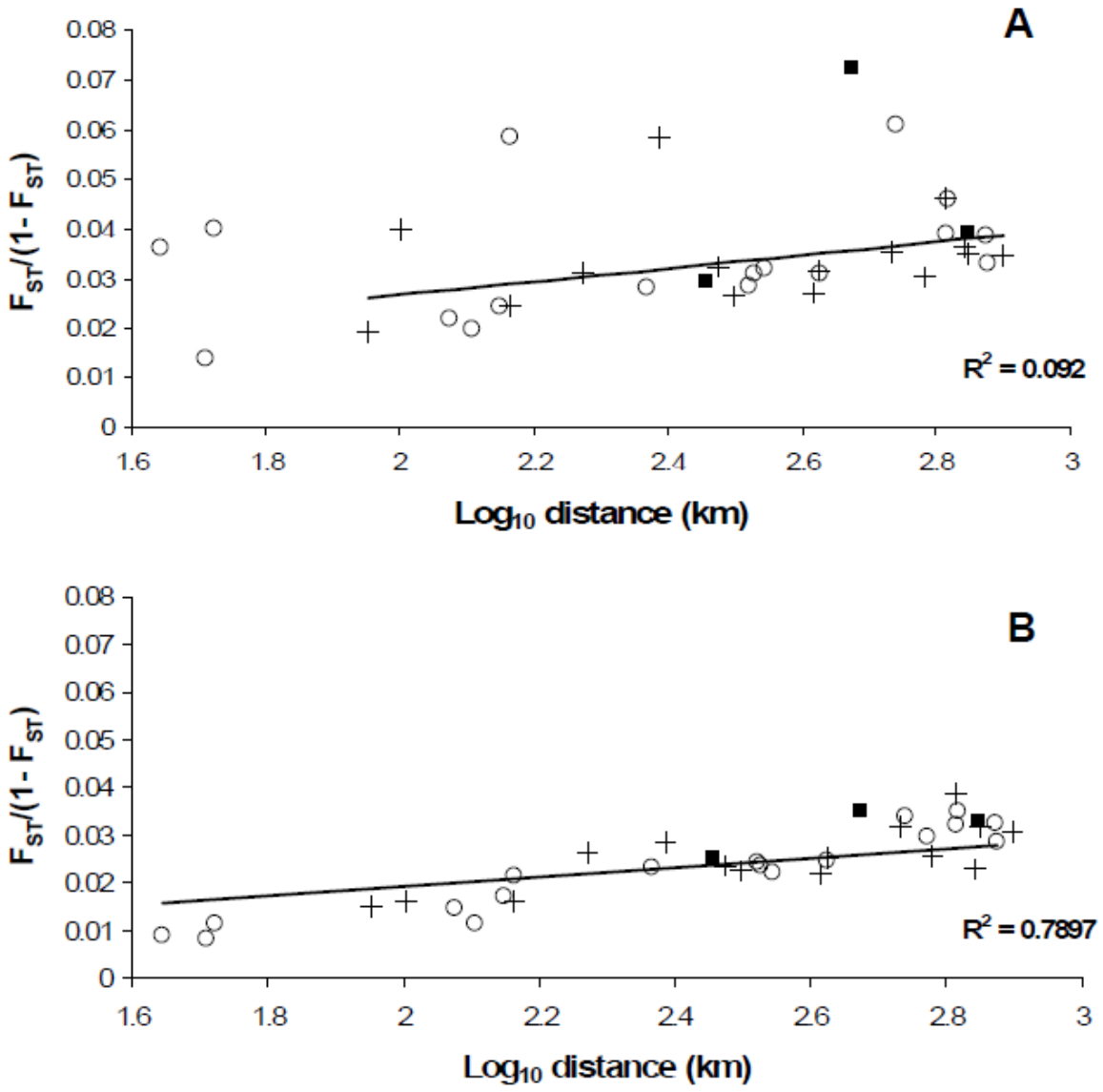


Figure 5. Geographic *versus* genetic distance. Relationship between genetic distance and geographic distance. Genetic distance was calculated from the mean F_{ST} for each pair of populations with (A) and without (B) outlier loci. Circles represent a pairwise comparison of a Superfund *versus* a reference site, squares represent a Superfund *versus* a Superfund site comparison, and crosses represent a reference *versus* a reference site comparison.

CHAPTER 2

SNP identification, verification, and utility for population genetics in a non-model genus

Larissa M. Williams[1], Xin Ma[2], Adam R. Boyko[2], Carlos D. Bustamante[2], and Marjorie F. Oleksiak[3]

1. Department of Environmental and Molecular Toxicology
Box 7633, North Carolina State University
Raleigh, NC 27695-7633 USA

2. Department of Biological Statistics and Computational Biology
Cornell University
Ithaca, NY 14853 USA

3. Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149 USA

Corresponding Author:
Marjorie Oleksiak
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149
Fax: 305-421-4600
Email: moleksiak@rsmas.miami.edu

Published in BMC Genetics 2010, **11**:32.

Abstract

Background: By targeting SNPs contained in both coding and non-coding areas of the genome, we are able to identify genetic differences and characterize genome-wide patterns of variation among individuals, populations and species. We investigated the utility of 454 sequencing and MassARRAY genotyping for population genetics in natural populations of the teleost, *Fundulus heteroclitus* as well as closely related *Fundulus* species (*F. grandis, F. majalis* and *F. similis*). **Results:** We used 454 pyrosequencing and MassARRAY genotyping technology to identify and type 458 genome-wide SNPs and determine genetic differentiation within and between populations and species of *Fundulus*. Specifically, pyrosequencing identified 96 putative SNPs across coding and non-coding regions of the *F. heteroclitus* genome: 88.8% were verified as true SNPs with MassARRAY. Additionally, putative SNPs identified in *F. heteroclitus* EST sequences were verified in most (86.5%) *F. heteroclitus* individuals; fewer were genotyped in *F. grandis* (74.4%), *F. majalis* (72.9%), and *F. similis* (60.7%) individuals. SNPs were polymorphic and showed latitudinal clinal variation separating northern and southern populations and established isolation by distance in *F. heteroclitus* populations. In *F. grandis*, SNPs were less polymorphic but still established isolation by distance. Markers differentiated species and populations. **Conclusions:** In total, these approaches were used to quickly determine differences within the *Fundulus* genome and provide markers for population genetic studies.

Background

High throughput sequencing and genotyping has become increasingly faster, less expensive and more accurate. In recent years this has lead to the establishment of myriad data sets ranging from increased coverage of variation in the human genome at the individual level (Bordoni *et al.*, 2008; Garber *et al.*, 2009; Ingman, Gyllensten, 2009; Turner *et al.*, 2009; Zheng *et al.*, 2009) to the sequencing of non-model prokaryotic and eukaryotic genomes and transcriptomes (Bontell *et al.*, 2009; De Schutter *et al.*, 2009; Iacono *et al.*, 2008; Novaes *et al.*, 2008; Vera *et al.*, 2008; Worden *et al.*, 2009). For many organisms sequencing of entire genomes is still unattained, but smaller, more targeted portions of the genome can be easily sequenced and genotyped. Such data can provide genome-wide sequence information which can be used to characterize population and selection pressure parameters as well as provide evolutionary insights that are broadly applicable (Luikart *et al.*, 2003).

One non-model genus, *Fundulus*, includes closely related species that range in physiology, environmental and habitat preference, and geographic locales; *Fundulus heteroclitus* and *Fundulus majalis* inhabit the Atlantic coast, and *Fundulus grandis* and *Fundulus similis* inhabit the Gulf Coast. Many *Fundulus* species and/or populations have extensive euryhaline capabilities, respond well to varying ranges of hypoxia (Diaz, 2001; Diaz, Rosenberg, 1995; Smith, Able, 2003), live along a steep thermocline, and have adapted to extremely polluted areas (Wirgin, Waldman, 2004). A variety of studies have investigated the underlying genetic basis of this teleosts' phenotypic plasticity. While some of the transcriptome is known for *F. heteroclitus* (Fisher, Oleksiak, 2007; Gonzalez *et al.*, 2006;

Meyer *et al.*, 2005; Oleksiak, 2008; Oleksiak *et al.*, 2002; Oleksiak *et al.*, 2005; Paschall *et al.*, 2004; Peterson, Bain, 2004; Roling *et al.*, 2006; Whitehead, Crawford, 2005; Whitehead, Crawford, 2006) much of the genome-wide variation within and between populations and species for this genus is relatively unknown.

Establishing a set of genetic markers, which can be used to assess regions of the genome involved in local adaptation and in speciation is important to understand fundamental similarities and differences between populations and species of *Fundulus*. Once markers are established they can be further studied to look for signatures of selection to any number of evolutionary forces (*e.g.*, pollution, hypoxia, salinity, temperature). A few studies have established genetic differences between populations of *F. heteroclitus* mainly with respect to phylogeographic constraints (Adams *et al.*, 2006; Bernardi *et al.*, 1993) or selection (Cashon *et al.*, 1981; Crawford *et al.*, 1989; Crawford, Powers, 1989; Crawford, Powers, 1992; McMillan *et al.*, 2006; Powers, Place, 1978; Powers *et al.*, 1986; Whitehead, 2009; Williams, Oleksiak, 2008). These studies used microsatellite, mitochondrial DNA, and AFLP analyses as well as targeted gene approaches. Single nucleotide polymorphisms (SNPs) are a useful starting point to scan large and disparate regions of the genome due to their abundance in both coding and non-coding regions, their co-dominant nature, and lack of ambiguity.

SNPs have been used to establish differences between individuals (Gill, 2001), populations (Paschou *et al.*, 2007; Weir *et al.*, 2005; Yamaguchi-Kabata *et al.*, 2008) and species (Kong *et al.*, 2008; Primmer *et al.*, 2002). They also are useful markers for propensity to disease (Amos *et al.*, 2008; Johnson *et al.*, 2007; Tomlinson *et al.*, 2007),

disease states (Poehlmann *et al.*, 2007), and evidence of the genetic basis of adaptation (Hoekstra *et al.*, 2006; Mauricio *et al.*, 2003; Moen *et al.*, 2008; Namroud *et al.*, 2008). In vertebrates, a SNP occurs on average every 100 to 1000 base pairs and often is in linkage disequilibrium with many other SNPs along the chromosome, forming strong haplotypes, which can be easily identified (Vignal *et al.*, 2002). Unfortunately, SNP resources are not readily available in the majority of non-model species lacking genomic resources. With this in mind, we set out to establish a set of SNP markers to identify differences between *Fundulus* populations and species.

Methods

Sample Collection and Extraction

F. heteroclitus were collected using minnow traps during the spring of 2005. Spleen and testes were sampled from 20 individuals from each of ten collection sites along the East coast of the United States (Figure 1). *F. grandis* were collected using minnow traps during the winter of 2009 (Figure 1). Fin clips were sampled from 15 individuals from each of the six collection sites along the Gulf Coast of the United States. Spleen from *F. majalis* was extracted from 13 individuals from Woods Hole, Massachusetts and 10 individuals from Sapelo Island, GA. Spleen also was extracted from *F. similis* collected from Pensacola, Florida (3 individuals) and Corpus Christi, Texas (8 individuals).

Genomic DNA from spleen and testes was extracted by phenol and chloroform as described in Wirgin *et al.*(Wirgin *et al.*, 1990), and DNA was resuspended in 50 µL 0.1X TE buffer. Genomic DNAs from fin clips were extracted using a modified version of Aljanabi

and Martinez (Aljanabi, Martinez, 1997) and DNA was resuspended in 50 µL 0.1X TE buffer.

DNA Pyrosequencing

F. heteroclitus genomic DNAs (500 ng) from eight individuals in each of ten collection sites (all sites except Point Judith, RI, Figure 1A) were digested individually with 1U BspE1 (New England Biolabs, MA) and 1U EcoRI (New England Biolabs, MA). Samples were incubated for three hours at 37º C in a total volume of 30 µL containing Buffer 3 (New England Biolabs, MA). Adaptors (Table 1) to each of the restriction sites, 25 mM ATP, and 1U of T4 DNA ligase (Epicentre) were added to reactions and incubated at 16º C overnight. A 2' O-methyl block was added to the 3' cytosine base on the adapter. This block assured that only those fragments digested with both BspEI and EcoRI would be amplified with PCR and prevented amplification of fragments with the same type of restriction site on both ends of the fragment.

Preselective PCR reactions with primers specific to adaptors (Table 1) were performed in a total volume of 25 µL containing 2 µL of diluted (1:10 in 0.1x Tris-EDTA buffer) ligation product with EcoRI primer (Integrated DNA Technologies; 10 pmol), BspE1 primer (Integrated DNA Technologies; 10 pmol) and 1U *Taq*. PCR conditions were 20 cycles of 94º C for 10 sec, 49º for 30 sec, and 72º C for one min. Following the preselective amplification, a selective amplification was carried out to decrease the number of fragments amplified in each individual to approximately 200 by extending the primer on the 3' end. Preselective PCR products were diluted (1:10) and 2 µL of diluted product was amplified with primers (Table 1) to EcoRI+ AAG (Integrated DNA Technologies; 10 pmol) and BspEI

+C (Integrated DNA Technologies; 10 pmol) with 1U *Taq* in a 25 µL total volume. PCR conditions in the first cycle were 94º C for 10 sec, 65º C for 30 sec, and 72º C for one minute with the annealing temperature reduced by 0.5º C for 20 cycles, then 25 cycles of 94º C for 10 sec, 55º C for 30 sec, and 72º C for one minute.

Primers (Table 1) specific to the EcoRI restriction site were generated with the goals of labeling the DNA fragments from each individual with specific nucleotide barcodes (Parameswaran *et al.,* 2007) and preparing those samples for emulsion-based amplification. Starting at the 5' end, 19 nucleotides (Table 1) complementary to the primer on the DNA capture beads used in the emulsion PCR reaction (Margulies *et al.*, 2005) were synthesized (Integrated DNA Technologies). Following those nucleotides, each primer had a distinct 10 base pair barcode (Parameswaran *et al.*, 2007) used to identify individuals (ten primers in total). The final 19 base pairs of the primer were specific to the EcoRI adapter. The BspE1 primer (Table 1) started at its 5' end with 19 nucleotides (Table 1), which were complementary to the primer on the DNA capture beads followed by 18 base pairs specific to the BspE1 adapter (Figure 2). All primers were HPLC purified. Amplified selective fragments were diluted (1:10) and added to both EcoRI and BspE1 primers (Integrated DNA Technologies; 10 pmol) in a 25 µL volume. PCR conditions were 94º C for 10 sec, 50º C for 30 sec, and 72º C for one minute and were carried out for 30 cycles. PCR reactions were pooled into eight wells, where each of the ten distinct barcodes was represented only once in each of the pools. Each pool of PCR products was purified using QIAquick PCR Purification Kit (Qiagen, USA). PCR products were further purified with AMPure (Agencourt).

Emulsion PCR was carried out on PCR products as described (Margulies *et al.*, 2005). Amplification of the PCR product on the bead was controlled for by quantifying and calculating the size of the amplicon pool using a Bioanalyzer 2100 so that there was a minimum of $2x10^6$ copies of DNA that ranged in size from 100 to 700 base pairs. Subsequent products were sequenced on a Roche/454 Life Sciences GS FLX Sequencer at the University of South Carolina's Environmental Genomics Core Facility. The PicoTiter plate was subdivided into eight regions with an expectation of 30,000 reads per region (Meyer *et al.*, 2008).

Assembly of pyrosequencing sequences and SNP Detection

Sequences were trimmed of their barcodes. All 626 sequences with at least one ambiguous base were removed since the presence of even a single ambiguous base is an effective indicator of low-quality sequence (Huse *et al.*, 2007). Because shorter than expected read lengths also correlate strongly with incorrect reads (Brockman *et al.*, 2008), another three percent of the sequences (whose lengths were smaller than 100 bp) were removed. The remaining reads were aligned using CAP3 (Huang, Madan, 1999). Quality scores were rescaled to be comparable to the usual Phred Score using ARACHNE (Batzoglou *et al.*, 2002).

SNPs were called at both the individual level and population level. At the individual level, SNPs were called using both a Bayesian method and a likelihood ratio test (LRT) method. For the Bayesian method, 10^{-4} was used as the prior for the mutation rate (Duvernell *et al.*, 2008). At the population level, for each locus on the contig, we simulated the error model and marked a locus as a potential SNP if it had a larger number of second alleles in

comparison to the critical value from the error model. Furthermore, a potential SNP site had to have at least three individuals sequenced to 2X at that locus unless another potential SNP site was within five basepairs or over 90% of the individuals had been classified as heterozygous at the individual level. This was done to minimize the rate of false positives caused by homologs.

Bayesian and LRT model for SNP calling at individual level

For the Bayesian model, for each contig, *Prior* = 1 x 10^{-4} represents the mutation rate; *N* represents the total number of unique mapping loci with multiple allelic types; A^i and a^i represent, respectively, the major and minor alleles at locus *i*; N_i represents the total number of alleles observed for locus i, and Y_j is the type of the j^{th} allele copy among these N_i alleles where $j = 0 \cdots N_i$; finally, e_j is the probability of error of the j^{th} allele where the error probability is computed as $10^{\frac{-Q}{10}}$ and where *Q* is the corresponding quality score after rescaling.

The posterior probability for the i^{th} locus being homozygous or heterozygous is:

$$\mathbf{P}(Hetero. \mid data) = \frac{\mathbf{P}(data|Hetero.)\text{x}\ \mathbf{P}\ (Hetero.)}{\mathbf{P}(data)}$$

$$\sim \mathbf{P}(data \mid hetero.)\ \text{x}\ \mathbf{P}(Hetero.)$$

$$\sim Prior \times (0.5)^{N}$$

$$\mathbf{P}(Homo. \mid data) = \frac{\mathbf{P}(data|Homo.)\text{x}\ \mathbf{P}\ (Homo.)}{\mathbf{P}(data)}$$

$$\sim \mathbf{P}(data \mid homo.)\ \text{x}\ \mathbf{P}(Homo.)$$

$$\sim (1\text{-}Prior) \times \prod_{j}^{Ni} (1-e_j)^{1(Y_j=A^i)} \quad \times e_j^{\,1(Y_j=a^i)}$$

Based on the posterior probabilities from above, we classified each of these N loci as homozygous or heterozygous exclusively. If a locus was classified as heterozygous, it was further tested using a likelihood ratio test (LRT) as follows:

For a particular locus *i* on the contig:

$$\mathbf{P}(X_j = A^i) = p$$
$$\mathbf{P}(X_j = a^i) = 1-p$$

where X_j stands for the true allele that we should have observed. For each Y_j, we have an error probability of e_j associated with it.

Then we have:

$$\mathbf{P}(Y_j = A^i \mid X_j = A^i) = 1-e_j$$
$$\mathbf{P}(Y_j = a^i \mid X_j = a^i) = 1-e_j$$
$$\mathbf{P}(Y_j = a^i \mid X_j = A^i) = e_j$$
$$\mathbf{P}(Y_j = A^i \mid X_j = a^i) = e_j$$

Therefore we have:

$$\mathbf{P}(Y_j = A^i \mid X_j = A^i) = (1-e_j) \times p + e_j \times (1-p)$$
$$\mathbf{P}(Y_j = a^i \mid X_j = A^i) = e_j \times p + (1-e_j) \times (1-p)$$

and

$$Y_j \sim Bernoulli(1,\ (1-e_j) \times p + e_j \times (1-p))$$

Based on all of the above, the likelihood of locus *I* was computed as:

$$L=\prod_{j=1}^{N_i}\{(1-e_j)\times p+e_j\times(1-p)\}^{I_j}\{e_j\times p+(1-e_j)\times(1-p)\}^{1-I_j}$$

Where $I_j = 1$ if $Y_j = A_i$; and $I_j = 0$ if $Y_j = a_i$

The LRT was performed with the hypothesis of H_O: p =0.5 *versus* H_a: p >0.5 and

$-2\times LRT \sim \chi^2(1)$.

Error model simulating

In order to call SNPs at the population level, we simulated the error model for each locus with multiple allelic types; we assumed that a particular locus was homozygous with major allele A^i and randomly simulated N_i number of alleles copies to be A^i or any of the other three allele types from a uniform distribution with probability (1 - e_j) and e_j respectively. We repeated this process 10,000 times and recorded the different numbers of second alleles found in the simulation. The critical value was chosen as the number of second alleles with a right-side p-value of 0.001.

Validation of SNPs

Multiplex assays targeting 458 SNPs in 250 *F. heteroclitus* individuals, 90 *F. grandis* individuals, 23 *F. majalis* individuals, and 21 *F. similis* individuals were attempted using the Sequenom MassARRAY technology. These consisted of 81 putative SNPs identified by the *F. heteroclitus* pyrosequencing, 350 putative SNPs previously identified in *F. heteroclitus* ESTs (Quackenbush *et al.*, 2000), and 27 putative SNPs from 22 genes containing, amongst others, SNPs in the aryl hydrocarbon receptor (Hahn *et al.*, 2004), lactate dehydrogenase B

(Bernardi *et al.*, 1993), and the proximal promoter of cytochrome P4501A (unpublished). Assays were designed using the MassARRAY Assay Design Software with the goal of maximizing multiplexing of 36 SNPs per well (Sequenom, San Diego, CA, USA). Only SNPs where 70 base pairs were annotated on either side of the polymorphism were included in the study. There were 14 SNPs previously identified with 454 pyrosequencing where this criterion was not met. If multiple SNPs were proximal (< 70 base pairs) to one another, one SNP was chosen and the other(s) was translated into a degenerate nucleotide (*e.g.*, K = G or T). Reaction conditions were performed by iPLEX chemistry as recommended by Sequenom across 13 plates at the University of Minnesota's BioMedical Genomics Center. SNP genotypes were called using the Sequenom System Typer 4.0 Analysis package. This software uses a three-parameter model to calculate the significance of each putative genotype. Based on the relative significance, a final genotype is called and assigned a particular name (*e.g.*, conservative, moderate, aggressive, user call). Non-calls also were noted (*e.g.*, low probability, bad spectrum).

Analysis of Genotype Data

Arlequin v.3.11 was used to calculate genetic diversity among populations (of *F. heteroclitus* and *F. grandis*) by calculating the percentage of polymorphic SNPs (P_O), observed (H_O) and expected heterozygosity (H_E), and the within-population fixation index (F) (Excoffier *et al.*, 2005). Fixation index deviations from zero were tested by 10,000 permutations of alleles between individuals. Hardy-Weinberg equilibrium also was tested in each population. An analysis of molecular variance (AMOVA) was performed to calculate the distribution of variance within populations, between North and South regions, and

between *F. heteroclitus* populations within North and South regions. For *F. grandis*, the AMOVA was performed to calculate the distribution of variance within populations as well as between populations longitudinally along the Gulf of Mexico. Since SNPs were initially identified from *F. heteroclitus* sequence data, a maximum of 5% missing data was used as a parameter for calculations involving *F. heteroclitus* and 10% for all others.

A Mantel test was performed to assess the assumption of isolation by distance using XLSTAT 2009 for *F. heteroclitus* and *F. grandis*.

STRUCTURE v.2.2 (Falush *et al.*, 2003; Pritchard *et al.*, 2000) was used to estimate the number of populations (K) in *F. heteroclitus*, *F. grandis*, *F. majalis* and *F. similis* along both the Western Atlantic and the Gulf of Mexico and to assign individuals to these populations. The Monte Carlo Markov Chain was run for 10^5 iterations following a burn-in period of 10^5 iterations for K = 1 to 14 using the correlated allele frequencies model and assumed admixture. Distruct v. 1.1 (Rosenberg, 2004) was used to generate bar plots to depict classifications with the highest probability under the model. JMP Genomics 3.2 for SAS 9.1.3 conducted principal component analysis on all samples to establish population structure.

Results

GS FLX Sequencing and Assembly

A total of 111,001 reads were obtained in one run of the GS FLX instrument producing 5,346,445 total bases of sequence (average read length of 218 bases) with 99.98 % of bases having a quality score of 20 or greater. Across the eight regions of the plate, there

were on average 1,982 reads per individual. The third barcode produced many less reads per region (<1,000) amongst all regions. All other barcodes performed very similarly with respect to the number of reads per individual across regions. Only 46% of the number of expected reads (111,001 instead of 240,000) were obtained from sequencing. Prior to sequencing, the amplification success of loci on the beads was checked for quality using a Bioanalyzer 2100, and all samples passed. However, three of the eight regions produced half the expected number of reads and a fourth region produced only 15% the expected number of reads. This indicated local problems in sequencing with respect to particular regions and the samples in those regions rather than the plate as a whole. All control beads passed the filter control with an average percentage of 90% across all regions, whereas the percentage of samples passing the filter control varied between regions and averaged 36%: regions with fewer than expected reads had fewer samples passing the filter control. Two regions had very high failure rates due to mixed samples, indicating more than one amplicon per bead.

Upon alignment 1,464 contigs were obtained with an average length of 213 base pairs. The average coverage across all loci was 22 reads per contig (Figure 3). Due to the low coverage of any one contig per individual, the detection of a SNP within a contig was mainly based on its presence across populations rather than at the individual level. Of the 1,464 contigs obtained, 96 contained SNPs. Within these contigs, 261 SNPs were identified. Among those contigs containing SNPs, the average length was 243 base pairs with an average coverage of 184 reads per contig (Figure 3). The observed rate of SNP detection is a function of depth, so as read counts per contig increased so did the number of SNPs detected. One third of all contigs with identified SNPs had only one SNP and 57% had two or fewer

SNPs per contig. SNPs were distributed approximately evenly along the position in the contig (R^2= 0.01).

Genotyping success

Of the initial 458 loci we attempted to amplify, 281 had a greater than 90% successful call rate among all individuals with no more than two alleles per SNP. In *F. heteroclitus* 74.4% of all loci amplified in greater than 95% of individuals. In *F. grandis,* 11% of SNPs did not amplify, and 74% of SNPs were monomorphic. 24% of the monomorphic SNPs in *F. grandis* also were monomorphic in *F. heteroclitus,* but for the alternative allele, indicating fixed differences between the two species.

On average, 80% (SD= ± 7.4%) of the putative SNPs identified with 454 pyrosequencing were amplified with MassARRAY in *F. heteroclitus* individuals: 72 of the 81 loci (88.8%) were polymorphic, 8 loci (9.8%) were monomorphic, and one locus did not amplify. Among all other putative SNPs genotyped with MassARRAY, 83% were successfully amplified. However, 13.5% of all loci in *F. heteroclitus*, 25.6% in *F. grandis*, 27.1% in *F. majalis* and 39.3% in *F. similis* did not amplify (Figure 4a). Many non-*heteroclitus* loci were also not polymorphic, and in *F. heteroclitus* 12.3% of all loci were monomorphic, as were 58.2% in *F. grandis*, 26.4% in *F. majalis*, and 29.7% in *F. similis* (Figure 4b). Due to the divergence between species resulting in unsuccessful amplification in non-*heteroclitus* individuals, locus amplification success was addressed on a species and population level for all remaining tests and not on the overall amplification success. Due to the low sample size, amplification rate, and predominant monomorphism of loci in *F. majalis* and *F. similis* samples, further characterizations of genetic parameters (with the exception of

population structure) were not carried out for these two species.

SNPs which were identified by Sequenom software as low probability in greater than 50% of all individuals were removed (17 SNPs in total). An additional 20 SNPs were excluded from analyses due to their excessive heterozygosity across individuals and populations of *F. heteroclitus*. These SNPs may represent segmental duplication where the two duplicate regions are identical, except that a SNP has been driven to high frequency or become fixed in one of the duplicates.

Genetic Diversity

The percentage of polymorphic SNPs (P_O) ranged from 3.7% to 67% (Table 3) among populations and species. The percentages of polymorphic SNPs were significantly different between northern and southern populations of *F. heteroclitus* where levels decreased in populations further north and east (p=0.035). Among *F. grandis* populations, the percentages of polymorphic SNPs did not significantly differ along latitude (p = 0.143) or longitude (p= 0.415). Among populations, most loci were in Hardy-Weinberg equilibrium (Table 3). Observed heterozygosity (H_O) among all populations ranged from 0.016 to 0.17 with a mean of 0.10 (Table 3). Observed heterozygosity was lower in northern *F. heteroclitus* in comparison to southern populations (p=0.04) and did not differ along latitude (p = 0.72) or longitude (p= 0.33) in *F. grandis*. Average expected heterozygosity (H_E) ranged from 0.019 to 0.20 with a mean of 0.11 (Table 3). The average within-population fixation index, *F*, averaged over all polymorphic loci was on average 0.16 in *F. heteroclitus* and 0.20 in *F. grandis* (Table 3).

SNPs identified *via* 454 sequencing did not have genetic parameters that differed

from SNPs identified in ESTs with the exception of Hardy-Weinberg equilibrium. 454-derived SNPs had a higher percentage of SNPs not in Hardy-Weinberg equilibrium due to a lack of heterozygosity (22% *versus* 9%).

Many SNP loci (60%) in *F. heteroclitus* had a frequency greater than 0.10 and were considered common SNPs (Table 4). In contrast, 90% of SNPs in *F. grandis* had low minor allele frequencies below 0.10.

Population Structure

The two independent tests of population stratification (STRUCTURE and principle component analysis (PCA)) identified species and population differences in all samples (Figure 5). STRUCTURE analysis, which uses a Bayesian MCMC clustering approach to assign individuals to clusters, separated populations into eight different clusters (Pr(K) = 0.37; Figure 4a). At the most probable clustering of the data (K=8), ten runs produced nearly identical membership coefficients which had pairwise similarity coefficients greater than 0.98. *F. heteroclitus* clustered north to south and *F. grandis* as its own separate cluster. Among *F. heteroclitus*, individuals from Maine and Georgia, the most northern and southern collection sites, formed their own distinct clusters. Individuals from sites between Maine and Georgia clustered with others from geographically similar sites. *F. majalis* and *F. similis* clustered together and away from the other two species. Similarly, in the PCA analysis, which does not rely on modeling the data, northern and southern *F. heteroclitus* stratified by latitude and were distinct from each other (Figure 5b) and each *F. heteroclitus* population was clustered together (Figure 5c). *F. grandis* made its own cluster and *F. majalis* and *F. similis* clustered together apart from other species (Figure 5b).

In *F. heteroclitus*, AMOVA showed that most of the variation was distributed within populations (59.05%), but another large proportion of variation (31.1%) was distributed among northern and southern regions. The remaining 9.85% of variation was explained by differences among populations within regions. In *F. grandis*, most of the variation was distributed within populations (82.4%), and a smaller proportion (17.6%) of variation was distributed longitudinally between populations across the Gulf of Mexico.

A Mantel test showed significant isolation by distance among *F. heteroclitus* populations ($p < 0.001$) and *F. grandis* populations ($p = 0.032$).

Discussion

We used high throughput sequencing and genotyping technology to identify and verify SNP markers in four non-model species within the *Fundulus* genus. Genotype data sharply differentiated northern and southern populations of *F. heteroclitus* as well as other species in this genus (*F. grandis*, *F. majalis*, and *F. similis*). Within the species where SNPs were originally annotated, most can be successfully verified and used to study population structure as well as the role and outcome of selection forces on a genome-wide scale.

Using the 454 FLX pyrosequencing system, we observed 111,001 reads yielding an average of 22x coverage across 1,464 contigs. Read lengths and quality scores were similar to many other studies using the 454 FLX system to sequence uncharacterized genomes (Novaes *et al.*, 2008; Wiedmann *et al.*, 2008), but we identified fewer SNPs. Two-hundred and sixty-one SNPs were identified in 96 of these contigs (81 were further verified with the Sequenom MassARRAY platform). The percentage of contigs containing SNPs did differ

between experiments: we obtained 0.07% of contigs containing SNPs while pyrosequencing of *Eucalyptus* ESTs identified 0.05% of contigs containing SNPs (Novaes *et al*., 2008) and pyrosequencing of size selected, genomic DNA from swine identified 11.4% of contigs contained SNPs (Wiedmann *et al.*, 2008).

Our 454 pyrosequencing of genomic DNA was originally designed to both discover and genotype SNPs within and among populations of *F. heteroclitus*. Thus, we attempted to perform genome reduction with selective PCR reactions to approximately 200 loci that could be sequenced in 10 populations of 8 individuals. With 30,000 reads per one-eighth of a 454 sequencing plate, each region would have 15X coverage per individual or 980X coverage across all populations, enabling accurate genotype calls for most individuals. However, preselective amplication was not perfect, and many more than 200 loci were sequenced; most amplified only a single time in a single individual (these singlets therefore were not useful for variant detection). Furthermore, we obtained only 46% of the expected number of reads. In the end, these problems led to the inability to directly call individual genotypes. We were hoping to both identify SNPs and genotype individuals in a single step, but a more successful approach (as evidenced by the swine group (Wiedmann *et al.*, 2008)) is to make reduced representation libraries from many pooled individuals for SNP discovery followed by individual genotyping. Because a pool of individuals is used, this approach identifies few singlets and thus enhances the number of reads per contig. Furthermore, improvements in both the number and length of reads using the Titanium series FLX 454 system compared to the original FLX system we used will increase the number of identified SNPs.

To increase our ability to measure population genetic parameters within and among

populations, we verified SNPs identified through 454 sequencing and additional SNPs annotated from *F. heteroclitus* cDNAs using the MassARRAY system. Similar percentages of 454 pyrosequencing derived SNPs and SNPs identified from ESTs were verified (80% and 83%, respectively). Of the 458 putative SNPs, 379 (82.75%) were polymorphic, but only 264 had a greater than 90% successful call rate among all individuals. Among *F. heteroclitus*, most SNPs amplified (61.3% were called in >95% of individuals) indicating that differences in amplification rate between species led to the lower overall call rate. In white spruce, 91% of SNPs verified with the Illumina SNP bead array platform (Fan *et al.*, 2003; Shen *et al.*, 2005) were true. Comparable to *F. heteroclitus*, 70% of SNPs in spruce were called in greater than 95% of individuals (Namroud *et al.*, 2008). Overall, verification of SNPs was powerful in providing information over many markers and individuals and was able to provide data to determine differences within populations, between populations and between species.

Species differentiation was demonstrated using principle component analysis (PCA) as well as STRUCTURE analysis. Both analyses showed separation between *F. heteroclitus*, *F. grandis* and *F. majalis* and *similis* as well as population structure within *F. heteroclitus* (Figure 4). These analyses provided the most resolution (even among distinguishing populations) in *F. heteroclitus* because the SNPs were originally identified in this species (*i.e.*, due to an ascertainment bias). PCA and STRUCTURE did not differentiate sister species, *F. similis* and *F. majalis*, from each other or establish population structure within these species. Small sample sizes (1 to 10 individuals per population), high levels of monomorphism (average of 28% of all SNPs), and the fact that only 10% of SNP alleles

differed between these two species, decreased the power to detect such differences when analyzed in conjunction with *F. heteroclitus* and *F. grandis*. Population structure also was masked in *F. grandis* when data was analyzed with other species. However, when *F. grandis* individuals were analyzed separately, they also showed distinct population structure (data not shown). One other study has reported multiple fixed differences in mitochondrial sequences between *F. heteroclitus* and *F. grandis* (Whitehead, 2009), but no other study to date has evaluated differences at many loci between all four species used in this study.

Within *F. heteroclitus* and *F. grandis* species, within-population fixation indices (F_{IS}, averaged across all loci) ranged from 0.09 to 0.32. Among *F. heteroclitus*, all populations had an overall significant deficiency of heterozygotes indicated by positive F_{IS} values. In these populations, approximately 10% of loci had similarly very large F_{IS} values (>0.5) across populations causing the skew in the average F_{IS} value for each population. Within a population, these loci were predominately homozygous for one allele with a complete absence of the heterozygote and one or a few individuals homozygous for the alternative allele. The loci which presented this pattern were called conservatively at both alleles by Sequenom software across all individuals indicating that genotyping error was not the main reason for this pattern. Furthermore, all northern populations were predominately homozygous for one allele and all southern populations were predominately homozygous for the alternative allele indicating strong demographic patterns in the data. The same demographic pattern was not found in *F. grandis*. Among *F. grandis* populations, most (70%) SNPs with high F_{IS} values were different between populations. This is in contrast to *F. heteroclitus* populations where loci with high F_{IS} values were shared across populations.

Within any one *F. grandis* population, one allele was predominant as a homozygote with one or a few individuals with the alternative homozygote. The most parsimonious explanation is that there is undetected substructure.

SNPs in Hardy-Weinberg were shown to be moderately polymorphic (average of 60%) in *F. heteroclitus*. In *F. grandis*, SNPs were shown to lack polymorphism (7.18%). The higher percentage of monomorphic loci in *F. grandis* likely is due to ascertainment bias in SNP discovery caused by only using *F. heteroclitus* populations. Many of the monomorphic loci (24%) represent fixed differences between *F. heteroclitus* and *F. grandis*. Thus, while SNP markers developed in *F. heteroclitus* are not necessarily polymorphic in other *Fundulus* species, they still can be used to differentiate *F. heteroclitus* from other species.

Among *F. heteroclitus* populations, genotype data revealed strong latitudinal clines between the Northern and Southern *F. heteroclitus* populations. PCA, STRUCTURE, F_{ST} values, and the isolation by distance test identified that individuals from Northern populations (above 40-41°N) were distinct from Southern populations. This split is centered around the southern-most extent of the Atlantic coastal advancement during the late Pleistocene (Mickelson DM, 1983). Specifically, observed heterozygosity and allelic richness across all loci is significantly lower ($p=0.043$, $p=0.042$, respectively) in the north than in the south. These differences have been shown previously in morphological features (Able, Felley, 1986) numerous allozyme loci (Cashon *et al.*, 1981; Powers, Place, 1978; Powers *et al.*, 1986; Ropson *et al.*, 1990) and microsatellites (Adams *et al.*, 2006). The larger historical population size of *F. heteroclitus* in the south (Adams *et al.*, 2006) would

maintain greater heterozygosity and allelic richness at shared loci; in the north, where population sizes are smaller, loci have a higher probability of becoming fixed.

Four STRUCTURE clusters encompass the six northern populations while only two clusters encompass the five southern populations (Figure 5A). Separate northern clusters may be driven by smaller population sizes in which drift is greater. When genetic drift has a larger effect it becomes easier to distinguish populations because the average difference in allele frequencies of a marker in different populations will be greater. This is illustrated by a larger average F_{ST} of 0.20 among northern populations in comparison to that of an average F_{ST} of 0.10 among southern populations. This statistic is also evident for the north and south split, where populations from respective regions had an extremely high F_{ST} value of 0.44 when compared against one another. Similar genetic divergence has been reported for *F. heteroclitus* using microsatellites (0.196 among northern populations, 0.117 among southern populations and 0.330 for the two most divergent populations, Nova Scotia and Georgia (Adams *et al.*, 2006)). Similar demographic patterns have been described in freshwater fish (Bernatchez, Wilson, 1998) and marine species such as goby (Gysels *et al.*, 2004) and blue crab (McMillen-Jackson, Bert, 2004), and, as in *Fundulus,* these patterns are attributed to Pleistocene events.

A similar latitudinal cline occurs between populations of *F. grandis*, and a Mantel test shows significant isolation by distance. However, there were no significant differences between either levels of polymorphism or observed heterozygosity along latitude or longitude. Williams *et al.*, 2008 reported significant isolation by distance as well as decreased allelic richness with increasing latitude. In this 2008 study, microsatellites were

used, and two additional sites southern to those used in our study were included. Since microsatellites have many more alleles than SNPs and two additional sites were found to have relatively higher allelic richness in comparison to all other sampling sites along the gulf, this may account for the differences found in levels of polymorphism.

Conclusions

By targeting SNPs contained in both coding and non-coding areas of the genome, we are able to better understand how evolutionary forces are shaping the *Fundulus* genome. Similar studies using high throughput methods to sequence SNP markers have been developed in Atlantic cod (Moen *et al.*, 2008), white spruce (Namroud *et al.*, 2008), *Eucalyptus* (Novaes *et al.*, 2008), and swine (Wiedmann *et al.*, 2008). Like our study, these studies expanded their own species' knowledge base with respect to potential markers for studying evolutionary adaptation (in the case of cod and spruce), genome-wide assessment of diversity (*Eucalyptus*) or for use in breeding programs (swine).

Acknowledgments

The authors thank G. Bozinovic and M. Everett for assistance in the collection of samples and D. Crawford for valuable input into methodology. *Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University which is partially funded by Microsoft Corporation.* Funding was partially provided by NIEHS Training Grant ES007046 award from the Department of Environmental and Molecular Toxicology at North Carolina State University to LMW, NIH 5 RO1 ES011588 to MFO, NSF DEB0948510 to ARB, and NIH R01 HG003229 CDB.

References

Able KW, Felley JD (1986) Geographical variation in *Fundulus heteroclitus* - Tests for concordance between egg and adult morphologies. *American Zoologist* **26**, 145-157.

Adams SM, Lindmeier JB, Duvernell DD (2006) Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. *Molecular Ecology* **15**, 1109-1123.

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research* **25**, 4692-4693.

Amos CI, Wu XF, Broderick P, *et al.* (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics* **40**, 616-622.

Batzoglou S, Jaffe DB, Stanley K, *et al.* (2002) ARACHNE: A whole-genome shotgun assembler. *Genome Research* **12**, 177-189.

Bernardi G, Sordino P, Powers DA (1993) Concordant mitochondrial and nuclear-DNA phylogenies for populations of the teloest fish *Fundulus heteroclitus. Proceedings of the National Academy of Sciences of the United States of America* **90**, 9271-9274.

Bernatchez L, Wilson CC (1998) Comparative phylogeography of nearctic and palearctic fishes. *Molecular Ecology* **7**, 431-452.

Bontell IL, Hall N, Ashelford KE, *et al.* (2009) Whole genome sequencing of a natural recombinant *Toxoplasma gondii* strain reveals chromosome sorting and local allelic variants. *Genome Biology* **10**, R53.

Bordoni R, Bonnal R, Rizzi E, *et al.* (2008) Evaluation of human gene variant detection in amplicon pools by the GS-FLX parallel Pyrosequencer. *Bmc Genomics* **9**, 464.

Brockman W, Alvarez P, Young S, *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research* **18**, 763-770.

Cashon RE, Vanbeneden RJ, Powers DA (1981) Biochemical genetics of *Fundulus heteroclitus* (L). Spatial variation in gene-frequencies of IDH-A, IDH-B, 6-PGDH-A, and EST-S. *Biochemical Genetics* **19**, 715-728.

Crawford DL, Constantino HR, Powers DA (1989) Lactate Dehydrogenase-B cDNA from the Teleost *Fundulus heteroclitus* - Evolutionary Implications. *Molecular Biology and Evolution* **6**, 369-383.

Crawford DL, Powers DA (1989) Molecular-Basis of Evolutionary Adaptation at the Lactate Dehydrogenase-B Locus in the Fish *Fundulus heteroclitus*. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 9365-9369.

Crawford DL, Powers DA (1992) Evolutionary Adaptation to Different Thermal Environments via Transcriptional Regulation. *Molecular Biology and Evolution* **9**, 806-813.

De Schutter K, Lin YC, Tiels P, *et al.* (2009) Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nature Biotechnology* **27**, 561-U104.

Diaz RJ (2001) Overview of hypoxia around the world. *Journal of Environmental Quality* **30**, 275-281.

Diaz RJ, Rosenberg R (1995) Marine benthic hypoxia: A review of its ecological effects and the behavioural responses of benthic macrofauna. *Oceanography and Marine Biology - an Annual Review, Vol 33* **33**, 245-303.

Duvernell DD, Lindmeier JB, Faust KE, Whitehead A (2008) Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, *Fundulus heteroclitus*. *Molecular Ecology* **17**, 1344-1360.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics*, 47-50.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.

Fan J-B, Oliphant A, Shen R, *et al.* (2003) Highly Parallel SNP Genotyping. *Cold Spring Harbor Symposia on Quantitative Biology* **68**, 69-78.

Fisher MA, Oleksiak MF (2007) Convergence and divergence in gene expression among natural populations exposed to pollution. *Bmc Genomics* **8**, 108.

Garber M, Zody MC, Arachchi HM, *et al.* (2009) Closing gaps in the human genome using sequencing by synthesis. *Genome Biology* **10**, R60.

Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International Journal of Legal Medicine* **114**, 204-210.

Gonzalez HO, Roling JA, Baldwin WS, Bain LJ (2006) Physiological changes and differential gene expression in mummichogs (*Fundulus heteroclitus*) exposed to arsenic. *Aquatic Toxicology* **77**, 43-52.

Gysels ES, Hellemans B, Pampoulie C, Volckaert FAM (2004) Phylogeography of the common goby, Pomatoschistus microps, with particular emphasis on the colonization of the Mediterranean and the North Sea. *Molecular Ecology* **13**, 403-417.

Hahn ME, Karchner SI, Franks DG, Merson RR (2004) Aryl hydrocarbon receptor polymorphisms and dioxin resistance in Atlantic killifish (*Fundulus heteroclitus*). *Pharmacogenetics* **14**, 131-143.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313**, 101-104.

Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877.

Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**, R143.

Iacono M, Villa L, Fortini D, *et al.* (2008) Whole-genome pyrosequencing of an epidemic multidrug-resistant *Acinetobacter baumannii* strain belonging to the European clone II group. *Antimicrobial Agents and Chemotherapy* **52**, 2616-2625.

Ingman M, Gyllensten U (2009) SNP frequency estimation using massively parallel sequencing of pooled DNA. *European Journal of Human Genetics* **17**, 383-386.

Johnson N, Fletcher O, Palles C, *et al.* (2007) Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility. *Human Molecular Genetics* **16**, 1051-1057.

Kong FR, Tong ZS, Chen XY, *et al.* (2008) Rapid identification and differentiation of Ttichophyton species, based on sequence Polymorphisms of the ribosomal internal transcribed spleacer regions, by rolling-circle amplification. *Journal of Clinical Microbiology* **46**, 1192-1199.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.

Margulies M, Egholm M, Altman WE, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Mauricio R, Stahl EA, Korves T, *et al.* (2003) Natural selection for polymorphism in the disease resistance gene Rps2 of *Arabidopsis thaliana*. *Genetics* **163**, 735-746.

McMillan AM, Bagley MJ, Jackson SA, Nacci DE (2006) Genetic diversity and structure of an estuarine fish (*Fundulus heteroclitus*) indigenous to sites associated with a highly contaminated urban harbor. *Ecotoxicology* **15**, 539-548.

McMillen-Jackson AL, Bert TM (2004) Mitochondrial DNA variation and population genetic structure of the blue crab *Callinectes sapidus* in the eastern United States. *Marine Biology* **145**, 769-777.

Meyer JN, Volz DC, Freedman JH, Di Giulio RT (2005) Differential display of hepatic mRNA from killifish (*Fundulus heteroclitus*) inhabiting a Superfund estuary. *Aquatic Toxicology* **73**, 327-341.

Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nature Protocols* **3**, 267-278.

Mickelson DM CL, Fullerton DS, Borns HW (1983) Late-Quaternary Environments of the United States. In: *The late Wisconsin glacial record of the Laurentide ice sheet in the United States* (ed. Wright HW Jr PS). University of Minnesota Press, Minneapolis.

Moen T, Hayes B, Nilsen F, *et al.* (2008) Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *Bmc Genetics* **9**, 18.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology* **17**, 3599-3613.

Novaes E, Drost DR, Farmerie WG, *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *Bmc Genomics* **9**.

Oleksiak MF (2008) Changes in gene expression due to chronic exposure to environmental pollutants. *Aquatic Toxicology* **90**, 161-171.

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. *Nature Genetics* **32**, 261-266.

Oleksiak MF, Roach JL, Crawford DL (2005) Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. *Nature Genetics* **37**, 67-72.

Parameswaran P, Jalili R, Tao L, *et al.* (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research* **35**, e130.

Paschall JE, Oleksiak MF, VanWye JD, *et al.* (2004) FunnyBase: a systems level functional annotation of Fundulus ESTs for the analysis of gene expression. *Bmc Genomics* **5**, 96.

Paschou P, Ziv E, Burchard EG, *et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *Plos Genetics* **3**, 1672-1686.

Peterson JSK, Bain LJ (2004) Differential gene expression in anthracene-exposed mummichogs (Fundulus heteroclitus). *Aquatic Toxicology* **66**, 345-355.

Poehlmann A, Kuester D, Meyer F, *et al.* (2007) K-ras mutation detection in colorectal cancer using the Pyrosequencing technique. *Pathology Research and Practice* **203**, 489-497.

Powers DA, Place AR (1978) Biochemical genetics of *Fundulus heteroclitus* (L). Temporal and spatial variation in gene-frequencies of LDH-B, MDH-A, GPI-B, and PGM-A. *Biochemical Genetics* **16**, 593-607.

Powers DA, Ropson I, Brown DC, *et al.* (1986) Genetic variation in *Fundulus heteroclitus*-geographic distribution. *American Zoologist* **26**, 131-144.

Primmer CR, Borge T, Lindell J, Saetre GP (2002) Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology* **11**, 603-612.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Quackenbush J, Liang F, Holt I, Pertea G, Upton J (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research* **28**, 141-145.

Roling JA, Bain LJ, Gardea-Torresdey J, Bader J, Baldwin WS (2006) Hexavalent chromium reduces larval growth and alters gene expression in mummichog (*Fundulus heteroclitus*). *Environmental Toxicology and Chemistry* **25**, 2725-2733.

Ropson IJ, Brown DC, Powers DA (1990) Biochemical genetics of *Fundulus heteroclitus* (L.) 6. Geographical variation in the gene frequencies of 15 loci. *National Geographic Research* **44**, 16-26.

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Shen R, Fan JB, Campbell D, *et al.* (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **573**, 70-82.

Smith KJ, Able KW (2003) Dissolved oxygen dynamics in salt marsh pools and its potential impacts on fish assemblages. *Marine Ecology-Progress Series* **258**, 223-232.

Tomlinson I, Webb E, Carvajal-Carmona L, *et al.* (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics* **39**, 984-988.

Turner EH, Lee CL, Ng SB, Nickerson DA, Shendure J (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315-316.

Vera JC, Wheat CW, Fescemyer HW, *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**, 1636-1647.

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34**, 275-305.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15**, 1468-1476.

Whitehead A (2009) Comparative mitochondrial genomics within and among species of killifish. *Bmc Evolutionary Biology* **9**, 11.

Whitehead A, Crawford DL (2005) Variation in tissue-specific gene expression among natural populations. *Genome Biology* **6**, R13.

Whitehead A, Crawford DL (2006) Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5425-5430.

Wiedmann RT, Smith TPL, Nonneman DJ (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *Bmc Genetics* **9**, 81.

Williams LM, Oleksiak MF (2008) Signatures of selection in natural populations adapted to chronic pollution. *Bmc Evolutionary Biology* **8**, 282.

Wirgin I, Waldman JR (2004) Resistance to contaminants in North American fish populations. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **552**, 73-100.

Wirgin II, Damore M, Grunwald C, Goldman A, Garte SJ (1990) Genetic Diversity at an Oncogene Locus and in Mitochondrial DNA between Populations of Cancer Prone Atlantic Tomcod. *Biochemical Genetics* **28**, 459-475.

Worden AZ, Panaud, Piegu (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **325**, 147-147.

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, *et al.* (2008) Japanese Population Structure, Based on SNP Genotypes from 7003 Individuals Compared to Other Ethnic Groups: Effects on Population-Based Association Studies. *American Journal of Human Genetics* **83**, 445-456.

Zheng JB, Moorhead M, Weng L, *et al.* (2009) High-throughput, high-accuracy array-based resequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6712-6717.

Table 1. Adapters and primers used in the amplification of genomic DNA. Star indicates location of 2' O-methyl block.

Adapters	
BspEI (5' to 3')	
GACGATGAGTCCTGAGC	
CCGGGCTCAGGACTCATCGTC	
EcoRI (5' to 3')	
CTGAGTCCTAGTAGCACCTCGTAGACTGCGTACC	
AATTGGTACGCAGTCTAC*	
Preselective Primers	
EcoRI (5' to 3')	
CTGAGTCCTAGTAGCACC	
BspEI (5' to 3')	
GACGATGAGTCCTGAGC	
Selective Primers	
EcoRI (5' to 3')	
GACTGCGTACCAATTCAAG	
BspEI (5' to 3')	
GACGATGAGTCCTGAGCC	
Barcoded Primers	
EcoRI (5' to 3')	
1	GCCTCCCTCGCGCCATCAGAGCCTAAGCTGACTGCGTACCAATTCAAG
2	GCCTCCCTCGCGCCATCAGAGTTCAAGTCGACTGCGTACCAATTCAAG
3	GCCTCCCTCGCGCCATCAGACTTGAACTGGACTGCGTACCAATTCAAG
4	GCCTCCCTCGCGCCATCAGACGGTAACGTGACTGCGTACCAATTCAAG
5	GCCTCCCTCGCGCCATCAGATCCGAATCGGACTGCGTACCAATTCAAG
6	GCCTCCCTCGCGCCATCAGATGGCAATGCGACTGCGTACCAATTCAAG
7	GCCTCCCTCGCGCCATCAGCAGGTCCAGTGACTGCGTACCAATTCAAG
8	GCCTCCCTCGCGCCATCAGCATTGCCATGGACTGCGTACCAATTCAAG
9	GCCTCCCTCGCGCCATCAGCTAAGCCTAGGACTGCGTACCAATTCAAG
10	GCCTCCCTCGCGCCATCAGCGAATCCGATGACTGCGTACCAATTCAAG
BspEI (5' to 3')	
GCCTTGCCAGCCCGCTCAGGACGATGAGTCCTGAGCC	

Table 2. Genotyping success of SNP markers using the MassARRAY multiplex assay

Category	Number of SNPs	Percentage of SNPs
SNPs called in >95% of *F. heteroclitus* individuals	259	61.4
SNPs called in <80% of all individuals	135	31.9
SNPs called in >90% but <95% of all individuals	101	23.9
Monomorphic SNPs called in >95% of all individuals	23	5.4
Polymorphic SNPs called in >95% of all individuals	163	38.6
SNPs called in <90 % of all individuals identified in 454	35	43.2
SNPs called in >90% of all individuals identified in 454	46	56.8

Table 3. Genetic parameters of sampled populations in two species of *Fundulus*. † $p \leq 0.01$ based on 10,000 permutations between individuals within the same populations.

Fundulus heteroclitus

Population	P_O	H_O	H_E	F	% Departure from HWE
Maine	33	0.08	0.09	0.13†	7.0
Sandwich	48	0.12	0.14	0.13†	9.3
New Bedford Harbor	57	0.13	0.15	0.12†	7.9
Point Judith	44	0.11	0.13	0.18†	10.8
Clinton	59	0.12	0.13	0.09†	6.0
Newark	65	0.17	0.19	0.11†	6.4
Tuckerton	67	0.16	0.21	0.25†	12.5
Magotha	66	0.17	0.20	0.17†	11.5
Elizabeth River	67	0.16	0.20	0.23†	12.3
Manteo	65	0.16	0.20	0.19†	12.9
Georgia	51	0.13	0.16	0.19†	13.2
Mean	56.54	0.14	0.16	0.16	9.98
Standard Deviation	11.28	0.03	0.04	0.05	2.76

Fundulus grandis

Population	P_O	H_O	H_E	F	% Departure from HWE
Weeks Bay	9.0	0.016	0.019	0.13†	1.1
Dauphin Island	5.9	0.016	0.024	0.23†	2.8
Leeville	5.9	0.017	0.020	0.32†	2.0
Hackberry	10.1	0.023	0.032	0.10	3.0
Port O'Connor	8.5	0.018	0.026	0.27†	4.2
Port Aransas	3.7	0.021	0.031	0.23†	2.1
Mean	7.18	0.02	0.03	0.20	2.53
Standard Deviation	2.4	0.003	0.005	0.11	1.06

Table 4. SNP minor allele frequencies (MAF) within *F. heteroclitus* and *F. grandis* populations.

		Average MAF							
	# SNP	Including Monomorphic	Excluding Monomorphic	Monomorphic	0≤MAF≤0.1	0.1≤MAF≤0.2	0.2≤MAF≤0.3	0.3≤MAF≤0.4	0.4≤MAF≤0.5
F. heteroclitus	398	0.2	0.25	0.21	0.19	0.11	0.16	0.15	0.18
F. grandis	410	0.03	0.13	0.74	0.16	0.03	0.02	0.02	0.03
Overall	404	0.12	0.19	0.48	0.18	0.07	0.09	0.09	0.11

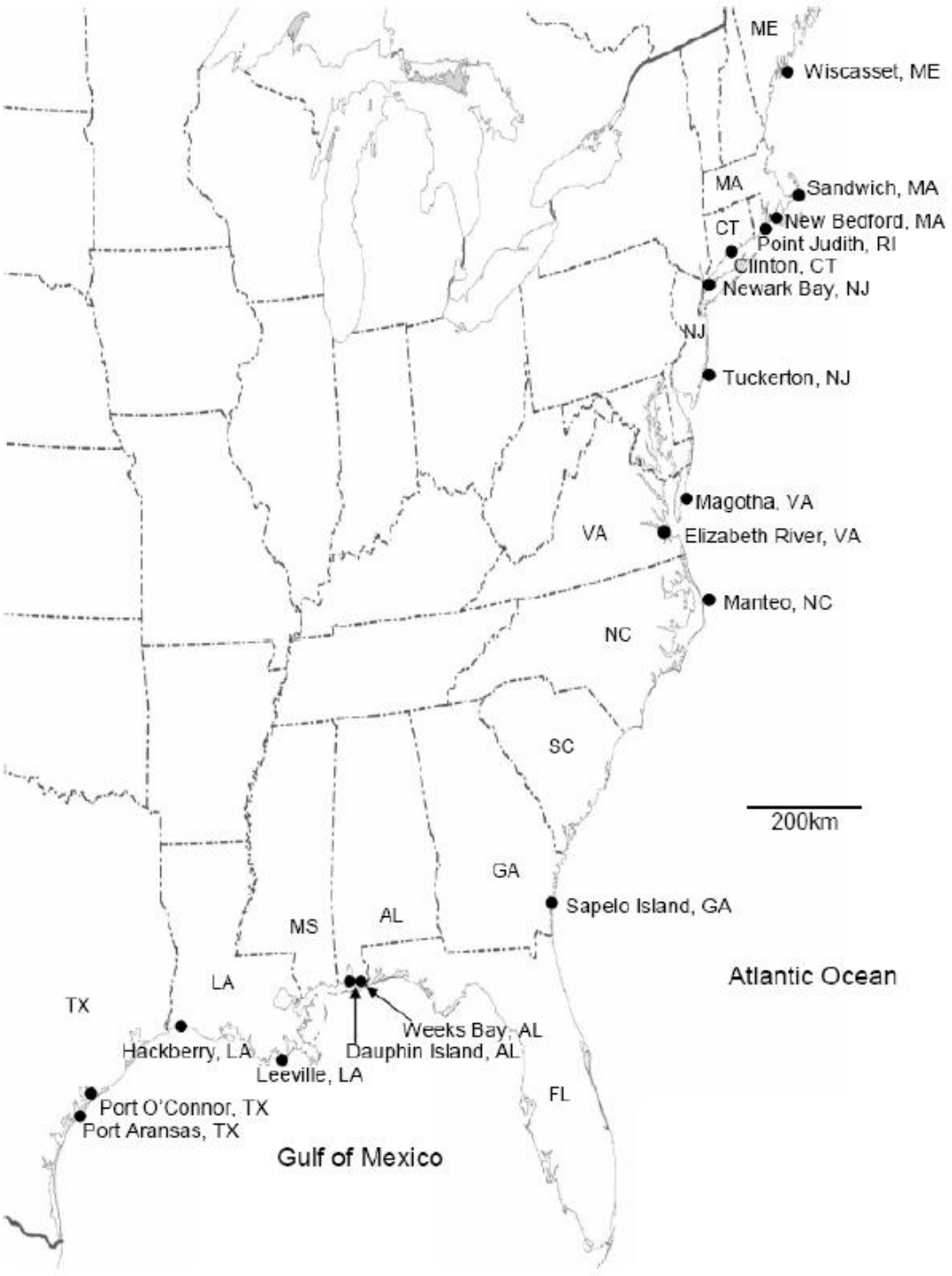


Figure 1. Sampling sites for *Fundulus* species. *F. heteroclitus* was collected along the east coast of the United States and *F. grandis* was collected along the Gulf of Mexico coast.

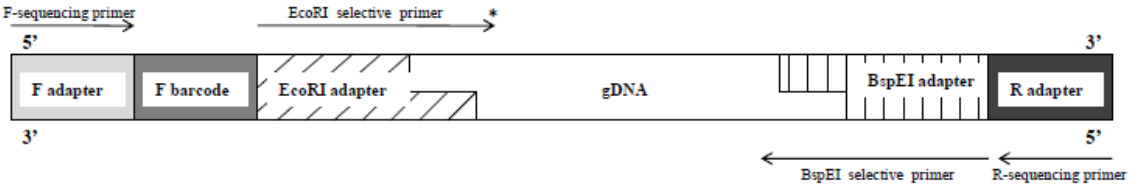


Figure 2. Design of 454 pyrosequencing contig generated from the digestion of genomic DNA with restriction enzymes (EcoRI and BspEI), the addition of restriction site specific linkers, an individual barcode and a 454 amplicon adapter.

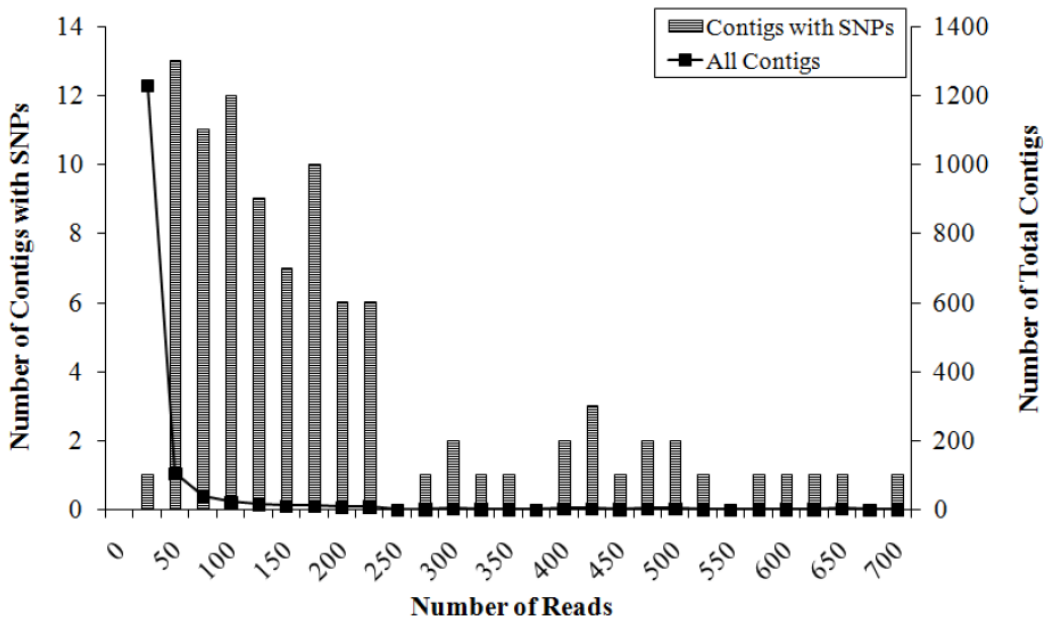


Figure 3. Contig totals *versus* number of reads per contig amongst those contigs with identified SNPs (bars) and all contigs (squares).

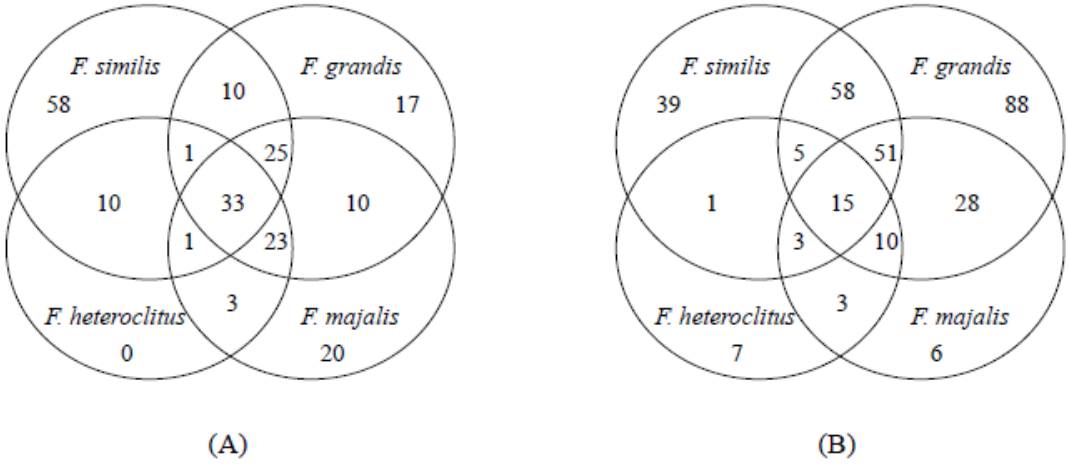


Figure 4. Non-amplified and non-polymorphic loci among *Fundulus* species. (A) Numbers of loci, which did not amplify with the MassARRAY platform among the four species of *Fundulus*. Not shown: loci shared between *F. majalis* and *F. similis* (8) and *F. heteroclitus* and *F. grandis* (12). (B) Numbers of loci, which were not polymorphic among the four species. Not shown: loci shared between *F. majalis* and *F. similis* (9) and *F. heteroclitus* and *F. grandis* (1).

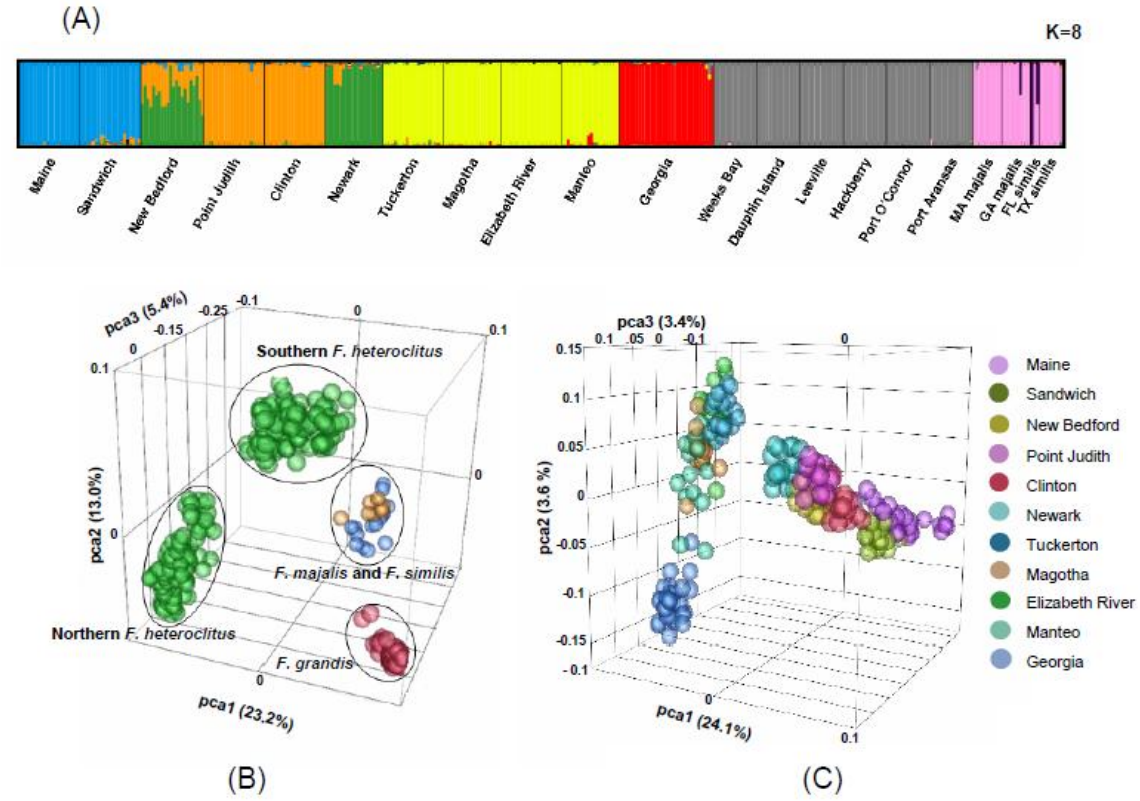


Figure 5. (A) Population structure as assessed by STRUCTURE. Bar plot was generated by DISTRUCT and depicts the classifications of the populations with the highest probability under the model. K indicates the number of clusters that maximized the probability of the model. Each individual is shown as a vertical bar. **(B)** Principal components PC1, PC2 and PC3 from all SNPs (as calculated in JMP Genomics 3.2) among all individuals. Species are separated from each other as well as northern and southern *F. heteroclitus* populations. Colors represent different species. **(C)** Principal components PC1, PC2, and PC3 from all SNPs among *F. heteroclitus* individuals. Colors represent different populations.

CHAPTER 3

Ecologically and Evolutionarily Important SNPs Identified in Natural Populations

Larissa M. Williams[1] and Marjorie F. Oleksiak[2]

1. Department of Environmental and Molecular Toxicology
Box 7633, North Carolina State University
Raleigh, NC 27695-7633 USA

2. Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149 USA

Corresponding Author:
Marjorie Oleksiak
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149
Fax: 305-421-4600
Email: moleksiak@rsmas.miami.edu

Abstract

Evolution by natural selection acts on natural populations amidst migration, gene-by-environmental interactions, constraints and tradeoffs, which affect the rate and frequency of adaptive change. We asked how many and how rapidly loci change in populations subject to severe, recent, environmental changes. To address these questions, we used genome-wide association approaches to identify SNPs with evolutionarily significant patterns in natural populations of *Fundulus heteroclitus* that inhabit and have adapted to highly polluted Superfund sites. Three statistical tests identified SNPs evolving by natural selection in three independent populations adapted to pollution: 1.6-4% of loci were significantly different from the neutral model in these populations. One SNP, in the xenobiotic metabolizing enzyme, cytochrome P4501A (CYP1A), was identified in all polluted populations using all tests. Extrapolating across the genome, these data suggest that rapid evolutionary change in natural populations will involve hundreds of loci, a few of which will be shared in independent events.

Introduction

Haldane estimated the mean rate of gene substitution as one per 300 generations (Haldane, 1990). Yet, due to human intervention, the global environment is changing rapidly (Kerr, 2007; Vitousek *et al.*, 1997) making adaptation in the 300 generations envisioned by Haldane unsustainable. One species, which has adapted in many fewer generations, is the estuarine fish *Fundulus heteroclitus*: it has adapted to anthropogenic contaminants in less than 15 generations (Nacci *et al.*, 1999) in at least three separate geographical locations along the east coast of the United States. These three populations are exposed to some of the highest concentrations of aromatic hydrocarbon pollutants of any vertebrate species (Wirgin, Waldman, 2004) and inhabit highly polluted Superfund sites (Elskus *et al.*, 1999; Meyer J, 2002; Nacci *et al.*, 1999; Ownby *et al.*, 2002), hazardous waste sites mandated for clean up by the Comprehensive Environmental Response, Compensation, and Liability Act of 1980. Such rapid adaptation indicates a strong selective force. Genome wide association (GWA) studies were used to identify SNPs associated with these polluted environments and thus provide an approach to investigate and understand rapid adaptive changes.

We sought to identify adaptive molecular variations that appear to be evolutionarily important within and among populations subject to anthropogenic stress using three statistical tests: an F_{ST} modeling approach, an association test, and a test on allele frequencies (MAF-F_{MAX}). We genotyped 367 SNPs from both coding and non-coding regions in 180 individuals from nine populations (Fig. 1). These populations have large effective population sizes and high genetic variation (Adams *et al.*, 2006). SNPs in each polluted population were compared with those in two flanking reference site populations located north and south of

each polluted site (triads, Figure 1). Populations within triads show little to moderate differentiation based on multilocus microsatellite estimates of F_{ST}, which range from 0.043 to 0.101 (Adams *et al.*, 2006). Our experimental design distinguishes pollutant effects from demographic ones because the genetic distance between the two clean reference populations is greater than the genetic distance between the polluted population and either reference population (Figure 1). Thus, the variation due to demography can be accounted for between the two distant reference populations, and divergence in a polluted population compared to both paired reference populations suggests that the contamination or other environmental factors associated only with the Superfund sites serves as the causative selective force.

Materials and Methods

F. heteroclitus were collected using minnow traps during the spring of 2005. Fin clips were sampled from 20 individuals from each of the nine collection sites along the Atlantic Coast of the United States (Figure 1). Three of the nine collection sites were EPA Superfund sites including New Bedford Harbor (EPA ID: MAD980731335), Newark (EPA ID: NJD980528996) and Elizabeth River (EPA IS: VAD990710410). To control for random processes such as drift, fish also were collected from populations from clean, control sites flanking each polluted site population.

Genomic DNA from fin clips was extracted using a modified version of Aljanabi and Martinez (Aljanabi, Martinez, 1997), and DNA was resuspended in 50 μL 0.1X TE buffer. We genotyped 180 *F. heteroclitus* at 458 SNPs using the MASSARRAY platform at the University of Minnesota as described (Williams *et al.*, 2010). We analyzed a subset of these

SNPs (367): SNPs that amplified in greater than 80% of *F. heteroclitus* and did not show an excess of heterozygosity. SNPs were from both coding and non-coding loci: 295 were coding and 72 were non-coding.

Arlequin v.3.11 was used to calculate F_{ST} values for each SNP within a population using the AMOVA function (Excoffier *et al.*, 2005). F_{ST} values were modeled to detect outliers using the FDIST2 program (Beaumont, Nichols, 1996). Simulations were run for each pair of populations using the average heterozygosity of the empirical data with 20,000 iterations assuming 10 demes, 2 populations, 20 individuals per sample and a stepwise mutation model. The 99^{th} percentile of simulation values was plotted against empirical data to determine the range of F_{ST} values in the neutral model. Those empirical values which exceeded simulation values and were shared outliers between each of the pair wise comparisons (polluted *versus* both reference sites) were considered to be outliers and potentially under selection by pollution or linked to loci under selection.

We used JMP Genomics 3.2 for SAS 9.1.3 to conduct SNP case control trait association tests. Tests were used to identify SNPs associated with pollution (trait) using a chi-square test with the assumption that individuals are unrelated in recent generations. A second case control trait association test was used to determine whether the differences in allele frequencies were due to geographical differences between clean sites. A likelihood ratio test was used to determine which of the associations, that of the polluted or clean sites, was the best fit for the data. SNPs with p-values < 0.01 in the association test and likelihood p-values < 0.01 for the polluted site model were identified as outliers. A Bonferroni correction was applied to each triad to correct for multiple testing.

Allele frequencies for each SNP were compared in each triad to determine whether there were significant differences in allele frequencies in a polluted site *versus* both reference sites. In each triad, the minor allele was determined on a per SNP basis. Within each population, the allele frequency of that overall minor allele was calculated for a random sampling of 15 out of 20 individuals 100 times. One-hundred random samplings of 15 out of 20 individuals has less than 1% probability that the same combination of individuals will be chosen more than once. A one-way ANOVA was performed on the iteratively derived values. To control for type I errors among all 100 iterations, we computed multiple test–corrected critical values for the F-statistic by the one-step adjustment method (Westfall, Young, 1993). This test used a random sampling of individuals (assuming no population structure) in the iterative ANOVA process described above to calculate the maximum F-value for all ANOVAs (F_{MAX}). Random sampling of individuals was carried out 1,000 times to determine the top 1% of the maximum F-statistics. F-statistics in the empirical data that exceeded the 1% F_{MAX} values were considered to be outliers.

Results and Discussion

The first statistical test used an F_{ST} modeling approach: empirical F_{ST} values of each SNP were compared against the 99th quantile of simulated, neutral distributions of F_{ST} values along the range of heterozygosity values (Figure 2) (Beaumont, Nichols, 1996). Selective loci were identified as outliers in each Superfund site population relative to its two reference site populations: polluted *versus* both references, where the union of polluted *versus* reference 1 and polluted *versus* reference 2 but not comparisons between the reference site

populations had p-values <0.01 (Williams, Oleksiak, 2008). Thirty outlier loci (8.2%) were found in the New Bedford Harbor triad, 9 (2.5%) in the Newark Bay triad, and 43 (11.7%) in the Elizabeth River triad. While F_{ST} tests relying on a simple island model of population differentiation have been shown to be robust, they are prone to a large excess of false positive loci when complex genetic structures exist (Excoffier *et al.*, 2009). To further assess the strength of our outliers, we performed two additional statistical tests.

We did association tests for each SNP within each triad and calculated p-values based on the strength of association with polluted populations (red line; Figure 3) or reference populations (blue line; Figure 3). We then used a likelihood ratio test to determine whether the polluted or reference model was a better fit of the data. In the New Bedford Harbor triad, 30 SNPs (8.2%) were significantly associated with the polluted model ($p \leq 0.01$). In the Newark Bay triad, fewer SNPs were associated with the polluted model (7; 1.9%). In the Elizabeth River triad, 29 SNP (7.9%) were associated with the polluted model. This association test detected many demographic effects in the Newark Bay triad (blue line, Figure 3), which is located at the historical introgression zone between northern and southern *F. heteroclitus* populations (Adams *et al.*, 2006; Bernardi *et al.*, 1993; Haney *et al.*, 2009). This large demographic effect may mask some of the effects of directional selection due to pollution.

We used a third test (MAF-F_{MAX} test) to determine whether the minor allele frequency (MAF) of each SNP was significantly different in a polluted site *versus* both the reference sites. For this test, we sampled 15 of the 20 individuals in each population 100 times and calculated the MAF of those subsets of individuals (15 polluted individuals and 30

reference individuals) on a SNP by SNP basis. Analysis of variance (ANOVA) was used on these 100 iterations to identify SNPs with MAF significantly different between the polluted and reference populations (Figure 4). An F_{MAX} (Westfall, Young, 1993) was used to control for type I errors among all 100 iterations: the distribution of F-values from permutations of the data assuming random population differentiation was used to determine the critical F-values that occur less than 1% of the time among all permuted F-values (1% F_{MAX} values (Westfall, Young, 1993)). In the New Bedford Harbor population, 22 SNPs (6.0%) had significantly different MAF between polluted and both reference populations (Figure 4). In the Newark Bay and Elizabeth River populations, 9 and 18 SNPs (2.5% and 4.9%, respectively) had significantly different MAF between polluted and reference populations (Figure 4).

The F_{ST} modeling approach, association test, and MAF-F_{MAX} test all identified similar percentages of SNPs with non-neutral patterns in all three triads. However, the MAF-F_{MAX} test often identified the fewest number of SNPs with non-neutral patterns (22, 9, and 18 for New Bedford Harbor, Newark Bay and Elizabeth River triads, respectively *versus* 30, 9 and 43 for the F_{ST} modeling approach and 30, 7 and 29 for the association test in these three triads). The MAF-F_{MAX} test asks whether a specific allele (the minor allele of the total population) occurs more frequently in the polluted population *versus* the flanking reference populations. This test differs from the association test, which associates an allele with the polluted population. Both tests differ from the F_{ST} modeling approach, which is based on differences in heterozygosities between populations. These tests use different aspects of the polymorphism spectrum measured by SNPs (minor allele frequency, frequency of each allele

and heterozygosity).

SNPs identified as outliers (exhibiting non-neutral behavior) in more than one test are statistically more powerful and less likely to suffer from type I errors. Within each triad, 6-15 SNPs were identified as outliers in all three tests: the New Bedford Harbor triad had 12, the Newark Bay triad had 6, and the Elizabeth River triad had 15 (Figure 1b, Supplemental Table S1). Among all triads, 15 of these SNPs occur in coding regions. Only one of these 15 SNPs, a SNP in β_2-microglobulin, is non-synonymous. β_2-microglobulin is vital to the immune response and is noncovalently associated with the heavy chain of the major histocompatibility complex (MHC) class I antigens (Bernabeu *et al.*, 1984). The G to A SNP in β_2-microglobulin changes an aspartic acid residue an asparagine predominantly in the polluted Newark Bay population. The other 14 SNPs that occur in coding regions all result in synonymous SNPs. Synonymous SNPs could affect mRNA splicing, translation or stability or could simply be linked to causative genetic polymorphisms. This can only be determined by functional, locus specific tests. Among the remaining annotated genes (Paschall *et al.*, 2004), five SNPs are in 3' untranslated regions and two are in 5' upstream regions. These SNPs potentially affect RNA stability and transcription.

If one assumes that the loci identified in all three tests are in fact undergoing natural selection, then 1.6% to 4.1% of the loci that we examined in the three triads is selectively important or linked to areas of the genome that are selectively important. Among these SNPs, we found no fixed SNPs between polluted and reference populations suggesting selection in favor of alleles that have not yet reached fixation (Voight *et al.*, 2006). This is not unexpected given the high migration rates among populations (Brown, Chapman, 1991)

and the recent selective change in the environment. In fact, despite the lethal concentrations of pollutants at these sites, these populations show no evidence of reduced genetic diversity (Williams, Oleksiak, 2008), most likely due to migrants.

A central question in outbred natural populations is whether similar or different solutions will evolve in response to comparable selective forces. Two SNPs that were significant in all three tests were shared between two different triads. A SNP in peroxiredoxin 6 was shared between the New Bedford Harbor and Elizabeth River populations, and a SNP in a sequence similar to mouse clone RP24-528E17 was shared between Newark Bay and Elizabeth River populations. Only one SNP was significant in all three tests and across all three triads: a SNP in the first intron of the phase I xenobiotic metabolizing enzyme CYP1A. CYP1A is integral to the detoxification pathway of many of the contaminants to which *F. heteroclitus* are continuously exposed in the three Superfund sites (Weis, 2002). These compounds, including polycyclic aromatic hydrocarbons (PAHs), dioxins and coplanar polychlorinated biphenyls (PCBs), induce CYP1A through the aryl hydrocarbon receptor (AhR) pathway (Hahn, 1998). In all three Superfund populations, CYP1A is refractory to induction by prototypical inducers (Bello *et al.*, 2001; Elskus *et al.*, 1999; Meyer, Di Giulio, 2002; Nacci *et al.*, 1999), and this trait is associated with resistance to PAH, PCB and dioxin toxicity (Bello *et al.*, 2001; Nacci *et al.*, 1999). Potentially, the SNP in the first intron of CYP1A affects transcription or is linked to SNPs affecting transcription.

GWA studies have been used to relate polymorphisms to disease and phenotypic traits (Hindorff *et al.*, 2009). For example, non-synonymous SNPs have been associated with

Crohn's disease, arthritis, freckles, and height (Hindorff *et al.*, 2009). Instead of relating SNPs to phenotypic traits, we asked which polymorphisms are associated with recent inhospitable environments (highly polluted Superfund sites). The analysis suggests that 1.6-4% of loci respond rapidly to change in the environment, 1.1-3% from mRNA encoding loci. These SNPs were randomly chosen and identified from expressed sequence tags (ESTs). If we assume 30,000 mRNA encoding genes, our results suggest that 330-900 loci have adaptively diverged in the last 50 years in each Superfund population. Relative to Haldane's expectation (Haldane, 1990), this is a large number of loci affected by natural selection over a short time. Possibly, the 367 polymorphic loci assayed have an unusually high frequency of adaptive change. Assuming one SNP every ~1000 bp (Vignal *et al.*, 2002), average gene sizes of ~1300 bp (Xu *et al.*, 2006), and *F. heteroclitus* EST lengths of 400 bp (modal size is closer to 550 bp), we missed SNPs in the remaining 800 or $2/3^{rds}$ of nucleotides. If we also assume that all $2/3^{rds}$ are evolving by neutral processes, then we overestimated the percentage of non-neutral loci by $2/3^{rds}$ by only sampling 1/3 of each gene. Thus, 0.37-1% rather than 1.1-3% of genic loci respond rapidly, and conservatively, 110-300 genes are involved in adaptation in natural populations in a short time. These data suggest that adaptive divergence can occur rapidly and involve hundreds of loci.

Acknowledgements

Partial funding for this work was received from NIH 5 RO1 ES011588 and NSF OCE 1008542. The Authors thank Douglas L. Crawford for helpful discussions and review of the manuscript.

References

Adams SM, Lindmeier JB, Duvernell DD (2006) Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. *Molecular Ecology* **15**, 1109-1123.

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res* **25**, 4692-4693.

Beaumont M, Nichols R (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London* **263**, 1619-1626.

Bello SM, Franks DG, Stegeman JJ, Hahn ME (2001) Acquired resistance to Ah receptor agonists in a population of Atlantic killifish (*Fundulus heteroclitus*) inhabiting a marine superfund site: in vivo and in vitro studies on the inducibility of xenobiotic metabolizing enzymes. *Toxicol Sci* **60**, 77-91.

Bernabeu C, van de Rijn M, Lerch PG, Terhorst CP (1984) Beta 2-microglobulin from serum associates with MHC class I antigens on the surface of cultured cells. *Nature* **308**, 642-645.

Bernardi G, Sordino P, Powers DA (1993) Concordant mitochondrial and nuclear DNA phylogenies for populations of the teleost fish Fundulus heteroclitus. *Proc Natl Acad Sci U S A* **90**, 9271-9274.

Brown BL, Chapman RW (1991) Gene flow and mitochondrial DNA variation in the killifish *Fundulus heteroclitus*. *Evolution* **45**, 1147-1161.

Elskus AA, Monosson E, McElroy AE, Stegeman JJ, Woltering DS (1999) Altered CYP1A expression in *Fundulus heteroclitus* adults and larvae: a sign of pollutant resistance? *Aquatic Toxicology* **45**, 99-113.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics*, 47-50.

Hahn ME (1998) The aryl hydrocarbon receptor: a comparative perspective. *Comp Biochem Physiol C Pharmacol Toxicol Endocrinol* **121**, 23-53.

Haldane JBS (1990) *The Causes of Evolution* Princeton University Press, Princeton.

Haney RA, Dionne M, Puritz J, Rand DM (2009) The comparative phylogeography of east coast estuarine fishes in formerly glaciated sites: Persistence versus recolonization in *Cyprinodon variegatus* ovinus and *Fundulus heteroclitus* macrolepidotus. *J Hered* **100**, 284-296.

Hindorff LA, Sethupathy P, Junkins HA, *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367.

Kerr RA (2007) Global Warming Is Changing the World. *Science* **316**, 188-190.

Meyer J DGR (2002) Patterns of heritability of decreased EROD activity and resistance to PCB 126-induced teratogenesis in laboratory-reared offspring of killifish (*Fundulus*

heteroclitus) from a creosote-contaminated site in the Elizabeth River, VA, USA. *Mar Environ Res* **54**, 621-626.

Nacci D, Coiro L, Champlin D, *et al.* (1999) (Adaptation of wild fish populations to persistent environmental contaminants. *Marine Biology* **134**, 9-18.

Ownby DR, Newman MC, Mulvey M, *et al.* (2002) Fish (*Fundulus hereroclitus*) populations with different exposure histories differ in tolerance of creosote-contaminated sediments. *Environmental Toxicology and Chemistry* **21**, 1897-1902.

Paschall JE, Oleksiak MF, VanWye JD, *et al.* (2004) FunnyBase: a systems level functional annotation of *Fundulus* ESTs for the analysis of gene expression. *Bmc Genomics* **5**, 96.

Vignal A, Milan D, SanCristobal M, Eggen A (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genet Sel Evol* **34**, 275-305.

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM (1997) Human Domination of Earth's Ecosystems. *Science* **277**, 494-499.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72.

Weis JS (2002) Tolerance to environmental contaminants in the mummichog, *Fundulus heteroclitus*. *Human and Ecological Risk Assessment* **8**, 933-953.

Westfall PH, Young SS (1993) *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment.* Wiley, New York.

Williams LM, Ma X, Boyko AR, Bustamante CD, Oleksiak MF (2010) SNP identification, verification, and utility for population genetics in a non-model genus. *BMC Genetics* **11**.

Williams LM, Oleksiak MF (2008) Signatures of selection in natural populations adapted to chronic pollution. *BMC Evolutionary Biology* **8**.

Wirgin I, Waldman JR (2004) Resistance to contaminants in North American fish populations. *Mutat Res* **552**, 73-100.

Xu L, Chen H, Hu X, et al. (2006) Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms. *Mol. Biol. Evol.* **23**, 1107-1108.

Table 1. Genes and loci with SNPs identified as outliers using the F_{ST} modeling approach, Association Test, and Minor Allele Frequency-F_{MAX} Test (MAF-F_{MAX}). P-values for the association test, the likelihood ratio test for model fit and MAF-F_{MAX} test are listed. P-values for the association test significant with a Bonferroni correction are bold. Genes and loci in blue are identified as outliers using both the F_{ST} modeling approach and the MAF-F_{MAX} test. Genes and loci in green are identified as outliers using the F_{ST} modeling approach and the Association test. Genes and loci in purple are identified as outliers using the Association test and the MAF-F_{MAX} test. Genes and loci in red are identified as outliers using all three tests.

New Bedford Harbor					
F_{ST} modeling approach	**Association Test**	***p-value***	***Likelihood Ratio Test p-value***	**MAF-F_{MAX} Test**	***p-value***
ATP synthase D chain, mitochondrial	AMBP protein precursor	$1.2x10^{-3}$	$1.1x10^{-2}$	ATP synthase D chain, mitochondrial	$1.3x10^{-4}$
ATP synthase subunit delta	Biotinidase fragment 2	$2.4x10^{-4}$	$1.4x10^{-3}$	ATP synthase subunit f, mitochondrial	$1.4x10^{-4}$
ATP synthase subunit f, mitochondrial	CYP1A_PP (-707)	$1.2x10^{-3}$	$2.1x10^{-3}$	CYP1A_PP (-618)	$2.3x10^{-3}$
Atrial natriuretic peptide	CYP1A_PP (-618)	**$1.1x10^{-5}$**	$1.3x10^{-5}$	CYP1A_PP (+268)	$7.6x10^{-5}$
CYP1A_PP (-618)	CYP1A_PP (+268)	$5.5x10^{-4}$	$1.5x10^{-3}$	FABP	$1.2x10^{-4}$
CYP1A_PP (+268)	Elastase 3	$2.6x10^{-3}$	$4.1x10^{-3}$	Fibrinogen, beta polypeptide	$1.3x10^{-9}$
Elastase 3	FABP	$1.5x10^{-4}$	$4.7x10^{-4}$	Glyceraldehyde 3-phosphate dehydrogenase	$8.7x10^{-5}$
FABP	Fibrinogen, beta polypeptide	**$2.4x10^{-10}$**	$7.4x10^{-10}$	Nuclease diphosphate kinase B	$9.9x10^{-5}$
Fibrinogen, betapolypeptide	Guanine nucleotide-binding protein subunit β-2-like 1	$1.2x10^{-3}$	$1.5x10^{-2}$	Peptidyl-prolyl cis-trans isomerase	$1.2x10^{-5}$
Glyceraldehyde 3-phosphate dehydrogenase	NADH dehydrogenase [ubiquinone] 1 β subcomplex subunit 2, mitochondrial precursor	$1.8x10^{-3}$	$6.0x10^{-3}$	Peroxiredoxin-6	$3.9x10^{-3}$
Guanine nucleotide-binding protein subunit β-2-like 1	NADH dehydrogenase 1 α subcomplex subunit 4	**$2.1x10^{-5}$**	$6.6x10^{-4}$	Pleurocidin-like peptide WF3 precursor	$4.6x10^{-15}$
NADH dehydrogenase [ubiquinone] 1 β subcomplex subunit 2, mitochondrial precursor	Nuclease diphosphate kinase B	**$7.3x10^{-5}$**	$8.2x10^{-5}$	Ribosomal protein L7a	$1.1x10^{-7}$

Table 1. Continued

Nuclease diphosphate kinase B	Peptidyl-prolyl cis-trans isomerase	**$9.8x10^{-6}$**	$1.0x10^{-5}$	Ribosomal protein S10	$4.9x10^{-15}$
Peptidyl-prolyl cis-trans isomerase	Peroxiredoxin-6	**$8.1x10^{-7}$**	$1.0x10^{-3}$	Ribosomal protein S29	$5.5x10^{-3}$
Peroxiredoxin-6	Ribosomal protein S29	**$2.9x10^{-7}$**	$6.8x10^{-6}$	Selenoprotein Pa precursor putative mRNA	$1.2x10^{-4}$
Ribosomal protein L7a	Selenoprotein Pa precursor putative mRNA	$2.9x10^{-4}$	$8.5x10^{-4}$	Serotransferrin precursor	$2.5x10^{-3}$
Ribosomal protein S15	Serotransferrin precursor	$1.6x10^{-3}$	$1.7x10^{-3}$	Similar to yeast hypothetical protein DSM70294	$5.3x10^{-8}$
Selenoprotein Pa precursor putative mRNA	Similar to yeast hypothetical protein DSM70294	**$3.8x10^{-9}$**	$5.8x10^{-9}$	Thioredoxin	$7.1x10^{-5}$
Similar to yeast hypothetical protein DSM70294	Thioredoxin	**$2.1x10^{-7}$**	$2.3x10^{-6}$	Unknown (AFLP 330_77)	$2.2x10^{-4}$
Thioredoxin	Trypsinogen Y	$1.4x10^{-3}$	$2.9x10^{-3}$	Unknown (AFLP 886_60)	$4.1x10^{-3}$
Trypsinogen Y	Unknown (AFLP 1230_196)	$2.8x10^{-3}$	$9.3x10^{-3}$	Vitellogenin I	$2.3x10^{-4}$
Unknown (AFLP 1230_196)	Vitellogenin	$1.6x10^{-3}$	$3.8x10^{-3}$	60S ribosomal protein L41	$5.0x10^{-6}$
Unknown (AFLP 210_146)	Vitellogenin	$5.1x10^{-4}$	$3.8x10^{-3}$		
Vitellogenin	Vitellogenin I	$4.3x10^{-4}$	$4.3x10^{-4}$		
Vitellogenin I	Vitellogenin-1 precursor	$2.8x10^{-4}$	$3.9x10^{-3}$		
Zona radiata-3	40S ribosomal protein S24	**$6.6x10^{-5}$**	$1.4x10^{-3}$		
40S ribosomal protein S24	40S ribosomal protein S26	$8.3x10^{-4}$	$6.1x10^{-3}$		
40S ribosomal protein S26	60S ribosomal protein L24	$2.2x10^{-4}$	$1.0x10^{-2}$		
60S ribosomal protein L24	60S ribosomal protein L27	$1.8x10^{-4}$	$1.2x10^{-2}$		
60S ribosomal protein L41	60S ribosomal protein L41	**$2.6x10^{-7}$**	$4.2x10^{-7}$		

Newark Bay

F_{ST} modeling approach	**Association Test**	***p-value***	***Likelihood ratio test p-value***	**MAF-F_{MAX} Test**	***p-value***
B2-microglobulin	B2-microglobulin	**$1x10^{-5}$**	$2.3x10^{-5}$	B2-microglobulin	$2.4x10^{-4}$
CYP1A_PP (+268)	CYP1A_PP (-618)	**$1.6x10^{-6}$**	$2.0x10^{-3}$	CYP1A_PP (+286)	$1.5x10^{-3}$
Ferritin, middle subunit (Ferritin M)	CYP1A_PP (+268)	$9.6x10^{-4}$	4.0x10-3	Ferritin, middle subunit (Ferritin M)	$4.8x10^{-3}$
Lysozyme C precursor	Ribosomal protein	$7.1x10^{-3}$	$1.0x10^{-2}$	Lysozyme C	$5.5x10^{-11}$

Table 1. Continued

	S2			precursor	
Ribosomal protein S2	Similar to gDNA of *Danio rerio* clone #CH211-27352	$6.2x10^{-3}$	$1.3x10^{-2}$	Ribosomal protein S2	$5.2x10^{-4}$
Similar to gDNA of Danio rerio clone #CH211-27352	Similar to mouse clone RP24-528E17	$3.1x10^{-3}$	$5.0x10^{-3}$	Similar to gDNA of *Danio rerio* clone #CH211-27352	$6.0x10^{-3}$
Similar to mouse clone RP24-528E17	Zona Pellucida binding protein	$5.3x10^{-3}$	$9.0x10^{-2}$	Similar to mouse clone RP24-528E17	$1.1x10^{-4}$
Thioredoxin				Unknown (TC17861_587)	$7.8x10^{-3}$
Zona pellucida binding protein				Zona Pellucida binding protein	$3.6x10^{-4}$

Elizabeth River

F_{ST} modeling approach	Association Test	*p-value*	*Likelihood Ratio Test p-value*	MAF-F_{MAX} Test	*p-value*
Actin, alpha cardiac	Astacin like metallo-protease	$2.6x10^{-3}$	$6.3x10^{-3}$	Astacin like metallo-protease	$4.9x10^{-3}$
Adult beta-type globin	B2-microglobulin	$8.5x10^{-5}$	$2.2x10^{-3}$	B2-microglobulin	$9.5x10^{-3}$
Astacin like metallo-protease	Chymotrypsinogen 2-like protein	$2.6x10^{-4}$	$2.6x10^{-4}$	Bf/C2 protein	$1.7x10^{-3}$
Bf/C2 protein	CYP1A_PP (+268)	$3.7x10^{-4}$	$6.9x10^{-4}$	Chymotrypsinogen 2-like protein	$1.3x10^{-3}$
Chymotrypsinogen 2-like protein	CYP1A_PP (-618)	$5.1x10^{-3}$	$2.7x10^{-2}$	CYP1A_PP (+268)	$4.7x10^{-4}$
CYP1A_PP (+268)	Drtp1	$1.7x10^{-3}$	$1.9x10^{-2}$	Fibrinogen, gamma polypeptide	$1.2x10^{-4}$
CYP1A_PP (-618)	Eukaryotic translation elongation factor 2	$7.3x10^{-3}$	$9.1x10^{-3}$	Leukocyte elastase inhibitor	$2.2x10^{-4}$
CYP1A_PP (-707)	Fibrinogen, gamma polypeptide	**$2.0x10^{-6}$**	$8.0x10^{-6}$	Nascent polypeptide-associated complex subunit α	$3.4x10^{-4}$
Drtp1	NADH dehydrogenase [ubiquinone] 1-β subcomplex subunit 2, mitochondrial precursor	$3.2x10^{-3}$	$5.7x10^{-3}$	Peroxiredoxin-6	$3.9x10^{-3}$
Eukaryotic translation elongation factor 2	Nascent polypeptide-associated complex subunit alpha	**$3.2x10^{-5}$**	$4.7x10^{-5}$	Ribosomal protein L7	$2.5x10^{-3}$
Fibrinogen, gamma polypeptide	Nuclease diphosphate kinase B	$6.3x10^{-3}$	$6.3x10^{-3}$	Ribosomal protein S9	$1.4x10^{-4}$

Table 1. Continued

Leukocyte elastase inhibitor	Pancreatic progenitor cell differentiation and proliferation factor b	$7.9x10^{-3}$	$2.5x10^{-2}$	Ribosomal protein S20	$2.1x10^{-4}$
L-SF precursor	Peroxiredoxin-6	$1.4x10^{-3}$	$1.9x10^{-3}$	SEC61, gamma subunit isoform 1	$9.7x10^{-5}$
NADH dehydrogenase [ubiquinone] 1-β subcomplex subunit 2, mitochondrial precursor	Ribosomal protein L7	$2.6x10^{-3}$	$3.2x10^{-2}$	Similar to chimp clone PTB-146D01	$4.7x10^{-4}$
Nascent polypeptide-associated complex subunit alpha	Ribosomal protein S9	$\mathbf{1.1x10^{-5}}$	$1.1x10^{-5}$	Similar to mouse clone RP24-528E17	$5.3x10^{-5}$
Nuclease diphosphate kinase B	Ribosomal protein S20	$\mathbf{4.0x10^{-6}}$	$8.0x10^{-6}$	Unknown (AFLP 122_129)	$3.9x10^{-3}$
Pancreatic progenitor cell differentiation and proliferation factor b	SEC61, gamma subunit isoform 1	$1.3x10^{-4}$	$6.3x10^{-3}$	Unknown (AFLP 56_82)	$3.9x10^{-3}$
Peroxiredoxin-6	Similar to chimp clone PTB-146D01	$3.8x10^{-4}$	$4.5x10^{-4}$	Unknown (AFLP 1461_65)	$3.5x10^{-3}$
Phosvitinless vitellogenin	Similar to mouse clone RP24-528E17	$\mathbf{3.4x10^{-5}}$	$3.4x10^{-5}$		
Ribosomal protein L34	TBT-binding protein	$\mathbf{1.4x10^{-5}}$	$2.2x10^{-5}$		
Ribosomal protein L7	TBT-binding protein 2	$2.2x10^{-3}$	$2.3x10^{-2}$		
Ribosomal protein S9	Trypsinogen 2	$5.5x10^{-4}$	$6.2x10^{-3}$		
Ribosomal protein S20	Unknown (AFLP 11_129)	$2.0x10^{-3}$	$5.8x10^{-3}$		
Ribosomal protein S6	Unknown (AFLP 56_82)	$2.8x10^{-3}$	$3.9x10^{-3}$		
SEC61, gamma subunit isoform 1	Unknown (AFLP 136_94)	$1.2x10^{-3}$	$4.1x10^{-3}$		
Selenoprotein Pa precursor putative mRNA	Unknown (AFLP 122_129)	$1.4x10^{-2}$	$4.1X10^{-2}$		
Similar to chimp clone PTB-146D01	Unknown (AFLP 525_183)	$5.2x10^{-4}$	$4.2x10^{-3}$		
Similar to *Gasterosteus aculeatus* clone VMRC26-21C14	Unknown (AFLP 987_45)	$1.7x10^{-3}$	$7.4x10^{-3}$		
Similar to mouse clone RP24-528E17	Unknown (AFLP 1079_383)	$2.1x10^{-3}$	$3.4x10^{-2}$		
TBT-binding protein	Unknown (AFLP 1461_65)	$2.9x10^{-3}$	$4.9x10^{-3}$		
Unknown (AFLP 11_129)					
Unknown (AFLP 56_82)					
Unknown (AFLP 122_129)					
Unknown (AFLP 136_94)					
Unknown (AFLP 280_83)					

Table 1. Continued

Unknown (AFLP 510_115)						
Unknown (AFLP 987_45)						
Unknown (AFLP 1197_172)						
Unknown (AFLP 1303_220)						
Unknown (AFLP 1328_135)						
Unknown (AFLP 1461_65)						
Unknown (TC16623_751)						
Zona pellucida binding protein						

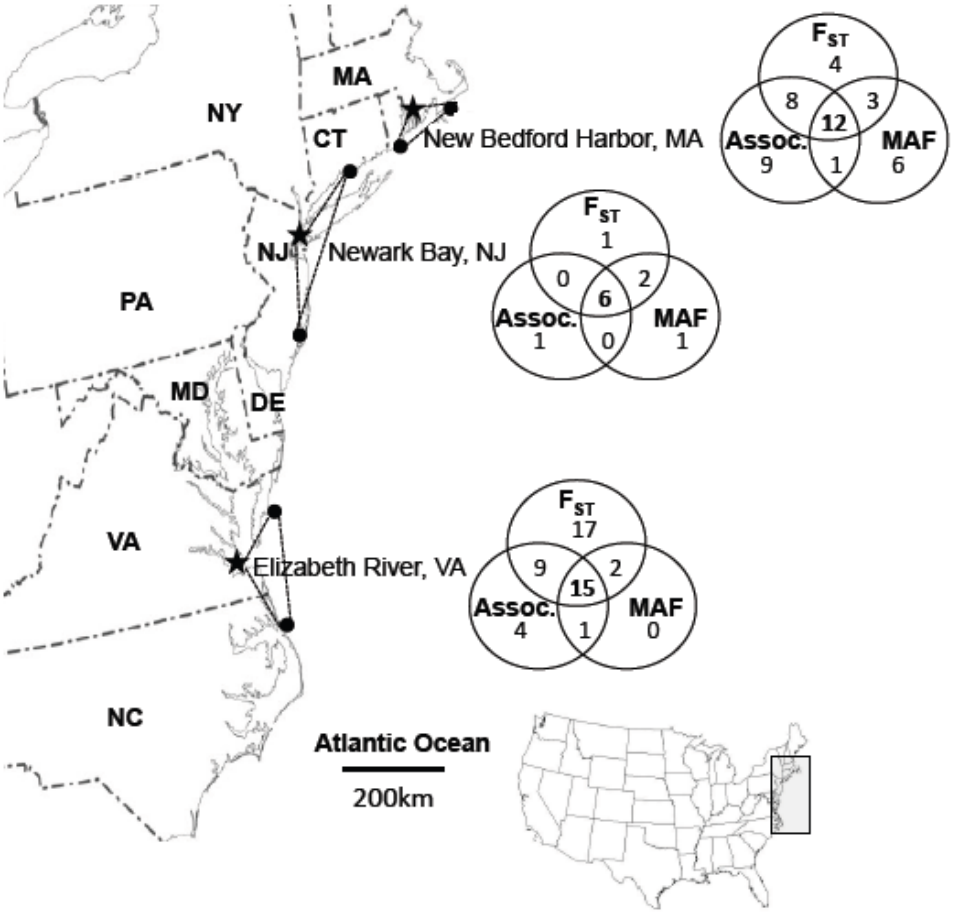
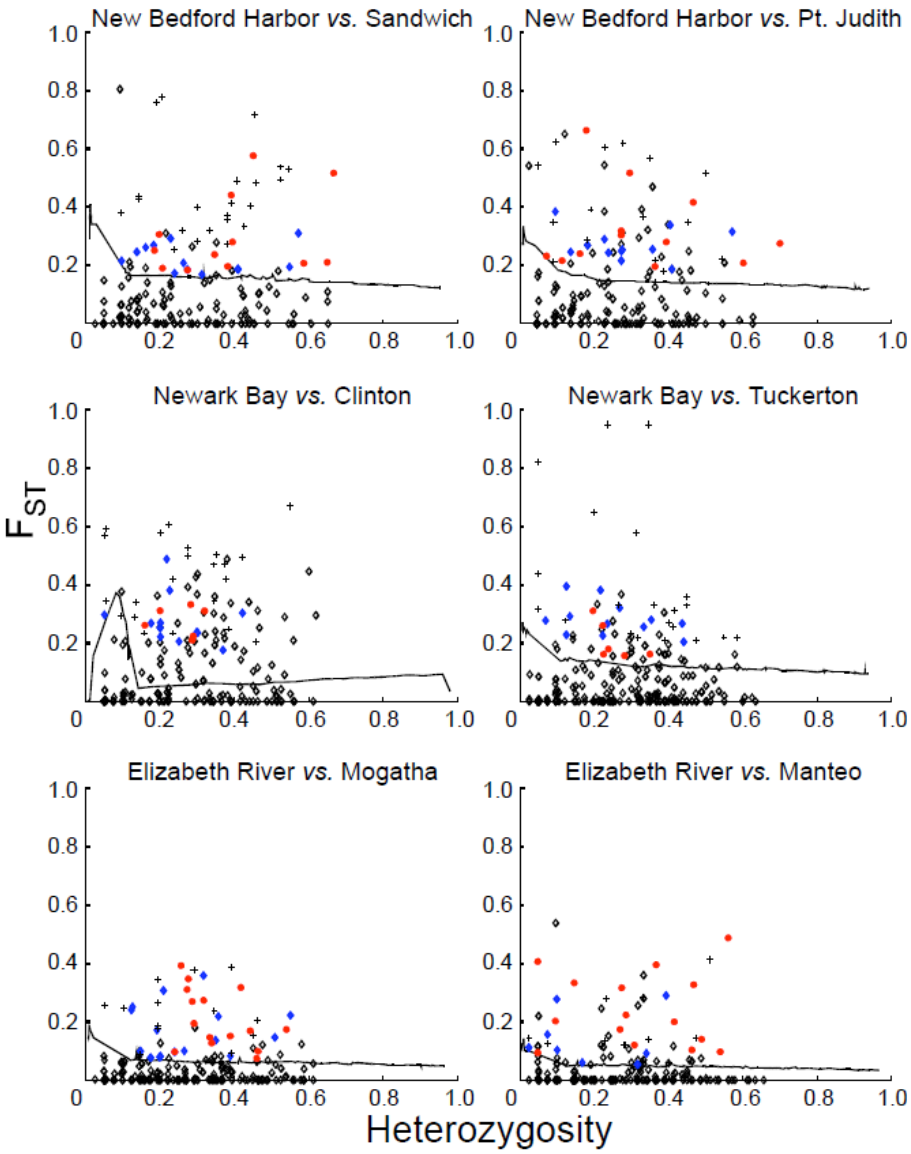


Figure 1. ***F. heteroclitus*** **sampling sites along the East coast of the United States.** Polluted sites are starred and flanked north and south by clean reference sites (circles) to form a triad. Venn diagrams indicate the number of SNPs exhibiting non-neutral behavior using the three statistical tests: the F_{ST}-modeling approach (F_{ST}), Association (Assoc.), and MAF-F_{MAX} (MAF).

Figure 2. F_{ST}-modeling approach to detect selection. Empirical F_{ST} values are plotted against heterozygosity. The line demarks the 99th percentile estimated from a simulation model. Blue diamonds indicate SNPs that are significantly different between the polluted population and both reference populations but not different for reference *versus* reference. Red dots are superimposed on blue diamonds if the SNP was also significant in the other two statistical tests. Less interesting are the crosses and open diamonds. Black crosses are outliers also in the reference *versus* reference comparison. Open diamonds represent outliers where the polluted population was only significant in comparison to one reference population.


New Bedford Harbor vs. Sandwich
New Bedford Harbor vs. Pt. Judith
Newark Bay vs. Clinton
Newark Bay vs. Tuckerton
Elizabeth River vs. Mogatha
Elizabeth River vs. Manteo
F_ST
Heterozygosity
0
0.2
0.4
0.6
0.8
1.0

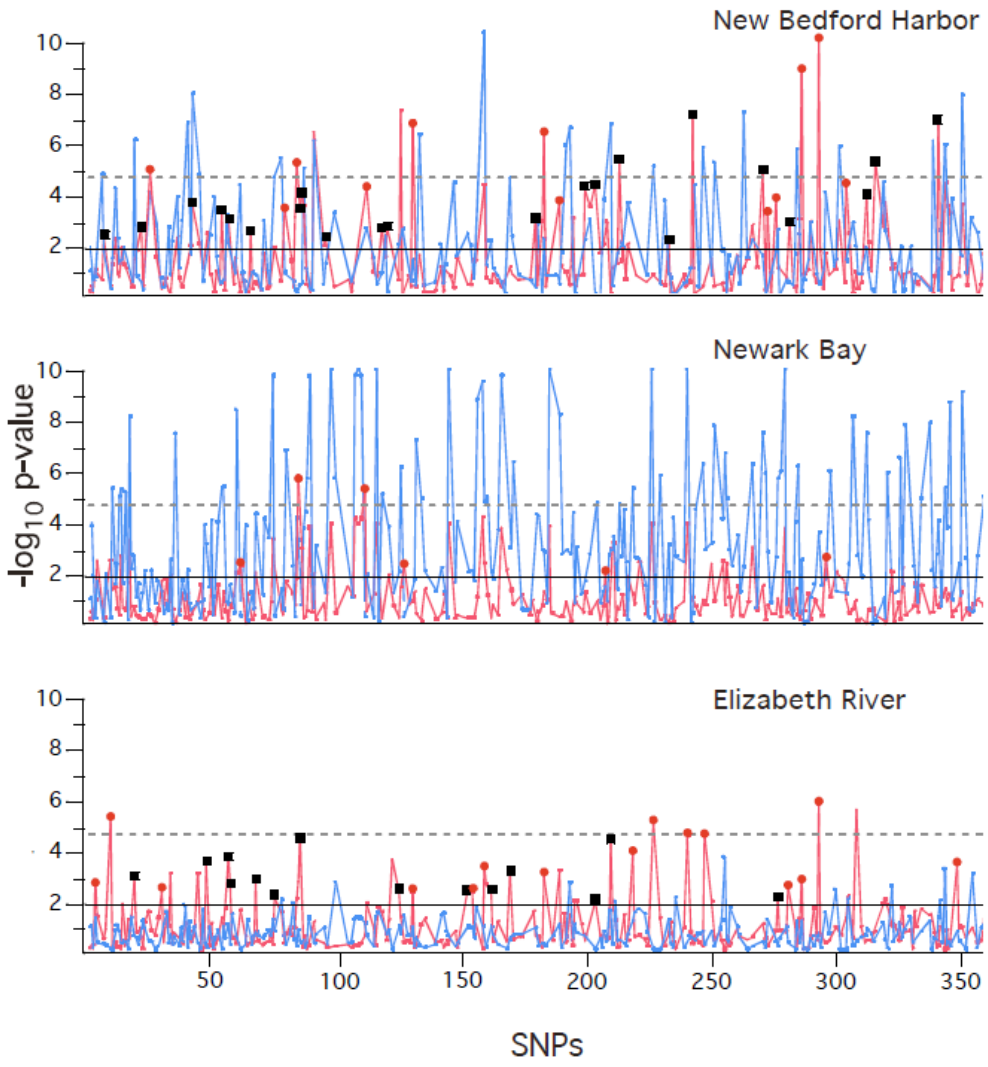


Figure 3. Association test for detection of selection. Likelihood of association of each SNP with either the polluted site (red line) or reference sites (blue line) as a $-\log_{10}$p-value. The $-\log_{10}$p-value of 2 is marked by a black line, and the Bonferroni correction for multiple testing is marked by the dotted grey line ($-\log_{10}$p-value of 4.56). SNPs are identified as outliers in polluted sites *versus* reference sites if the polluted association value is greater than 2 and the likelihood ratio test p-value of polluted *versus* reference association is ≤0.01 (supplemental Table S1). Black squares indicate those SNPs where the likelihood model for pollution significantly exceeds the model based on divergence among reference sites. Red dots are superimposed on black squares if the SNP was also significant in the other two statistical tests.

Figure 4. MAF-F_{MAX} test for detection of differences in SNP allele frequencies between polluted and reference sites. The allele frequency of the triad-wide minor allele was calculated and plotted for all SNPs. Columns are collection sites arranged north to south, and each row represents an individual SNP. SNPs with allele frequencies significantly different in an ANOVA using F_{MAX} (Westfall, Young, 1993) to control for type I errors among iterations (F_{MAX}: empirical F-value exceeds the top 1% of all permutated F-values assuming random population differentiation) between polluted and both reference sites are plotted. Red dots denote SNPs exhibiting non-neutral behavior in all three statistical tests. The SNP exhibiting non-neutral behavior in all three triads and using all tests (CYP1A +268) is boxed.


North
South
0 0.125 0.25 0.375 0.5
Minor Allele Frequency
New Bedford Harbor
FABP
Serotransferrin precursor
Similar to zebrafish clone CH73-181B19
Similar to human clone CTB-7N3
Glyceraldehyde 3-PD
ATP synthetase D chain, mito.
Selenoprotein Pa precursor protein
ATP synthetase subunit F, mito.
CYP1A PP (-618)
Nuclease diphosphate kinase B
Similar to human clone RP11-119H12
Pleurocidin-like peptide WF3 precursor
CYP1A (+268)
Peroxiredoxin-6
Ribosomal protein L7a
Similar to yeast hyp. Prot. DSM70294
Fibrinogen, beta
Peptidyl-prolyl cis-trans isomerase
Vitellogenin I
Thioredoxin
60S ribosomal protein L41
Ribosomal S10
Newark Bay
CYP1A (+268)
B2-microglobulin
Similar to mouse clone RP24-528E17
Unknown (TC17861_587)
Similar to zebrafish clone CH21127352
Lysozyme C precursor
Ribosomal protein S2
Zona pellucida binding protein
Ferritin, middle subunit (Ferritin M)
Elizabeth River
CYP1A (+268)
Similar to chimp. clone PRB146D01
40S Ribosomal protein S9
Nascent polypeptide-assoc. complex α
Ribosomal protein S20
Chymotrypsinogen 2-like protein
Ribosomal protein L7
Similar to mouse clone RP24-528E17
Astacin-like metallo-protease
Unknown (AFLP 1461_65)
B2-microglobulin
Unknown (AFLP 56_82)
Fibrinogen, gamma
Unknown (AFLP122_129)
Peroxiredoxin-6
SEC61, gamma subunit isoform 1
Bf/C2 protein
Leukocyte elastase inhibitor mRNA

CHAPTER 4

Cytochrome P4501A promoter polymorphisms and activity in natural populations

Larissa M. Williams[1] and Marjorie F. Oleksiak[2]

1. Department of Environmental and Molecular Toxicology
Box 7633, North Carolina State University
Raleigh, NC 27695-7633 USA

2. Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149 USA

Corresponding Author:
Marjorie Oleksiak
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149
Fax: 305-421-4600
Email: moleksiak@rsmas.miami.edu

Abstract

Cytochrome P4501A (CYP1A) has been shown to be refractory to prototypic inducers in populations of the estuarine minnow, *Fundulus heteroclitus*, adapted to chronic anthropogenic pollution. Two SNPs in the promoter and first intron of the CYP1A promoter were previously found to be under selection, indicating that natural selection was acting on the CYP1A promoter or loci linked to these SNPs. In order to understand the role of the CYP1A promoter and these selectively important SNPs, the promoter and first intron and exon were sequenced in multiple individuals in one polluted (New Bedford Harbor) population resistant to PCBs and other anthropogenic contaminants and two reference populations north and south of the polluted site. The CYP1A promoter was extremely variable (an average of 21% of the promoter nucleotides varied among all populations) and there were no fixed differences between populations. There was little variation in known and predicted transcription factor binding sites between populations. There was also no signature of selection along the promoter, and the SNPs found to be under selection in a prior study were not in linkage disequilibrium with the remainder of the promoter or the first intron and exon. The inducibility of the promoter was also explored; the promoter was induced by a prototypic PAH, 3-methylcholanthrene, in a dose-dependent manner. When promoters from multiple individuals per population were tested, the CYP1A promoter was most induced in the polluted New Bedford Harbor population as compared to both reference populations. This is in contrast to CYP1A expression *in vivo* which is refractory to induction in New Bedford Harbor individuals. Overall, the variation in the promoter does not explain the variation in *in vitro* promoter expression in all tested populations. These results indicate that

the underlying mechanism for the *in vivo* transcriptional phenotype may lay further upstream or downstream of the CYP1A promoter region which was sequenced (1600bp), or involve proteins which were not available in the cell line (top minnow) in which the transfection assays were completed.

Introduction

The estuarine minnow, *Fundulus heteroclitus*, inhabit several of the most heavily polluted estuaries in the world (Wirgin, Waldman, 2004), including New Bedford Harbor, Massachusetts, a Superfund site polluted with high levels of polychlorinated biphenyls (PCBs) (Weaver, 1983). While this site has only been contaminated for approximately 50 years, *F. heteroclitus* has adapted to its seemingly inhospitable conditions. Recently, two SNPs in the proximal promoter of the hepatic monooxygenase, cytochrome P4501A (CYP1A), were shown to be under selection in the New Bedford Harbor population as compared to clean, reference populations (Williams, Oleksiak, 2010) indicating a genetic underpinning for adaptation. Furthermore, *F. heteroclitus* individuals from NBH are refractory to CYP1A induction by prototypic inducers (Bello *et al.*, 2001; Nacci *et al.*, 1999) and have lower CYP1A activity as compared to reference populations. The basis for this refractory phenotype has not been characterized, but could involve the SNPs under selection either directly or through linkage to areas responsible for modulating CYP1A transcription.

CYP1A has been used as a biomarker of environmental exposure to anthropogenic contamination (M. Nilsen *et al.*, 1998) such as polycyclic aromatic hydrocarbons (PAH; (Arinc *et al.*, 2000; Collier *et al.*, 1998; Hahn, 1998; Huggett *et al.*, 2006; M. Nilsen *et al.*,

1998; Nacci *et al.*, 2002a; Wirgin, Waldman, 2004), polychlorinated biphenyls (PCB) (Chen *et al.*, 2009; Fiedler *et al.*, 1998; Gunawickrama *et al.*, 2008; Hahn, 1998; Lake *et al.*, 1995; Lubet *et al.*, 1992; Wirgin, Waldman, 2004) and other planar halogenated aromatic hydrocarbons (PHAH) (Garrick *et al.*, 2006; Hahn, 1998; Powell *et al.*, 2000; Wilson *et al.*, 2005; Wirgin, Waldman, 2004). CYP1A transcription is mainly regulated through the aryl hydrocarbon receptor (AHR) pathway (Whitlock *et al.*, 1996). The AHR is a highly conserved ligand-activated transcription factor that is primarily found in the cytoplasm, complexed with Hsp90 and immunophilin-like Ara9 (also known as XAP2 or AIP). Upon interaction with xenobiotic ligands such as PCBs and PAHs (Denison, Nagy, 2003), AHR translocates to the nucleus, disassociates with Hsp90, and dimerizes with the aryl hydrocarbon receptor nuclear translocator (ARNT) (Reyes *et al.*, 1992). The AHR-ARNT heterodimer interacts with gene-regulatory elements commonly referred to as dioxin response elements (DRE) or xenobiotic response elements (XRE) (Murre *et al.*, 1989), disrupting chromatin structure and facilitating the interaction of additional transcription factors and co-regulatory proteins with the promoters of target genes, including CYP1A (Beischlag *et al.*, 2008).

CYP1A transcription, total protein and enzyme activity has been well described in the estuarine minnow species, *F. heteroclitus.* *F. heteroclitus* lives along the east coast of the United States in both urban and pristine estuaries where it is one of the most abundant intertidal marsh fishes (Burnett *et al.*, 2007). *F. heteroclitus* is a non-migratory species (Skinner *et al.*, 2005), living in large subpopulations (Adams *et al.*, 2006). These large subpopulations maintain high standing genetic variation (Oleksiak, 2010), allowing quick

adaptation to changing environmental conditions. Over the last 150 years, the east coast has become highly urbanized, leading to anthropogenic contamination of the waters and sediment of the estuaries. Many of these urban sites have been designated as Superfund sites: hazardous waste sites mandated for clean up by the Comprehensive Environmental Response, Compensation, and Liability Act of 1980.

One notable Superfund site is New Bedford Harbor (NBH), also known as the Acushnet River Estuary. From the 1940s to 1978, two capacitor manufacturing facilities discharged PCBs and heavy metals both directly and indirectly (through the sewer system) into the harbor (Weaver, 1983). This activity made NBH one of the most PCB contaminated sites in the United States, with concentrations of PCBs in the sediment exceeding 100,000 parts per million (Weaver, 1983) in some areas and 100-2,000 μg/g dry sediment in other areas of the estuary (Lake *et al.*, 1995; Pruell *et al.*, 1990). NBH also contains high levels of chlorinated dibenzofurans, PAHs, and metals such as Cr, Cu, Cd, and Pb (Pruell *et al.*, 1990; Shine *et al.*, 1995; Weaver, 1983). While much of the flora and fauna that lived in NBH pre-urbanization can no longer survive there, *F. heteroclitus* has flourished in this seemingly toxic environment (Nacci *et al.*, 2001; Nacci *et al.*, 2002b), despite bioaccumulating whole body PCB concentrations of 1,370 μg/g dry weight and non-*ortho* and mono-*ortho* PCB congener concentrations of 30-35 μg/g dry weight in the liver (Black *et al.*, 1998; Lake *et al.*, 1995). The mechanism(s) by which *F. heteroclitus* has adapted to this environment has been the attention of much research for several decades and is focused around the CYP1A xenobiotic metabolizing gene and associated pathways.

F. heteroclitus individuals from NBH are refractory to CYP1A induction by prototypic inducers (Arzuaga, Elskus, 2010; Bello *et al.*, 2001; Nacci *et al.*, 1999) and have lower CYP1A activity as compared to reference populations. NBH fish and offspring are also insensitive to the lethal effects of PCBs (Nacci *et al.*, 2002a) and accumulate fewer DNA adducts when exposed to the PAH, benzo(a)pyrene (Nacci *et al.*, 2002b). The mechanism of differential sensitivity between NBH and reference populations has been proposed to involve the AHR-ARNT-CYP1A signaling pathway and may involve both heritable and non-heritable mechanisms. The genetic basis of the resistance phenotype has not been described. Efforts have been made to explore the role of the AHR pathway; no changes in gene expression of AHR (1 and 2), ARNT, or repressor (AHRR) have been found between NBH and reference populations (Karchner *et al.*, 2002; Powell *et al.*, 2000). While the AHR1 (Hahn *et al.*, 2002; Hahn *et al.*, 2004) and AHR2 (Hahn *et al.*, 2005) loci are polymorphic, and there are different allele frequencies in NBH and reference populations that may have been driven by selective pressures related to polluted environments (Hahn *et al.*, 2002; Hahn *et al.*, 2004), no variants have been shown to be functionally important *in vitro* (Hahn *et al.*, 2004).

Single nucleotide polymorphisms in the CYP1A promoter have recently been shown to be under selection or linked to areas of the genome under selection in *F. heteroclitus* from NBH as compared to two reference sites (Williams, Oleksiak, 2010). One selectively important SNP is located at -670 bp upstream of the +1 transcriptional start site and the other at 173 bp into the first intron of the CYP1A gene. These selectively important SNPs may

alter transcription factor binding or be linked to areas affecting transcription and may be responsible for the refractory induction of CYP1A in the NBH population.

To further explore the CYP1A promoter and the role of these SNPs in resistance to anthropogenic contaminant mixtures, we sequenced 1,600 bp of the promoter and the first (non-coding) exon and intron in multiple individuals from NBH and reference sites to assess promoter-wide patterns of variation and to conduct evolutionary analyses among and between populations. We also tested the inducibility of each promoter to a prototypic PAH using a transient luciferase transfection assay to determine CYP1A promoter activities within and between populations.

Materials and Methods

Genomic Clones

F. heteroclitus were collected using minnow traps during the spring of 2005. Fin clips were sampled from eight individuals from each of the three collection sites along the Atlantic Coast of the United States (Figure 1). One of the three collection sites was an EPA Superfund site: New Bedford Harbor (EPA ID: MAD980731335) and the other two represent clean reference sites which flanked the polluted site equidistant north and south. *Fundulus grandis*, a sister species to *F. heteroclitus*, was collected from Port O'Connor, Texas in the spring of 2008.

Genomic DNA from fin clips was extracted using a modified version of Aljanabi and Martinez (Aljanabi, Martinez, 1997), and DNA was resuspended in 50 µL 0.1X TE buffer. 2kb of the CYP1A promoter and first intron and two exons were amplified using PCR from

the genomic DNA of each individual. The forward primer (5'-AGTTATAGCCACAGTCCAGTCATTT-3') was located 1575 bp upstream of the transcriptional start site and the reverse primer (5'-CAAGGCTATCAAACCCTCAGACAC-3') was 526 bp downstream of the transcriptional start site in the second exon (Powell *et al.*, 2004). Loci were ligated into a Promega pGEM-T vector overnight at 16°C with T4 DNA ligase, transformed into electrocompetant JM109 *E.coli* cells, and grown on carbenicillin selective plates. Colonies were screened for insert with PCR using vector specific primers. Insert PCR products from three clones per individual were sequenced in both directions using ABI BigDye terminator chemistry providing 6x coverage of each nucleotide. Sequences were aligned in MacVector 8.0 (Olson, 1994) and CAP3 (Huang, Madan, 1999).

Sequence Analysis

CYP1A sequences from each of eight individuals in NBH, Point Judith and Sandwich and one *F. grandis* individual were analyzed with DnaSP (Librado, Rozas, 2009) to determine patterns of sequence variation along the promoter as well as in the first and second intron and first exon of the CYP1A gene. To examine evolutionary processes affecting the CYP1A promoter, Tajima's D was calculated across the promoter in all three populations, as well as for each individual nucleotide (Tajima, 1989). Fu and Li's test for natural selection (Fu, Li, 1993) was also performed and the D and F statistics were calculated using *F. grandis* as the outgroup. Average nucleotide substitutions per site within ($P_1$1 and $P_1$2) and between populations (Dxy) was also calculated and plotted using a sliding window with a 50 bp-wide window and a 10 bp step size. A modification of the McDonald-Kreitman test (Mcdonald,

Kreitman, 1991) was used to detect selection following a procedure described by Crawford *et al*. (Crawford *et al.*, 1999) which compares the patterns of sequence variation among functional and nonfunctional regions of the promoter. Functional regions (xenobiotic and glucocorticoid response elements) were ascribed *a priori* (Karchner *et al.*, 1999; Powell *et al.*, 1999; Powell *et al.*, 2004) but not tested for functionality in this study. Additional potential regulatory regions within the promoter were determined with AliBaba 2.1 (http://www.gene-regulation.com/pub/programs/alibaba2/index.html), which predicts binding sites of transcription factors in unknown DNA sequences using binding sites collected in TRANSFAC (Matys *et al.*, 2006), a database containing eukaryotic cis-acting regulatory DNA elements and trans-acting factors. The McDonald-Kreitman test distinguishes between natural selection and neutral evolutionary processes by testing the ratio between fixed functional differences to fixed non-functional differences relative to the ratio of polymorphisms between taxon. A two-tailed Fisher's exact test was used to test the null hypothesis that sequence variation is random. Arlequin 3.5 (Excoffier, Lischer, 2010) was used to determine linkage disequilibrium among nucleotides along the promoter.

Phylogenetic analysis for the whole promoter, functional and non-functional regions was applied to the proximal promoter for each of the eight *F. heteroclitus* individuals from the three collection sites as well as the outgroup species, *F. grandis* using maximum parsimony methods in PAUP* 4.0 (Swofford, 2003).

Reporter Gene constructs

Reporter gene constructs for transfection experiments were generated by digesting

CYP1A inserts from the pGEM-T vector first with ApaI, blunting the insert with T4 DNA polymerase and further digesting with SacI restriction enzymes. Promega pGL3-Basic vector containing the Firefly luciferase gene was linearized by digesting with KpnI, blunting with T4 DNA polymerase and a final digestion with SacI restriction enzymes. These cuts assisted in the directional cloning of the CYP1A insert into the pGL3-Basic vector upstream of the luciferase gene. The promoter construct was 6,831 bp in length, with 2013 bp being the CYP1A promoter, first intron and first exon (Supplementary Figure 1). The 2013 bp which consisted of the CYP1A promoter, first intron and first exon, the start of transcription occurred at 1630 bp. There was an intact TATA box at 1596bp. Upstream of the transcriptional start site there were three XRE sites at 799bp, 848bp and 1405bp. Digested insert DNA and pGL3-Basic vector was gel purified using Zymoclean Gel DNA recovery kit and ligated together overnight at 16°C. Ligation product was transformed into electrocompetant JM109 *E.coli* cells, and plated on carbenicillin selective plates. Colonies were screened for insert with PCR using vector specific primers.

Cell Culture and transfection

Fish hepatoma cells PLHC-1 (*Poeciliopsis lucida* hepatoma cell) (Ryan, Hightower, 1994) were grown in Gibco CO_2 independent media supplemented with 5% FBS and 50 U ml^{-1} of penicillin G/50 μg ml^{-1} streptomycin in a 30°C incubator. PLHC-1 cells have an intact and inducible CYP1A system and possess properties similar to those observed in *F. heteroclitus* liver (Hahn *et al.*, 1993). Cells were usually split every 4 days by dissociating with 0.05% (w/v) trypsin and 0.5 mM EDTA and subcultured at $2x10^7$ cells per each $75cm^2$

tissue culture flask. Using Fugene 6 transfection reagent (Boehringer), cells at 50–80% confluence were transiently co-transfected in 12-well plates at a density of $2x10^6$ cells per well with the reporter gene vector and an internal control vector expressing *Renilla* luciferase under the control of a cytomegalovirus promoter (pRL-CMV) (Promega). Cells were treated with 3-methylcholanthrene (3-MC) dissolved in DMSO, or with DMSO only. 3-MC is a prototypic inducer of CYP1A through the AHR pathway and evokes a TCDD and related halogenated aromatic hydrocarbon pattern of induction (Poland, Knutson, 1982). Luciferase and *Renilla* activities were measured using the Promega Dual-Luciferase Reporter Assay System.

Results

Proximal and First Exon and Intron Sequences of CYP1A

1630 nucleotides of the CYP1A proximal promoter were sequenced in eight individuals from each of three populations (Sandwich, New Bedford Harbor, and Point Judith). The promoter has a substantial amount of variation, with 21% of the promoter containing variable nucleotides between individuals and populations although there are no fixed differences between populations. One hundred and fifty-seven parsimony informative sites occur along the promoter upstream of the transcriptional start site: 17 are fixed differences between *F. heteroclitus* and its sister species *F. grandis* (Figure 2). The three xenobiotic response elements (Powell *et al.*, 2004) were mainly conserved and had few polymorphisms across populations and species (Supplementary Figure 1). XRE3 had several polymorphisms present only in the New Bedford Harbor population at nucleotides 1406 and

1407, where individuals NBH6 and NBH13 have a TT rather than an AC. The average nucleotide diversity (excluding gaps; N= 1474 bp) for the proximal promoter (Jukes, Cantor, 1969) for *F. heteroclitus* among all three populations is 0.0150 or theta (θ) per site of 0.027. Per population, the average nucleotide diversity is 0.012 for Sandwich (N=1514 bp), 0.021 for New Bedford Harbor (N=1494 bp), and 0.0016 for Point Judith (N=1538 bp).

The first exon and intron also were sequenced, as one of the SNPs under selection (Williams, Oleksiak, 2010) was located in the first intron (at nucleotide 1892). The first exon contained 11 parsimony informative sites (of which one was a fixed difference between *F. heteroclitus* and *F. grandis*), and the first intron contained 34 parsimony informative sites (of which two were fixed difference between *F. heteroclitus* and *F. grandis*). No polymorphisms in the first exon resulted in a nonsynonymous change to the deduced amino acid sequence. The average nucleotide diversity for the first exon was 0.018 for Sandwich (N=111 bp) and 0.009 for New Bedford Harbor (N=112 bp); Point Judith contained no polymorphisms. The average nucleotide diversity for the first intron was 0.017 for Sandwich (N=268), 0.040 for New Bedford Harbor (N=261), and 0.005 for Point Judith (N=264).

Sequence variation distinguished population divergence and established genetic distance by geographical distance (Table 1). Between the two most geographically close sites, Sandwich and New Bedford, the G_{ST} was 0.028 and F_{ST} was 0.054. Between New Bedford Harbor and Point Judith, the G_{ST} was 0.037 and F_{ST} was 0.405. Sandwich and Point Judith, the two furthest sampling sites from each other, had a G_{ST} of 0.040 and a F_{ST} of 0.410. The average number of pairwise nucleotide differences (Kxy) were similar in all three

pairwise comparisons, with 42.11 for Sandwich *vs.* New Bedford, 41.14 for Sandwich *vs.* Point Judith, and 46.09 for New Bedford *vs.* Point Judith (Table 1).

Linkage disequilibrium was compared for all polymorphic nucleotides along the promoter and the first exon and intron. For Sandwich, 558 of the total 12,483 pairwise comparisons of polymorphic loci (4.5%) were in significant LD (Figure 3). New Bedford had 2,175 significant pairwise comparisons in LD, out of a total of 33,871 pairwise combinations (6.4%) (Figure 4). Point Judith had the fewest significant pairwise comparisons (158) but had the highest percentage of significant comparisons of 46.76% among the 338 pairwise comparisons (Figure 5). Sandwich and New Bedford Harbor sequences showed strong LD among nucleotides in the 5' end of the promoter and also among a conserved area of the 3' end of the first intron. At the 3' end of the first intron, the sequence 'ACTT' spanning nucleotides 1970 to 1973, was in LD with a large portion of the promoter. This was also true in the Point Judith population, although the LD of this area with the rest of the promoter spanned fewer nucleotides than was the case in Sandwich and New Bedford Harbor. Point Judith sequences have a large, strong, block of LD in the 5' end of the promoter where a conserved, ancestral insertion occurs. There is no LD among the SNPs (929, 1892 bp) found to be under selection in the previous study (Williams, Oleksiak, 2010) with the remainder of the promoter, intron one or exon one. No functional regions (Supplementary Table 1) were found to be in significant LD with other nucleotides along the promoter or beyond the transcriptional start site with the exception of XRE3, which was in significant LD in the New Bedford Harbor population with nucleotides 1336-1361 (just upstream of the binding XRE3 binding site at 1405-1410).

Evolutionary Analyses of Promoter

Evolution by natural selection does not seem to be a driving force behind the sequence variation found within and among *F. heteroclitus* populations. A McDonald-Kreitman test, which compares functional and non-functional, fixed and polymorphic changes between species (or in this case, also between populations), failed to reject the null hypothesis that the CYP1A promoter is not evolving by natural selection (Table 2). Most of the changes in the promoter were found to be polymorphic in nonfunctional regions.

The variation between *F. heteroclitus* can be visualized in sliding window comparisons (Figure 6). In the comparison of Sandwich to New Bedford Harbor, similar areas of the promoter are equally variable between populations except between nucleotides 1553-1619 (upstream of the TATA box at 1596 bp) where there is a difference between the nucleotide substitutions per site between the two populations. In this area there are several polymorphisms that only exist in one individual in one population, increasing the Dxy value. There is also significant variation within the Point Judith population, although this variation decreases at the very 3' end of the promoter. When compared with New Bedford Harbor, Dxy exceeds the nucleotide substitution per site among populations in the 3' end of the promoter spanning from 1466-1555. In this area, Point Judith individuals do not have any sequence variation, and there is moderate variation in three individuals (4, 6, 13) in the New Bedford Harbor population. When clean, reference sites, Sandwich and Point Judith, are compared, the same area that had a large Dxy value in the Point Judith vs. New Bedford Harbor comparison (1466-1555bp) exhibits the same pattern. In this area, there is no

sequence variation in the Point Judith population, but there is moderate variation in the Sandwich population. There are no clear patterns of variation near functional regions of the promoter between New Bedford Harbor and its clean, reference sites.

A similar lack of a clear pattern in variation along the CYP1A promoter can be visualized by plotting Tajima's D and Fu and Li's D statistics by population (Figure 7). The Tajima test for data from a single locus (Tajima, 1989) compares the estimate of θ based on the average number of pairwise differences to that based on the number of segregating sites. The test of Fu and Li (Fu, Li, 1993) with an outgroup (*F. grandis* in this case) compares estimates of the number of segregating sites to the number of mutations on external branches expected under neutrality. Tajima's D was -0.202 for Sandwich, 0.034 for New Bedford Harbor and 0.496 for Point Judith populations. None of these values were significant at a p-value of 0.05 and no sliding window region was significant along the promoter. Fu and Li's D also was calculated for the promoter in each population and was 2.10 for Sandwich ($p < 0.02$), 2.04 for New Bedford Harbor ($p < 0.02$) and 0.96 for Point Judith ($p > 0.10$) populations. Sandwich, New Bedford Harbor and Point Judith sequences have generally positive D statistics along the promoter signifying low levels of both low and high frequency polymorphisms and indicating a decrease in population size and/or balancing selection. Fu and Li's D statistic was found to be significantly positive in the Sandwich population at the sliding window encompassing nucleotides 1492-1541. This area falls upstream of the TATA box but does not contain any known or predicted transcription factor binding sites. The New Bedford Harbor population also had a significantly positive Fu and Li's D statistic in the sliding windows of nucleotide regions 1241-1353 bp and 1406-1490 bp. Notably, XRE3

falls in the second window. In the Point Judith population, similar patterns of positive D values were found in the Tajima's D and Fu and Li's calculation, although this pattern of variation was not significant.

Phylogenetic analysis also revealed apparently random variation along the promoter. Maximum parsimony was used to construct relationships between individuals, populations and species for the CYP1A promoter (Figure 8). When the full length promoter is used to construct a phylogenetic analysis, Point Judith individuals (which have little sequence variation) fall out together and New Bedford Harbor and Sandwich individuals are interspersed with each other (Figure 8A). One individual from the New Bedford Harbor population (NBH307) falls out with the outgroup, *F. grandis*. This individual has many ancestral nucleotides and indels, which sets it apart from the other individuals both in its own population and in comparison to other populations. A similar, random, pattern is seen when non-functional regions of the promoter are used to construct the phylogenetic analysis (Figure 8B). As noted in the McDonald-Kreitman test, there is little sequence variation in the functional regions of the CYP1A promoter. This is reflected in the phylogenetic analysis, where all *F. heteroclitus* form a monopyletic group (Figure 8C).

Functional Assays

The induction of luciferase expression from two *F. heteroclitus* CYP1A proximal promoter reporter gene contructs by several doses of 3-MC in PLHC-1 cells was tested (Figure 9). Transiently transfected cells were dosed with 3-MC or treated with vehicle (DMSO) control as described in Williams *et al*. (Williams *et al.*, 2000), at concentrations

ranging from 0.001 μM to 2 μM. This test was conducted to assess whether luciferase expression increased in a dose-dependent manner in the clean (Sandwich individual S27) and/or polluted (New Bedford Harbor individual NBH6) CYP1A proximal promoter constructs. At lower doses (<0.05 μM), luciferase expression was similar for both constructs, and not significantly different from vehicle-treated control. At 0.05 μM the New Bedford Harbor construct had significant fold induction over control. At doses of 0.1 μM and above, both Sandwich and New Bedford Harbor constructs had significant fold induction over control. New Bedford also had a significantly higher ($p \leq 0.05$) fold induction in comparison to Sandwich, ranging from 1.5 to 1.9 times greater than Sandwich. A horizontal asymptote occurs at the 1μM dose, indicating that maximal induction of the promoter occurred at or around that dose.

To determine if there are individual and population differences between the inducibility of CYP1A promoters, luciferase activities from four individual promoter constructs were compared within and between the three populations (Figure 10) at a dose of 1 μM 3-MC. The pattern observed was very similar to that of the dose-response curve: reference populations, Sandwich and Point Judith, had on average a fold induction over control of 2.66 and 2.39, respectively, which was lower than the New Bedford Harbor population which had an average fold induction over control of 4.54. Within population induction variability was minimal, ranging from 0.14 to 0.30 over control. A one-way ANOVA determined a significant difference between populations ($p < 0.0001$), and a Tukey's HSD post-hoc test determined that New Bedford Harbor was significantly different

from Sandwich ($p < 0.0001$) and Point Judith ($p < 0.0001$), but Sandwich and Point Judith were not significantly different from each other ($p = 0.258$).

Discussion

The role of the CYP1A proximal promoter in the resistance phenotype to environmental, anthropogenic contamination in the New Bedford Harbor *F. heteroclitus* population had not been investigated prior to this study. Furthermore, an in-depth survey of genetic variation in this promoter in natural populations also had not been conducted. This study sought to determine which regions, if any, of the CYP1A promoter are evolving by natural selection and whether the promoter plays a functional role in the refractory CYP1A inducibility phenotype that has been described in the New Bedford Harbor population (Bello *et al.*, 2001; Nacci *et al.*, 1999).

The CYP1A proximal promoter, cloned from eight individuals from each of three populations, is extremely variable. Up to 21% of the nucleotides in the promoter vary between individuals. A McDonald-Kreitman test indicated that the promoter was not evolving by natural selection due to the fact that there were no nucleotides that were fixed or near fixation. The variation observed for the CYP1A promoter is much greater relative to other promoters that have been sequenced between populations: *Ldh-B* proximal promoter in *F. heteroclitus* (10%; (Crawford *et al.*, 1999); the first intron of *Adh* (1.7%; (Kreitman, 1983), the *eve* enhancer (0.87%; (Ludwig, Kreitman, 1995), and G6pd (1.0%; (Eanes *et al.*, 1993) in *Drosophila melanogaster*; G protein-coupled receptor kinase 4 (0.94%;(Hasenkamp *et al.*, 2008) and CCR5 (2.1%; (Bamshad *et al.*, 2002) in humans; endo16 in the purple sea

urchin (10%; (Balhoff, Wray, 2005); the chalcone synthetase promoter in *Arabidopsis thaliana* (1%; (de Meaux *et al.*, 2005). However, unlike the studies listed above, which showed many of the polymorphisms in functional regions of the promoter, almost all of the variation (92%) for CYP1A exists in regions without a described function.

Plots of variation between populations (Figure 6) revealed that each population surveyed had a substantial and similar amount of variation. It is possible that the observed changes do not alter the pattern of transcription-factor binding and are thus entirely neutral. It has also been hypothesized that stabilizing selection on transcriptional output allows slightly deleterious mutations to persist, compensated for by adaptive changes elsewhere in the promoter and resulting in continuous binding-site turnover (Balhoff, Wray, 2005; Ludwig *et al.*, 2000). This alternative hypothesis is supported for Sandwich and New Bedford populations by the Fu and Li's D test for selection which found significantly positive values for the promoter. Significantly positive values for this statistic reflect an excess of intermediate-frequency alleles, which can result from population bottlenecks, structure and/or balancing selection. Conversely, Tajima's D for each population was not significant. Tajima's D has been found to be more powerful than Fu and Li's D test to test for very strong, recent directional selection (Braverman *et al.*, 1995; Simonsen *et al.*, 1995). However, it has been shown that Tajima's D and Fu and Li's D test have very low power to detect balancing selection (Charlesworth *et al.*, 1995). The one test that does have slightly higher power to detect this type of selection, Fu and Li's F-test (Fu, Li, 1993), found no significant patterns of variation along the CYP1A promoter in any population (data not shown). Given that only one of the three tests for selection was significant, balancing

selection most likely is not acting on the CYP1A promoter to compensate for slightly deleterious mutations, but rather the pattern of variation is neutral. Variation is most likely maintained due to large population sizes and migration. The pattern, but not the significant nature, of variation was entirely different for known functional regions of the promoter.

The regions of known or described function were mainly conserved across populations and species (between *F. heteroclitus* and its sister species *F. grandis*). There were no fixed differences between species or populations and only 13 polymorphisms total in functional regions among all individuals and populations. Despite very high levels of conservation within functional sites, there was no significant selection signal on those areas in Tajima's D or Fu and Li's D. If balancing selection were maintaining those sites, one would also expect to find high linkage disequilibrium between the nucleotides of each site and other functional sites (Charlesworth, 2006; Wall, 1999); this was not the case. In the functional regions where there were polymorphisms, the majority of the polymorphisms were in one or a few individuals across populations. There was a polymorphism in the XRE3, which binds the AHR-ARNT complex less tightly than XRE1 (Powell *et al.*, 2004). In New Bedford Harbor individuals 6 and 13, the XRE site was "C**TT**GCGA" rather than the consensus sequence "C**AC**GCGA". However, the difference in the XRE sequence did not significantly alter the inducibility of the CYP1A promoter construct *in vitro* (for individual NBH6) compared to other NBH individuals assays, which have the consensus XRE3 consensus sequence. While the XREs described in the CYP1A promoter do bind the AHR-ARNT complex *in vitro* (Powell *et al.*, 2004), the functional significance of each of these XREs is still unclear with respect to inducing CYP1A transcription. Deletion constructs

containing all, two, one or none of the XREs coupled to a reporter gene would have to be conducted *in vitro* in order to assess their ability to support AHR-regulated transcription. Furthermore, in order to more accurately assess their role in binding AHR-ARNT *in vivo*, an *in vivo* footprinting assay (Schulte *et al.*, 1995) would be necessary.

For a DNA sequence that has to bind many conserved transcription factors in order to induce the transcription of CYP1A, the promoter is quite variable. However, transfection assays with CYP1A promoter constructs coupled to the luciferase gene showed that the CYP1A promoter functioned *in vitro* to stimulate transcription and was inducible in a dose-responsive manner to a prototypic PAH, 3-MC. That inducibility of luciferase expression differed between two different individual constructs, one from Sandwich and one from New Bedford Harbor given several dosages of 3-MC. Fold induction over control was similar between the constructs below 0.05 μM 3-MC, but at 0.05 μM 3-MC the New Bedford Harbor construct had significant fold induction over control. At doses of 0.1 μM 3-MC and above, both Sandwich and New Bedford Harbor constructs had significant fold induction over control. New Bedford also had a significantly higher ($p \leq 0.05$) fold induction in comparison to Sandwich, ranging from 1.5 to 1.9 times greater than Sandwich. This NBH individual had a similar dose- response curve to CYP1A promoter constructs derived from European flounder (*Platichthys flesus*) which were transfected into HepG2 cells (Williams *et al.*, 2000). In flounder it was shown that doses as low as 0.05 μM 3-MC had a significant fold induction of luciferase expression over vehicle control. Flounder CYP1A was cloned from fish unexposed to environmental contaminants. These observations indicate that New Bedford Harbor derived-CYP1A promoters may function in a manner similar to other teleosts,

regardless of exposure to environmental contaminants. However, since the fold induction of the Sandwich individual was significantly less than that of the New Bedford Harbor individual at the same doses, individual variation in inducibility due to genetic differences (individual or population-level) may be the driving force behind CYP1A promoter activity. Cardiac gene expression variation between *F. heteroclitus* individuals within a population is very high but has been shown to predict variation in metabolism (Oleksiak *et al.*, 2005). Similarly, genetic variation among the CYP1A promoter may predict the inducibility of the promoter by PAHs and PCBs. To explore this theory, we performed a series of transfection assays on constructs derived from four individuals from each of the populations to determine if the pattern of inducibility was based on individual differences or was a shared characteristic of a population.

The CYP1A promoter inducibility was not significantly variable within a population, but was significantly different between polluted and reference populations. New Bedford Harbor had significantly higher fold induction over control as compared to each reference population (Figure 10). The reference populations did not have significantly different average fold induction from each other. There are no fixed differences along the entire promoter or in known or described transcription factor binding sites to explain the significant difference in average fold induction over control between New Bedford Harbor populations and its reference site populations. Given the refractory CYP1A transcriptional phenotype exhibited by New Bedford Harbor fish (Bello *et al.*, 2001; Nacci *et al.*, 1999), it was expected that the CYP1A promoter for this population would be less inducible by prototypic PAHs as compared to populations without environmental contamination histories—the

opposite was found. CYP1A transcription is controlled mainly through the AHR pathway. While there are no known functional differences found in the ability of AHR1 variants to bind TCDD between New Bedford Harbor and Sandwich (Hahn *et al.*, 2004), there are variants specific to individuals from polluted sites that have yet to be tested for functionality in the AHR2 gene as well as the AHR repressor (Hahn *et al.*, 2005). Variants of either of these two proteins may function solely or with other proteins to give both the inducibility and CYP1A refractory phenotypes described. There are also many other enhancers and repressors that may serve to enhance the inducibility of the CYP1A gene while still maintaining the refractory transcriptional phenotype. In humans, the vasoactive intestinal peptide was shown to be regulated by numerous enhancers and repressors in a 600 bp region 4kb upstream of the transcriptional start site (Liu *et al.*, 2001). Thus, the CYP1A promoter may be regulated through many different proteins that have not been identified because they are further up- or downstream from the promoter sequence used for this study, and these areas may contain variants specific to the populations. These could also be in linkage disequilibrium with the two SNPs found to be under selection (Williams, Oleksiak, 2010) but shown in this study not to be in LD with any other nucleotides in the sequenced area. LD has been measured in humans from a few kilobases to in some instances greater than 100 kb (Patil *et al.*, 2001). Without a sequenced genome, the detection of such long distance LD in *F. heteroclitus* is not possible. Thus, if other SNPs responsible for modulating CYP1A transcription are in LD with our selectively important SNPs but are beyond the 1.5 kb sequenced area of this study, they will not be detected. To determine whether the SNPs identified as being under selection were important to the transcriptional phenotype, site-

directed mutagenesis could be used. If the mutagenesis of either or both of the SNPs resulted in a polluted individual having a reference phenotype and vice-versa, the role of these SNPs in the transcriptional regulatory phenotype could be attributed to those particular SNPs.

Several other genes, such as basic leucine zipper nuclear factor 1, CYP1B1, and guanidinoacetate N-methyltransferase, amongst others, have shown a similar gene expression profile in the New Bedford Harbor population as compared to reference populations as CYP1A (Oleksiak *et al*., unpublished). This pattern indicates that there may be shared transcriptional regulation of these genes and others which may involve a singular, or combination of several different, transcription factors, enhancers and/or repressors.

Epigenetic factors may also be causative in this phenotype. One particular epigenetic mechanism, that of cytosine methylation at CpG sites, was explored for the CYP1A promoter in *F. heteroclitus* populations residing in the Elizabeth River, VA Superfund site (Timme-Laragy *et al.*, 2005). Cytosine methylation was not detected at any of the 34 CpG sites examined in any of the total eight adult fish examined, including 3 CpG sites that are part of putative XREs. Thus, this particular epigenetic mechanism could not explain the refractory phenotype of the CYP1A gene in the Elizabeth River polluted population. Similarly Arzuaga *et al*. (Arzuaga *et al.*, 2004) found that the DNA de-methylating agent 5-azacytidine does not restore CYP1A induction in the polluted *F. heteroclitus* population in the Superfund site of Newark Bay suggesting DNA methylation is not responsible for the refractory CYP1A phenotype in that population. These studies do not rule out methylation as a potential mechanism in the New Bedford Harbor population. Due to the shared phenotype between

the three polluted populations, it does however imply it may not be the underlying factor. There are other epigenetic mechanisms which could contribute to this refractory phenotype including gene silencing, position effect, bookmarking, maternal effects, regulation of histone modification and heterochromatin structure, paramutation, and transvection. In order to explore these possibilities, among other possibilities, the use of chromatin immunoprecipitation, epigenetic microarray technologies, DNA adenine methyltransferase identification, fluorescent *in situ* hybridization, and bisulfite sequencing could be used.

The high inducibility of the New Bedford Harbor CYP1A constructs result could also be explained as an artifact of the cell line type. It is possible that the PLHC-1 cell type does not contain all the same transcription factors as does *F. heteroclitus*. Thus, if there is a necessary repressor found in *F. heteroclitus* but not top minnow (*Poeciliopsis lucida*) which would create a lower inducibility of the promoter in the New Bedford Harbor fish, we would not observe this in our study. However, this study did show population differences between New Bedford Harbor and its reference sites, which are not an artifact of the cell culture.

The CYP1A promoter is extremely variable, mainly in non-functional areas and is not evolving through natural selection based on the McDonald-Kreitman test, Tajima's D calculations, and Fu and Li's D calculations. Patterns of variation differ between each of the tested populations; however all are equally variable. Known and predicted functional transcription factor binding sites are well conserved across populations. One site that differed, XRE3 in two New Bedford Harbor individuals, did not seemingly contribute to the inducibility of the CYP1A promoter *in vitro* to a prototypic inducer. Inducibility of the

promoter was dose-dependent, and significantly higher in New Bedford Harbor individuals. This study opens the door to the investigation of other mechanisms involved in the refractory CYP1A transcriptional phenotype observed in the New Bedford Harbor population.

Acknowledgements

Partial funding for this work was received from NIH 5 RO1 ES011588 and NSF OCE 1008542. The authors thank the University of Miami's Molecular Core for sequencing.

References

Adams SM, Lindmeier JB, Duvernell DD (2006) Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. *Molecular Ecology* **15**, 1109-1123.

Aljanabi SM, Martinez I (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research* **25**, 4692-4693.

Arinc E, Sen A, Bozcaarmutlu A (2000) Cytochrome P4501A and associated mixed-function oxidase induction in fish as a biomarker for toxic carcinogenic pollutants in the aquatic environment. *Pure and Applied Chemistry* **72**, 985-994.

Arzuaga X, Calcano W, Elskus A (2004) The DNA de-methylating agent 5-azacytidine does not restore CYP1A induction in PCB resistant Newark Bay killifish (*Fundulus heteroclitus*). *Marine Environmental Research* **58**, 517-520.

Arzuaga X, Elskus A (2010) Polluted site killifish (*Fundulus heteroclitus*) embryos are resistant to organic pollutant-mediated induction of CYP1A activity, reactive oxygen species, and heart deformities. *Environmental Toxicology and Chemistry* **29**, 676-682.

Balhoff JP, Wray GA (2005) Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8591-8596.

Bamshad MJ, Mummidi S, Gonzalez E, *et al.* (2002) A strong signature of balancing selection in the 5 ' cis-regulatory region of CCR5. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10539-10544.

Beischlag TV, Morales JL, Hollingshead BD, Perdew GH (2008) The aryl hydrocarbon receptor complex and the control of gene expression. *Critical Reviews in Eukaryotic Gene Expression* **18**, 207-250.

Bello SM, Franks DG, Stegeman JJ, Hahn ME (2001) Acquired resistance to Ah receptor agonists in a population of Atlantic killifish (*Fundulus heteroclitus*) inhabiting a marine superfund site: In vivo and in vitro studies on the inducibility of xenobiotic metabolizing enzymes. *Toxicological Sciences* **60**, 77-91.

Black DE, Gutjahr-Gobell R, Pruell RJ, *et al.* (1998) Reproduction and polychlorinated biphenyls in *Fundulus heteroclitus* (Linnaeus) from New Bedford Harbor, Massachusetts, USA. *Environmental Toxicology and Chemistry* **17**, 1405-1414.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The Hitchhiking Effect on the Site Frequency-Spectrum of DNA Polymorphisms. *Genetics* **140**, 783-796.

Burnett KG, Bain LJ, Baldwin WS, *et al.* (2007) *Fundulus* as the premier teleost model in environmental biology: Opportunities for new insights using genomics. *Comparative Biochemistry and Physiology D-Genomics & Proteomics* **2**, 257-286.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *Plos Genetics* **2**, 379-384.

Charlesworth D, Charlesworth B, Morgan MT (1995) The Pattern of Neutral Molecular Variation under the Background Selection Model. *Genetics* **141**, 1619-1632.

Chen D, Zhang X, Mai B, *et al.* (2009) Polychlorinated biphenyls and organochlorine pesticides in various bird species from northern China. *Environmental Pollution* **157**, 2023-2029.

Collier TK, Anulacion BF, Bill BD (1998) Hepatic CYP1A in winter flounder (*Pleuronectes americanus*) along the northeast coast: Results from the National Benthic Surveillance Project. *Marine Pollution Bulletin* **37**, 86-91.

Crawford DL, Segal JA, Barnett JL (1999) Evolutionary analysis of TATA-less proximal promoter function. *Molecular Biology and Evolution* **16**, 194-207.

de Meaux J, Goebel U, Pop A, Mitchell-Olds T (2005) Allele-specific assay reveals functional variation in the chalcone synthase promoter of *Arabidiopsis thaliana* that is compatible with neutral evolution. *Plant Cell* **17**, 676-690.

Denison MS, Nagy SR (2003) Activation of the aryl hydrocarbon receptor by structurally diverse exogenous and endogenous chemicals. *Annual Review of Pharmacology and Toxicology* **43**, 309-334.

Eanes WF, Kirchner M, Yoon J (1993) Evidence for Adaptive Evolution of the G6pd Gene in the *Drosophila melanogaster* and *Drosophila simulans* Lineages. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 7475-7479.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Fiedler H, Cooper K, Bergek S, *et al.* (1998) PCDD, PCDF, and PCB in farm-raised catfish from southeast United States - Concentrations, sources, and CYP1A induction. *Chemosphere* **37**, 1645-1656.

Fu YX, Li WH (1993) Statistical Tests of Neutrality of Mutations. *Genetics* **133**, 693-709.

Garrick RA, Woodin BR, Wilson JY, Middlebrooks BL, Stegeman JJ (2006) Cytochrome P4501A is induced in endothelial cell lines from the kidney and lung of the bottlenose dolphin, *Tursiops truncatus*. *Aquatic Toxicology* **76**, 295-305.

Gunawickrama S, Aarsaether N, Orbea A, Cajaraville MP, Goksoyr A (2008) PCB77 (3,3 ',4,4 '-tetrachlorobiphenyl) co-exposure prolongs CYP1A induction, and sustains oxidative stress in B(a)P-exposed turbot, *Scophthalmus maximus*, in a long-term study. *Aquatic Toxicology* **89**, 65-74.

Hahn ME (1998) Mechanisms of Innate and Acquired Resistance to Dioxin-like Compounds. *Reviews in Toxicology. Series B: Environmental Toxicology.* **2**, 395-443.

Hahn ME, Karchner SI, Franks DG (2002) Aryl hydrocarbon receptor polymorphisms and dioxin resistance in Atlantic killifish (*Fundulus heteroclitus*). *Marine Environmental Research* **54**, 409.

Hahn ME, Karchner SI, Franks DG, *et al.* (2005) Mechanism of PCB- and dioxin-resistance in fish in the Hudson River esturary: role of receptor polymorphisms. (ed. H.R.F.G.A.).

Hahn ME, Karchner SI, Franks DG, Merson RR (2004) Aryl hydrocarbon receptor polymorphisms and dioxin resistance in Atlantic killifish (*Fundulus heteroclitus*). *Pharmacogenetics* **14**, 131-143.

Hahn ME, Lamb TM, Schultz ME, Smolowitz RM, Stegeman JJ (1993) Cytochrome-P4501a Induction and Inhibition by 3,3',4,4'-Tetrachlorobiphenyl in an Ah Receptor-Containing Fish Hepatoma-Cell Line (PLHC-1). *Aquatic Toxicology* **26**, 185-208.

Hasenkamp S, Telgmann R, Staessen JA, *et al.* (2008) Characterization and functional analyses of the human G protein-coupled receptor kinase 4 gene promoter. *Hypertension* **52**, 737-746.

Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877.

Huggett RJ, Neff JM, Stegeman JJ, *et al.* (2006) Biomarkers of PAH exposure in an intertidal fish species from Prince William Sound, Alaska: 2004-2005. *Environmental Science & Technology* **40**, 6513-6517.

Jukes TH, Cantor CR (1969) *Evolution of protein molecules* Academic Press, New York.

Karchner SI, Franks DG, Powell WH, Hahn ME (2002) Regulatory interactions among three members of the vertebrate aryl hydrocarbon receptor family: AHR repressor, AHR1, and AHR2. *Journal of Biological Chemistry* **277**, 6949-6959.

Karchner SI, Powell WH, Hahn ME (1999) Identification and functional characterization of two highly divergent aryl hydrocarbon receptors (AHR1 and AHR2) in the teleost *Fundulus heteroclitus* - Evidence for a novel subfamily of ligand-binding basic helix loop helix-Per-ARNT-Sim (bHLH-PAS) factors. *Journal of Biological Chemistry* **274**, 33814-33824.

Kreitman M (1983) Nucleotide Polymorphism at the Alcohol-Dehydrogenase Locus of *Drosophila melanogaster*. *Nature* **304**, 412-417.

Lake JL, Mckinney R, Lake CA, Osterman FA, Heltshe J (1995) Comparisons of Patterns of Polychlorinated Biphenyl Congeners in Water, Sediment, and Indigenous Organisms from New Bedford Harbor, Massachusetts. *Archives of Environmental Contamination and Toxicology* **29**, 207-220.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452.

Liu DH, Krajniak K, Chun D, *et al.* (2001) VIP gene transcription is regulated by far upstream enhancer and repressor elements. *Biochemical and Biophysical Research Communications* **284**, 211-218.

Lubet RA, Nims RW, Beebe LE, *et al.* (1992) Induction of hepatic CYP1A activity as a biomarker for environmental exposure to Aroclor® 1254 in feral rodents. *Archives of Environmental Contamination and Toxicology* **22**, 339-344.

Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564-567.

Ludwig MZ, Kreitman M (1995) Evolutionary Dynamics of the Enhancer Region of Even-Skipped in *Drosophila*. *Molecular Biology and Evolution* **12**, 1002-1011.

M. Nilsen B, Berg K, Goksor A (1998) Induction of Cytochrome P4501A (CYP1A) in Fish: A Biomarker for Environmental Pollution, pp. 423-438.

Matys V, Kel-Margoulis OV, Fricke E, *et al.* (2006) TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108-D110.

McDonald JH, Kreitman M (1991) Adaptive Protein Evolution at the Adh Locus in Drosophila. *Nature* **351**, 652-654.

Murre C, Mccaw PS, Vaessin H, *et al.* (1989) Interactions between Heterologous Helix-Loop-Helix Proteins Generate Complexes That Bind Specifically to a Common DNA-Sequence. *Cell* **58**, 537-544.

Nacci D, Coiro L, Champlin D, *et al.* (1999) Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. *Marine Biology* **134**, 9-17.

Nacci D, Gleason T, Gutjahr-Gobekkm R, Huber M, Munns WRJ (2002a) *Effects of environmental stressors on wildlife populations* CRC Press/Lewis Publishers, Washington, D.C.

Nacci D, Jayaraman S, Specker J (2001) Stored retinoids in populations of the estuarine fish *Fundulus heteroclitus* indigenous to PCB-contaminated and reference sites. *Archives of Environmental Contamination and Toxicology* **40**, 511-518.

Nacci DE, Kohan M, Pelletier M, George E (2002b) Effects of benzo[a]pyrene exposure on a fish population resistant to the toxic effects of dioxin-like compounds. *Aquatic Toxicology* **57**, 203-215.

Oleksiak MF (2010) Genomic approaches with natural fish populations. *Journal of Fish Biology* **76**, 1067-1093.

Oleksiak MF, Roach JL, Crawford DL (2005) Natural variation in cardiac metabolism and gene expression in Fundulus heteroclitus. *Nature Genetics* **37**, 67-72.

Olson SA (1994) MacVector: Sequence Comparisons Using a Matrix Method, pp. 215-225.

Patil N, Berno AJ, Hinds DA, *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719-1723.

Poland A, Knutson JC (1982) 2,3,7,8-Tetrachlorodibenzo-Para-Dioxin and Related Halogenated Aromatic-Hydrocarbons - Examination of the Mechanism of Toxicity. *Annual Review of Pharmacology and Toxicology* **22**, 517-554.

Powell WH, Bright R, Bello SM, Hahn ME (2000) Developmental and tissue-specific expression of AHR1, AHR2, and ARNT2 in dioxin-sensitive and -resistant populations of the marine fish *Fundulus heteroclitus*. *Toxicological Sciences* **57**, 229-239.

Powell WH, Karchner SI, Bright R, Hahn ME (1999) Functional diversity of vertebrate ARNT proteins: Identification of ARNT2 as the predominant form of ARNT in the marine teleost, *Fundulus heteroelitus*. *Archives of Biochemistry and Biophysics* **361**, 156-163.

Powell WH, Morrison HG, Weil EJ, *et al.* (2004) Cloning and analysis of the CYP1A promoter from the atlantic killifish (*Fundulus heteroclitus*). *Marine Environmental Research* **58**, 119-124.

Pruell RJ, Norwood CB, Bowen RD, *et al.* (1990) Geochemical Study of Sediment Contamination in New Bedford Harbor, Massachusetts. *Marine Environmental Research* **29**, 77-101.

Reyes H, Reiszporszasz S, Hankinson O (1992) Identification of the Ah Receptor Nuclear Translocator Protein (Arnt) as a Component of the DNA-Binding Form of the Ah Receptor. *Science* **256**, 1193-1195.

Ryan JA, Hightower LE (1994) Evaluation of Heavy-Metal Ion Toxicity in Fish Cells Using a Combined Stress Protein and Cytotoxicity Assay. *Environmental Toxicology and Chemistry* **13**, 1231-1240.

Schulte PM, Segal JA, Crawford DL, Powers DA (1995) Rapid in-Vivo Footprinting Method for the Detection of DNA-Protein Interactions in Isolated-Nuclei. *Molecular Marine Biology and Biotechnology* **4**, 200-205.

Shine JP, Ika RV, Ford TE (1995) Multivariate Statistical Examination of Spatial and Temporal Patterns of Heavy-Metal Contamination in New-Bedford Harbor Marine-Sediments. *Environmental Science & Technology* **29**, 1781-1788.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. *Genetics* **141**, 413-429.

Skinner MA, Courtenay SC, Parker WR, Curry RA (2005) Site fidelity of mummichogs (*Fundulus heteroclitus*) in an Atlantic Canadian estuary. *Water Quality Research Journal of Canada* **40**, 288-298.

Swofford DL (2003) Phylogenetic Analysis Using Parsimony (*and Other Methods). . Sinauer Associates, Sunderland, Massachusetts.

Tajima F (1989) Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585-595.

Timme-Laragy AR, Meyer JN, Waterland RA, Di Giulio RT (2005) Analysis of CpG methylation in the killifish CYP1A promoter. *Comparative Biochemistry and Physiology C-Toxicology & Pharmacology* **141**, 406-411.

Wall JD (1999) Recombination and the power of statistical tests of neutrality. *Genetical Research* **74**, 65-79.

Weaver G (1983) PCB contamination in and around New Bedford, Mass. *Environmental Science & Technology* **18**, 22A-27A.

Whitlock JP, Okino ST, Dong LQ, *et al.* (1996) Cytochromes P450 .5. Induction of cytochrome P4501A1: A model for analyzing mammalian gene transcription. *Faseb Journal* **10**, 809-818.

Williams LM, Oleksiak MF (2010) Ecologically and evolutionarily important SNPs identified in natural populations. *Molecular Biology and Evolution* **in review**.

Williams TD, Lee JS, Sheader DL, Chipman JK (2000) The cytochrome P450 1A gene (CYP1A) from European flounder (*Platichthys flesus*), analysis of regulatory regions and development of a dual luciferase reporter gene system. *Marine Environmental Research* **50**, 1-6.

Wilson JY, Cooke SR, Moore MJ, *et al.* (2005) Systemic effects of arctic pollutants in Beluga whales indicated by CYP1A1 expression. *Environmental Health Perspectives* **113**, 1594-1599.

Wirgin I, Waldman JR (2004) Resistance to contaminants in North American fish populations. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **552**, 73-100.

Table 1. Measures of population divergence between *F. heteroclitus* populations. *K*xy = πD (average number of pairwise nucleotide differences), G_{ST} is Nei's coefficient of gene variation and defined as 1-(H_S/H_T) where H_S and H_T are the mean heterozygosity within populations and in the entire species, respectively, and F_{ST} is Wright's inbreeding coefficient and is defined as (H_T-H_S)/H_T.

Population 1	**Population 2**	**Kxy**	**G_{ST}**	**F_{ST}**
Sandwich	New Bedford	42.11	0.028	0.054
Sandwich	Point Judith	41.14	0.040	0.410
New Bedford	Point Judith	46.09	0.037	0.405

Table 2. Neutrality test for the pattern of sequence variation (Mcdonald, Kreitman, 1991). Probability that this pattern of sequence variation is due to random (H_O) or selective (H_1) evolutionary processes. A two-sided Fisher's exact test failed to reject the null hypothesis (p=0.4906) for table **A** which tests between species. A two-sided Fisher's exact test failed to reject the null hypothesis (p=1.0) for table **B** which tests between polluted and reference populations of *F. heteroclitus*.

A

	Fixed	Polymorphic
Functional	0	13
Non-functional	17	142

B

	Fixed	Polymorphic
Functional	0	13
Non-functional	0	142

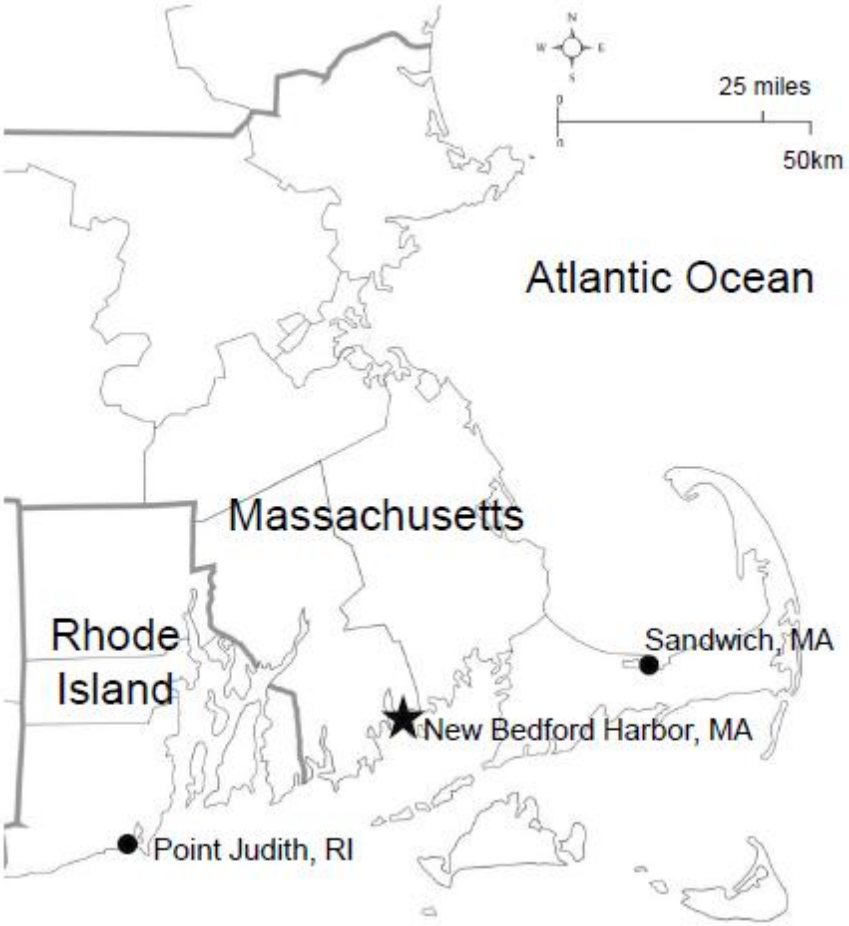
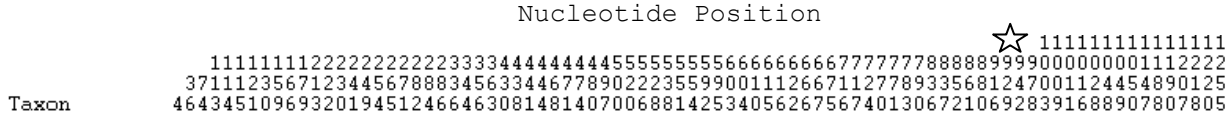


Figure 1. *F. heteroclitus* collection sites. Superfund site (New Bedford Harbor, MA) is denoted by a star, and reference sites north and south denoted by a circle.

Figure 2. 221 parsimony informative sites among and between *F. heteroclitus* populations and between *F. heteroclitus* and *F. grandis* in the sequenced portion of the CYP1A promoter, exon and intron 1. Within the promoter region (upstream of basepair 1630), there are 157 parsimony informative sites. There are no fixed differences between *F. heteroclitus* populations. 20 fixed differences between *F. heteroclitus* and *F. grandis* are marked by a star. SNPs (929 and 1892bp) found to be under selection in Williams and Oleksiak (2010) are starred over the nucleotide number.

Nucleotide Position




```
Sandwich 6    CGTACCTTGAGGATTCTCTTGGACTCATACCCAGAAAATGGACATTTGCGGTCGCGGGTCTACCGCCATCTCTAAATATTTAAC
Sandwich 7    CGTGCCTTGAGGGTTCTCTTGGACTCTTACCCAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTAAC
Sandwich 8    CGTGCCTTGAGGGTTCTCTTGGACTCTTACCCGGGCTATAGACCTTTGTGAGAGCGGGTCTGACGCGATCTCTTAATATTTAAC
Sandwich 11   CGTNCCTTGAGGGTTCCCTTGGACTCNTACCCAGAAAATAAACATTTGTGAGAACGGGTCTACCAAGAGTTCGTAATATTTAAC
Sandwich 12   CGCGCCTGGAGGGTTCTCTTGGACTCTTACCCAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTAAC
Sandwich 14   CGTNCCTTATGGGTTCTCTCGGACTCTTACCCAGAAAATAGACATTTGTGGTCGCGGGTCTGCCGCGATTTCTTAATATTTAAC
Sandwich 27   CGCCCCTGGAGGGTACTCTTGGACTCNTACCCAGAAAATAAACATTTGTGGTCGCGGGTCTACCGCCATCTCTTAATATTTAAC
Sandwich 32   CGTCCGTCGAGGGTTCTCTTGGACTCATACCCAGAAAATAAACATCTGTGGTCGCGGGCCTGCCGCGATTTCTTAATATTTAAC
NBH 3         CGTGCCTTGAGGGTTCTCTTGGACTCTTACCCAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCCATCTCTTAATATTTAAC
NBH 4         CGTGCCTTGAGGGTTCTCTTGGACTCTTACCCAGAAAATAAACATTTGTGGTCCGGGGTCTGCCGCNATNTCTTAATATTTAAC
NBH 6         CGTGCCTTGAGGGTTCTCTTGGACTCTTACCCAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCCATTTCTTAATATTTAAC
NBH 7         CGTGCCTTGAGGGTACTCTTGGACTCGTACCCAGAAAATAAACATTCGTGGTCGCGGGTCTGCCGCGATTTCTTAATATTTAAC
NBH 9         CGTNCCTTGAGGGTTCTCTTGGACTCNTNCCCAGAAAATAAACATTTGTGGTCGCGGATCTGCCGCGATTTCTTAATATTTAAC
NBH 12        CGNNCCTNGAGGGTTCTCTTGGACTCATGCAGGGGCTATAGACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTCCT
NBH 13        CGTGCCTTGAGGGTTCTCTGGGACTCTTACCCAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCCATTTCTTAATATTTAAC
NBH 307       CTCGATGGGAATGCTTTACTTAATCCTTAATCATAATAGAAACATTTATAGTCGCGGGTCTGCCGCCATCTCTTGCAATTTAAC
PJ 1          CGTGCCTTGAGGGTTCTCTTGGACTTATACACAGAAAATAAACATTTGTGGTCGCGAGTCTGCCGCGATCTCTTAATATTTAAC
PJ 2          CGTGCCTTGAGGGTTCTCTTGGACTCATACACAGAAAATAAACATTTGTGGTCGCGAGTCTGCCGCGATCTCTTAATATTTAAC
PJ 3          CGTGCCTTGAGGGTTCTCTTGGGCTCATACACAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTAAC
PJ 4          CGTGCCTTGAGGGTTCTCTTGGGCTCATACACAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTAAC
PJ 5          CGTGCCTTGAGGGTTCTCTTGGGCTCATACACAGAAAATAAACATTTGTGGTCGCGAGTCTGCCGCGATCTCTTAATATTTAAC
PJ 6          CGTGCCTTGAGGGTTCTCTTGGACTCATACACAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTAAC
PJ 19         CGTGCCTTGAGGGTTCTCTTGGACTCATACACAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTAAC
PJ 26         CGTGCCTTGAGGGTTCTCTTGGGCTCATACACAGAAAATAAACATTTGTGGTCGCGGGTCTGCCGCGATCTCTTAATATTTAAC
F.grandis 13  TGCGCTTGGAAGGTTCTATTGGACTTTCAATCAGGATGGAAGGACCTATGATCGCAGGTTCACGGCCGTCCACTAAATAAAAAC
              *                         *         *   ** *          *   ** *   *  **     ****
```

Nucleotide Position

```
              1111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
              2222222222222222333333333333333333344444444444444444444455555555555555555566666666666
              5566677788999999000000111333456668900122333444556788899001112223334445568800112223444
Taxon         6926867805013689036790124791002610676053576790199256274825923403801717317594646732 38
```

```
Sandwich 6    ACTCAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
Sandwich 7    ACTCAAAAGCATAATTGTCCTATCACGCCTATCAACTCAAAACACAATGCAATGATAGCCGAGGAGGCTGAAGACCCAAATAAA
Sandwich 8    ACTTAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
Sandwich 11   ACTCAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
Sandwich 12   ACTCAAAAGCATAATTGTCCTATCACGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGGGCGAA
Sandwich 14   ACTCAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
Sandwich 27   ACTCAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
Sandwich 32   ACTCAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
NBH 3         ACTCAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
NBH 4         ACTCAAAAGCATAATTGTCCTATCNCGCCTACGAACTCAAAACACAATNNAANTGTNNNCANGGNTGANGAGGTTTGGAGTGAA
NBH 6         ACTCAAAAGCATAATTGTCCTATCATAATCCGGTTTCGCCCCCTTTTTCTTCCATGGAAAGAGAAGTGACGCGATTTAGAGTGCA
NBH 7         ACTCAAAAGCATAATTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
NBH 9         ACTCAAAAGCATAATTGTCCTATCNCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
NBH 12        GTCCGGTCATGGGCCCCCGTGGATCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAG
NBH 13        ACTCAAAAGCATAACTGTCCTATCATAATCCGGTTTCGCCCCCTTTTTCTTCCATGGAAAGAGAAGTGACGCGATTTAGAGTGCA
NBH 307       ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
PJ 1          ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGRTGATGAGGTTTGGAGTGAA
PJ 2          ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGGTGATGAGGTTTGGAGTGAA
PJ 3          ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGGTGATGAGGTTTGGAGTGAA
PJ 4          ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGGTGATGAGGTTTGGAGTGAA
PJ 5          ACTCAAAAGCATAACTGTCCTATCNCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGATAACCAGGTTTGGAGTGAA
PJ 6          ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGGTGATGAGGTTTGGAGTGAA
PJ 19         ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGATGACCAGGTTTGGAGTGAA
PJ 26         ACTCAAAAGCATAACTGTCCTATCCCGCCTAGGAACTCAAAACACAATTTAAAGGTAAACAAGGGTGATGAGGTTTGGAGTGAA
F.grandis 13  ACTCAAAAGCATAACTGTTCTATCACGCCTAGGAACTCAAAACACAATTTAAATGTGAACAGGGGTGATGAGGTTTGGAGTGAA
```

Nucleotide Position

```
                1111111111111111111111111111111111111111111
                6666777777777777888888888888899999999999999
                6889001147789999011344556899900112446889999
Taxon           2693140173570257048807019112308791574481 49
-------------------------------------------------------------

Sandwich 6      CGTCGTTTAAAGTGTGCGAGTAATAAACCTCCAACAGTTGAT
Sandwich 7      CCCGGGCCAAGGTGTGCGAGTAATAAATCTTCAAGTGTTGAT
Sandwich 8      CGTTGTTCAAAGTGTGCGANTAATAAACCTTCAAATGTTGAT
Sandwich 11     CGTTGTTCAAGGTGTGCGAGTAATAAACCTTCAAGTGTTGAT
Sandwich 12     CGTTGTTCTTAGTACACTCATACTAAATCTTCAGGTGNTAAT
Sandwich 14     CGTTGTTCAARGTGTGCGAGTAATAAACCTTCAAGTGTTGAT
Sandwich 27     CGTTGTTCAAAGTGTGCTCGTACTAAACCTTCAGGTGTTGAT
Sandwich 32     CGTTGTTCAAAGTGTGCGAGTAATAAACCTTCAAGTGTTGAT
NBH 3           CGTTGTTCTAAGTATACTCATACTAAACCTTCAGGTGTTGAT
NBH 4           CGTTGTTCAAAGTGTGCTCGTTATAAATCGTCGAGAGCCGGC
NBH 6           CGTTCTTCTAAGTACACTCAGACCATTCCTTAAGRTCTTGAT
NBH 7           CGTTGTTCAAAGTGTGTTCATACTAAATCTTCAGGTGTTGAT
NBH 9           CGTTGTTCAAAGTGTGCTCATACTAAACCTTCAGGTGTTGAT
NBH 12          CGTTGTTCTAAGAATACGANTAATAAATCTTCAGGTGTTGAT
NBH 13          CGTTCTTCTAAGTACACTCAGACCATTCCTTAAGGCCTTGAT
NBH 307         CGTTGTTCTAAGTATACTCATACTAAATCTTCAGGTGTTGAT
PJ 1            CGTTGTTCAAAGTGTGCGMATACTAAACTTTCAGGTGTTGAT
PJ 2            CGTTGTTCAAAGTGTGCGCATACTAAACTTTCAGGTGTTGAT
PJ 3            CGTTGTTCAAAGTGTGCGCATACTAAACTTTCAGGTGTTGAT
PJ 4            CGTTGTTCAAAGTGTGCGCATACTAAATCTTCAGGTGTTGAT
PJ 5            CGTTGTTCAAAGTGTGCTCATACTAAACTTTCAGGTGTTGAT
PJ 6            CGTTGTTCAAAGTGTGCGCATACTAAACCTTCAGGTGTTGAT
PJ 19           CGTTGTTCAAAGTGTGCGCATACTAAATTTTCAGGTGTTGAT
PJ 26           CGTTGTTCAAAGTGTGCGCATACTAAACTTTCAGGTGTTGAT
F.grandis 13    TGTTGTTCTAATTATACTCATACTGAACTTTCAGGTTTTGAT
                *          *          *
```

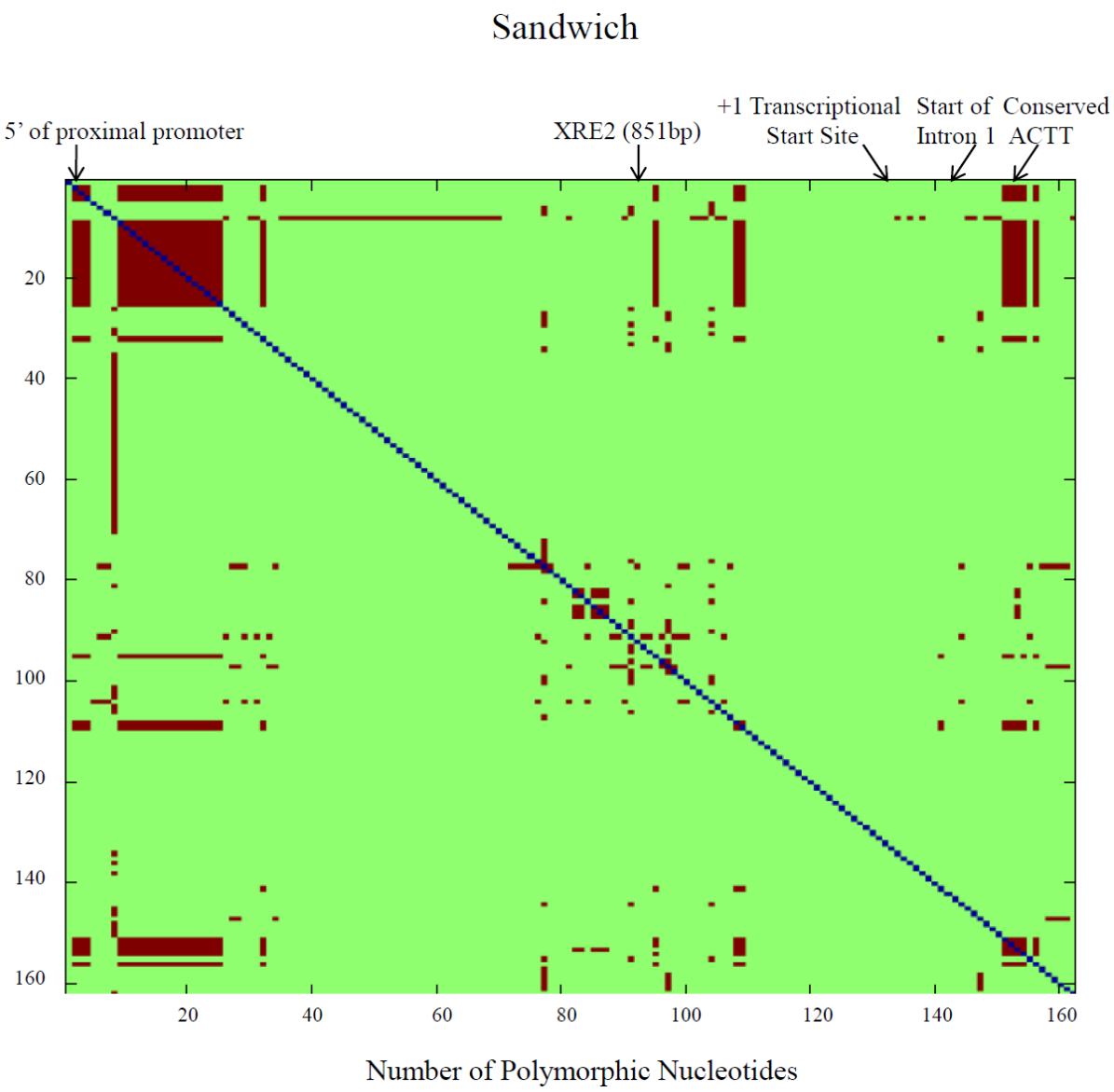


Figure 3. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for Sandwich. Green indicates no significant LD, red indicates significant LD ($p \leq 0.05$), and the blue line indicates the diagonal.

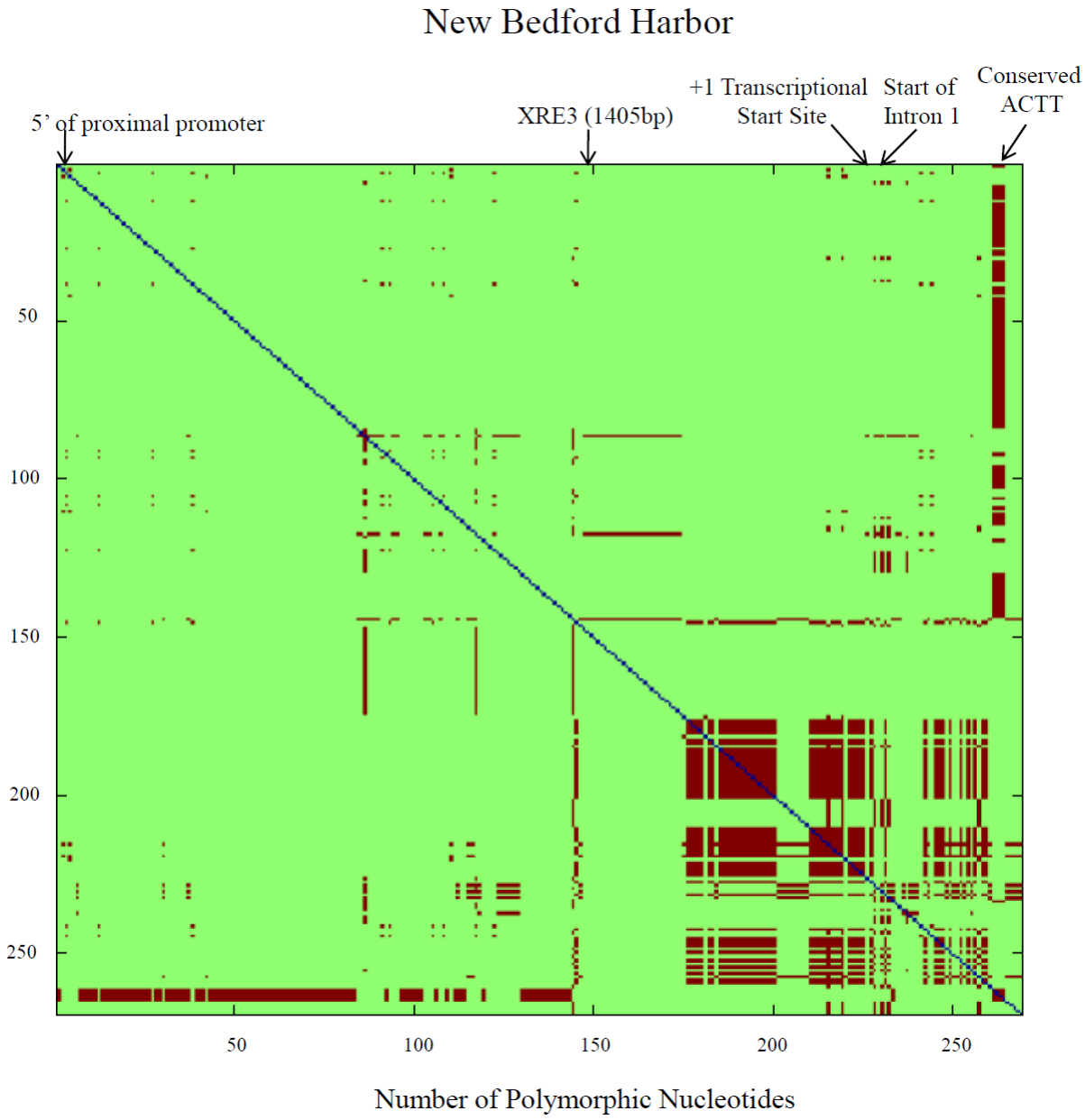


Figure 4. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for New Bedford Harbor. Green indicates no significant LD, red indicates significant LD ($p \leq 0.05$), and the blue line indicates the diagonal.

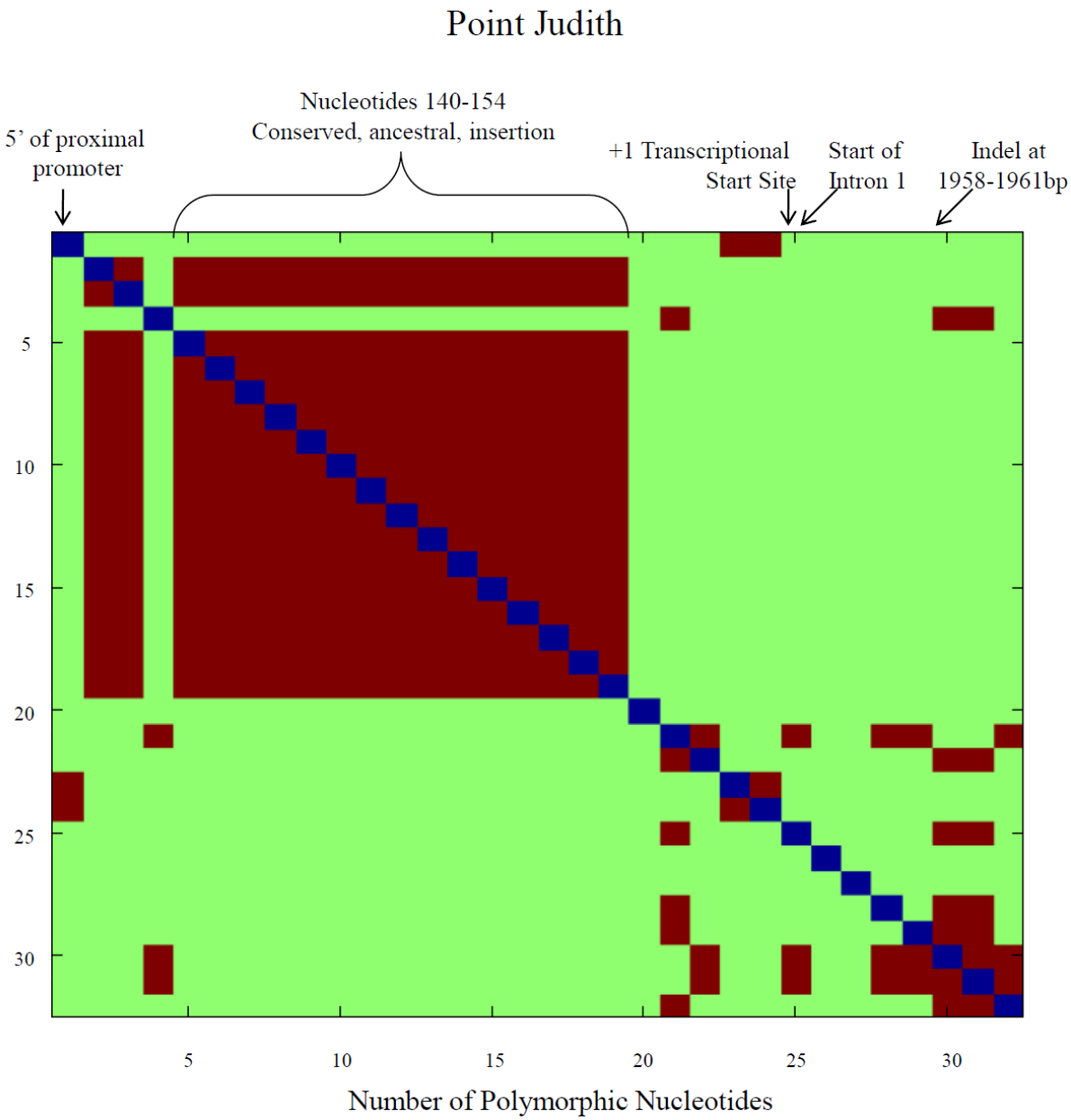


Figure 5. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for Point Judith. Green indicates no significant LD, red indicates significant LD ($p \leq 0.05$), and the blue line indicates the diagonal.

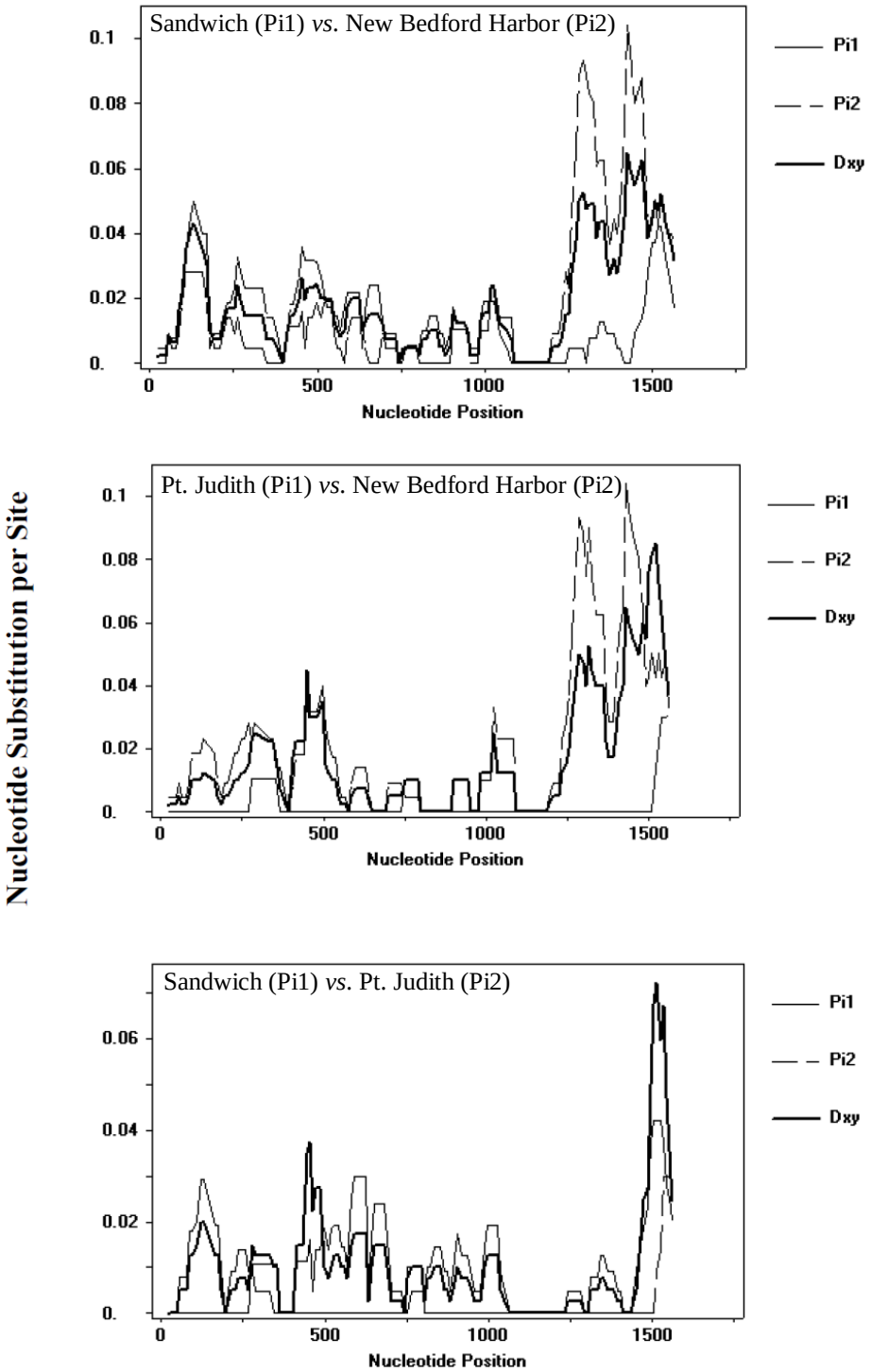


Figure 6. Sliding window comparisons of average nucleotide substitutions per site within *F. heteroclitus* populations (Pi) and between populations (Dxy). Plots for the sliding window use 50bp-wide window and 10-bp step size.

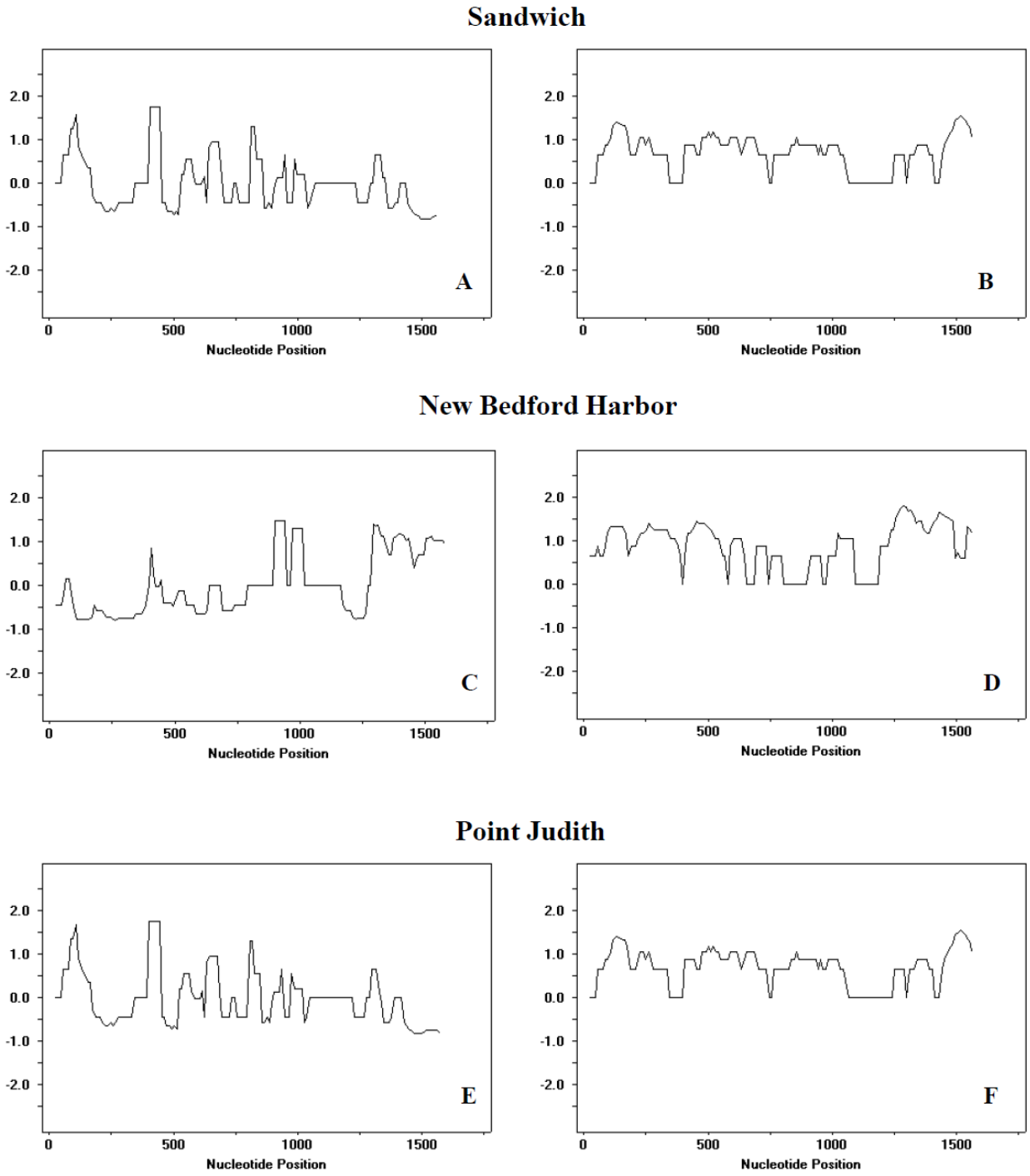
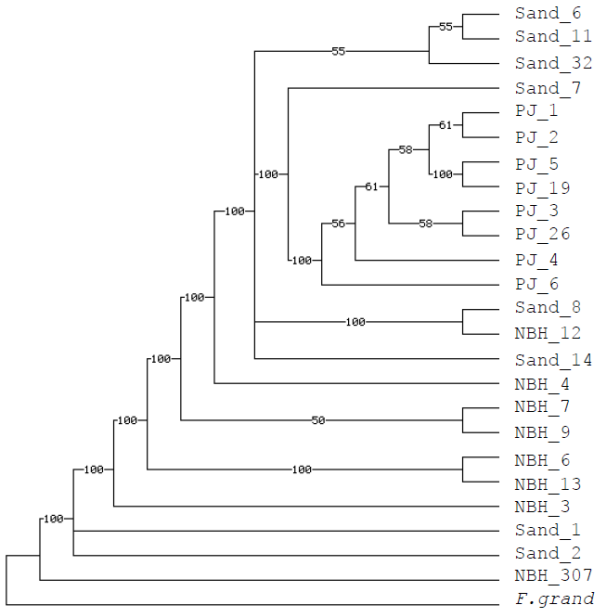


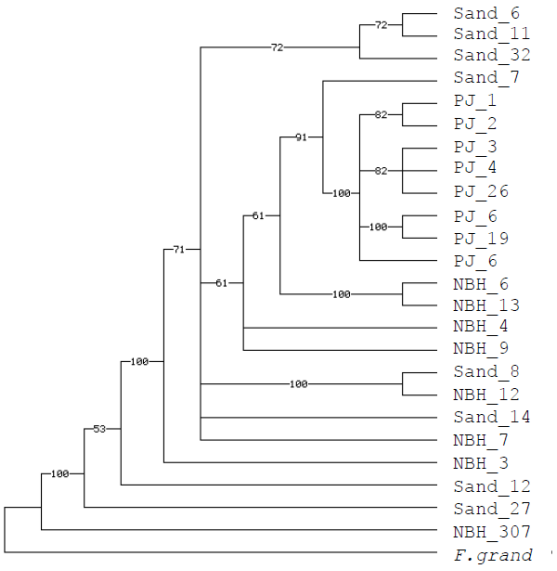
Figure 7. Tajima's D calculations (panels A,C,E) and Fu and Li's D calculations (panels B, D, and F) using a sliding window with a 50bp-wide window and a 10-bp step size for each *F. heteroclitus* population.

Figure 8. Phylogenetic analyses of the CYP1A proximal promoter using PAUP*4.0 (Swofford, 2003). A. Phylogenetic tree using all nucleotides of the proximal promoter. B. Phylogenetic tree using functional regions of the proximal promoter. C. Phylogenetic tree using non-functional regions of the proximal promoter. Bootstrap values (N=500) are listed. Probabilities are maximum parsimony values.

A. Full length proximal promoter



B. Non-Functional regions of the proximal promoter



C. Functional regions of the proximal promoter

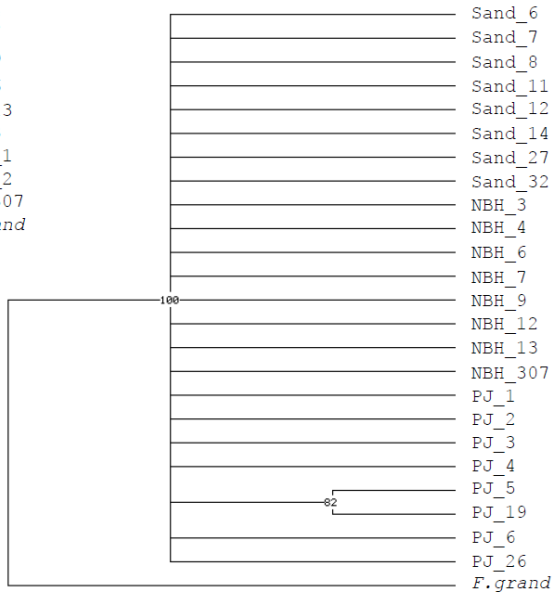
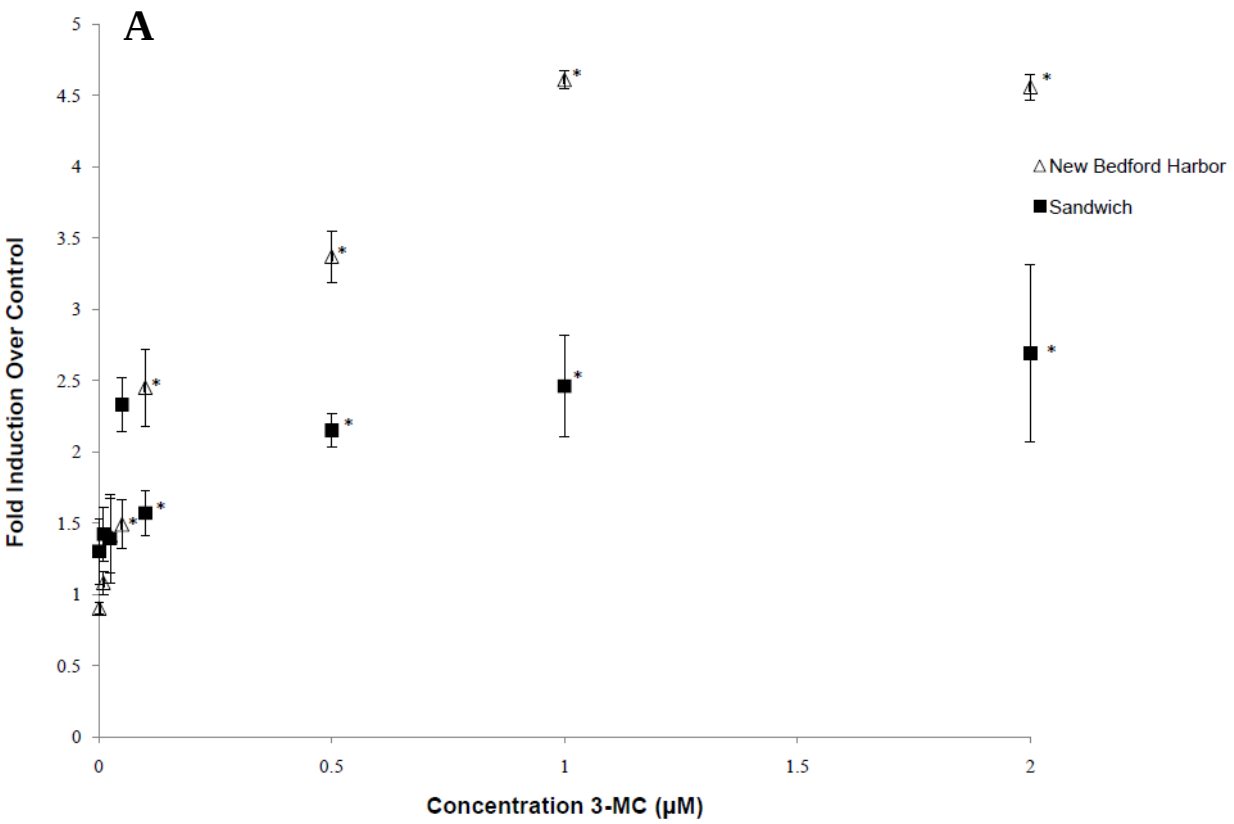
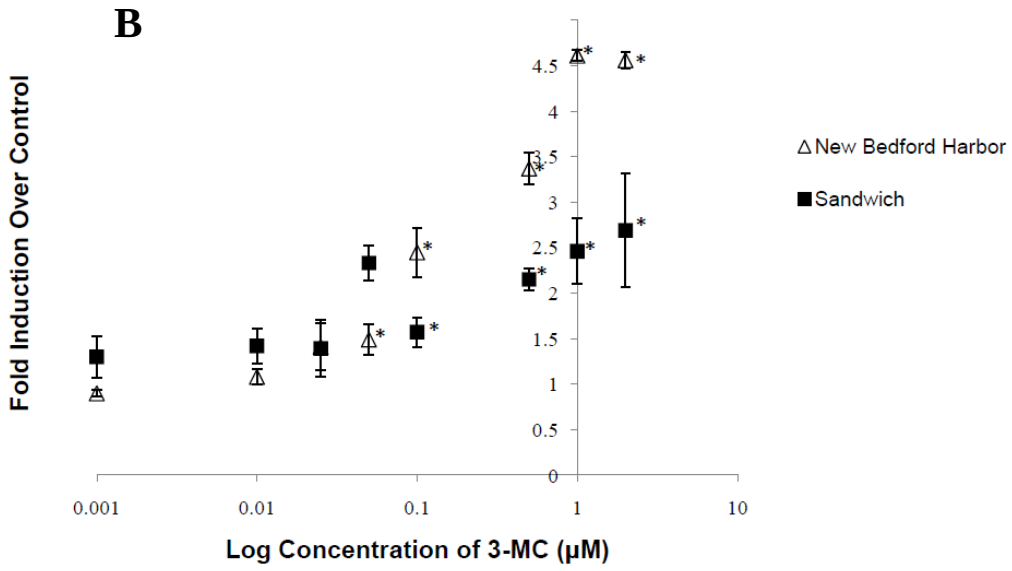


Figure 9. Induction of luciferase activity from *F. heteroclitus* CYP1A reporter gene constructs from a New Bedford and Sandwich individual by 3-MC in PLHC-1 cells plotted on a non-logarithmic (A) and logarithmic scale (B). A star indicates a significant fold induction over control ($p<0.05$).


A
Fold Induction Over Control
Concentration 3-MC (μM)
△New Bedford Harbor
■Sandwich
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5


B
Fold Induction Over Control
Log Concentration of 3-MC (μM)
△New Bedford Harbor
■Sandwich
0.001
0.01
0.1
1
10

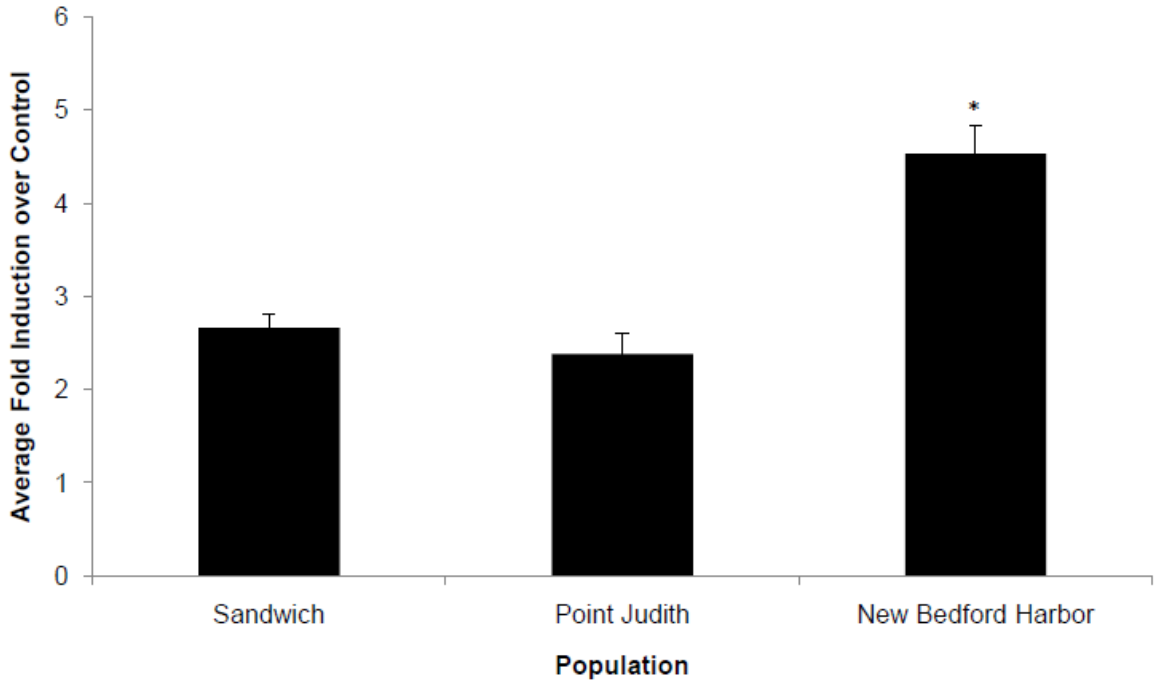


Figure 10. Average fold induction of the CYP1A promoter over control for three populations (N=4) at a 1 μM dose of 3-MC. A one-way ANOVA found a significant difference in the average fold induction between populations ($p < 0.0001$) and a Tukey's HSD post-hoc test determined that New Bedford Harbor was significantly different from Sandwich ($p < 0.0001$) and Point Judith ($p < 0.0001$), but Sandwich and Point Judith were not significantly different from each other ($p = 0.258$). Significant induction in the New Bedford Harbor population over the reference populations is denoted by a star.

Supplementary Table 1. Predicted transcription factor binding sites along the CYP1A promoter by AliBaba2.1 by constructing matrices from TRANSFAC 4.0. XRE and GRE consensus sequences were ascribed and tested for functionality as described in Powell et al. (2004).

Nucleotide Position	Transcription Factor	Consensus Sequence
121-130	GATA-1	TTTTTTTTAT/GC
230-238	C/EBPα	TAAGCAATC
399-404	GRE 1/2 Site	AGGACA
403-413	Oct-1	CATGTAAACA
729-740	Sp-1	CTCCTCCCCTCC
799-805	XRE 1	CACGCAA
848-854	XRE 2	CACGCAA
1027-1036	MEB-1 and GLO	TTATTTTTAA
1101-1111	CREB AND C/EBPα and AP1	CCTGCGTCAGC
1212-1221	C/EBPα	TAAAATAAAT
1324-1333	OCT2.1	GCTGATTTGC
1391-1401	GR	ACACTTTGTA
1405-1410	XRE 3	CACGCGAA
1528-1539	Sp-1	GGAGGAGGGGAGA
1538-1546	CREB	GATGATGTC
1596-1601	TATA Box	TATAAA

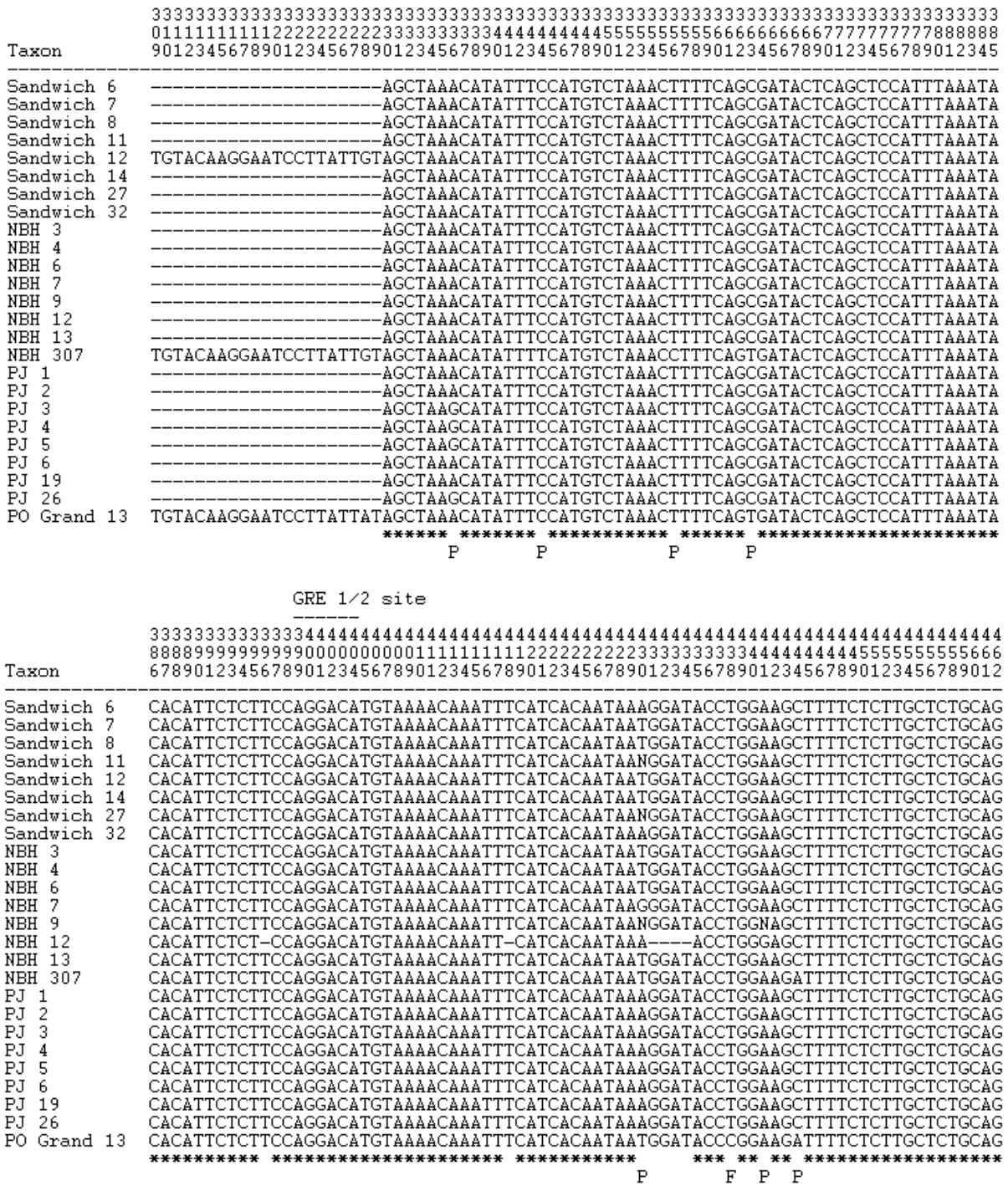
Supplementary Figure 1. Full length proximal promoter and exon 1 and intron 1. Star indicates conservation in the nucleotide across all individuals, P indicates a parsimony informative site, and F indicates a fixed difference between *F. heteroclitus* and *F. grandis*. GRE and XRE binding sites are noted, as well as the TATA box, start of transcription, and intron and exonic boundaries. Big stars on nucleotides 929 and 1892 indicate SNPs under selection as described in Williams and Oleksiak (2010).

```
                    1111111111222222222233333333334444444444555555555566666666667777777
Taxon          12345678901234567890123456789012345678901234567890123456789012345678901234567
-------------------------------------------------------------------------------------------------
Sandwich 6     GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
Sandwich 7     GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
Sandwich 8     GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
Sandwich 11    GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
Sandwich 12    GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAACTTT
Sandwich 14    GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
Sandwich 27    GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAACTTT
Sandwich 32    GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
NBH 3          GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
NBH 4          GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTNGTTCCACACTTTTTCCTTCCACTAAATTTT
NBH 6          GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
NBH 7          GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
NBH 9          GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
NBH 12         GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAANTTT
NBH 13         GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
NBH 307        GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTATTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAACTTT
PJ 1           GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
PJ 2           GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
PJ 3           GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
PJ 4           GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
PJ 5           GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTT-GTTCCACACTTTTTCCTTCCACTAAATTTT
PJ 6           GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
PJ 19          GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
PJ 26          GTGCGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAATTTT
PO Grand 13    GTGTGTTTCTCTGGTGGTCAAAACGTTGATTCTTAGTGCACCTTTTTGTTCCACACTTTTTCCTTCCACTAAACTTT
               *** ******************************** ********** ************************** ***
                  F                                P                                    P

                                                  1111111111111111111111111111111111111111111111111111111111
               7788888888889999999999000000000011111111112222222222333333333344444444445555
Taxon          8901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234
-------------------------------------------------------------------------------------------------
Sandwich 6     CCACCATCATATTTTGTTACTGCACCAT--GAAAAACCGACTCTTTTTT--ATCAA----------------CGACCT
Sandwich 7     CCACCATCATATTTTGTTACTGCACCAT--GAAAAGCCGACTCTTTTTTTTATCAA----------------CGACCT
Sandwich 8     CCACCATCATATTTTGTTACTGCACCAT--GAAAAGCCGACTCTTTTTTTTATCAA----------------CGACCT
Sandwich 11    CCACCATCATATTTTGTTACTGCACCAT--GAAAANCCGACTCTTTTTTTTATCAA----------------CGACCT
Sandwich 12    CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCCGACTCTTTTTTTTTAGCAACGACCTCTTGTGGTTCATCCT
Sandwich 14    CCACCATCATATTTTGTTACTGCACCAT--GAAAANCCGACTCTTTTTTTTATCAA----------------CGACCT
Sandwich 27    CCACCATCATATTTTGTTACTGCACCATGTGAAAAACCGACTCTTTTTTTTAGCAACGACCTCTTGTGGTTCATCTT
Sandwich 32    CCACCATCATATTTTGTTACTGCACCAT--GAAAAACCGACTCTTTTTTTTAGCAA----------------CGACCT
NBH 3          CCACCATCATATTTTGTTACTGCACCAT--GAAAAGCCGACTCTTTTTTTTATCAA----------------CGACCT
NBH 4          CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCCGACTCTTTTTTTTATCAA----------------CGACCT
NBH 6          CCACCATCATATTTTGTTACTGCACCAT--GAAAAGCCGACTCTTTTTTTTATCAA----------------CGACCT
NBH 7          CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCCGACTCTTTTTTTTATCAA----------------CGACCT
NBH 9          CCACCATCATATTTTGTTACTGCACCATG--AAAANCCGACTCTTTTTTTTATCAA----------------TCGACCT
NBH 12         CCACCATCATATTTTGTTACTGCACCATGTGAAAANCCGACTCTTTTTTTTTANCAA----------------CGACCT
NBH 13         CCACCATCATATTTTGTTACTGCACCAT--GAAAAGCCGACTCTTTTTTTTATCAA----------------CGACCT
NBH 307        CCACCATCATATTTTGTTACTGCACCATGTGAAAAGATGACTCGTTTTT--AGCAACGACCTCTTGTGGTTCATCCT
PJ 1           CCACCATCATATTTTGTTACTGCACCATG--AAAAGCCGACTCTTTTTTTTATCAACGACCTC---------------
PJ 2           CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCCGACTCTTTTTTTTTATCAACGACCTCTTGTGGTTCATCYT
PJ 3           CCACCATCATATTTTGTTACTGCACCATG--AAAAGCCGACTCTTTTTTTTTATCAACGACCTC---------------
PJ 4           CCACCATCATATTTTGTTACTGCACCATG--AAAAGCCGACTCTTTTTTTTTATCAACGACCTC---------------
PJ 5           CCACCATCATATTTTGTTACTGCACCATG--AAAAGCCGACTCTTTTTTT-ATCAACGACCTC---------------
PJ 6           CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCCGACTCTTTTTTTTTATCAACGACCTCTTGTGGTTCATCCT
PJ 19          CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCCGACTCTTTTTTTTTATCAACGACCTCTTGTGGTTCATCCT
PJ 26          CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCCGACTCTTTTTTTTTATCAACGACCTCTTGTGGTTCATCCT
PO Grand 13    CCACCATCATATTTTGTTACTGCACCATGTGAAAAGCTGACTCTTTTTT--AGCAATAACCTCTTGTGGTTCATCCT
               ****************************   ****   ***** *****  * ***
                                              PPP     P        P
```

```
                1111111111111111111111111111111111111111111111112222222222222222222222222222222222
                5555566666666667777777777888888888899999999990000000000111111111122222222233
Taxon           5678901234567890123456789012345678901234567890123456789012345678901234567890123456789
-------------------------------------------------------------------------------------------------------
Sandwich 6      CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAGACTTCTATA
Sandwich 7      CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
Sandwich 8      CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
Sandwich 11     CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
Sandwich 12     CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
Sandwich 14     CCTGATAGAATTAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
Sandwich 27     CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
Sandwich 32     CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 3           CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 4           CCTGGTA-AATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 6           CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 7           CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 9           CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 12          CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTNATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 13          CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
NBH 307         CCTGGTAGAATAAAATGCTCAAATAAGCATTGTCTTTTTCATGATTGTATTGGTCATGTCTATAATAAACTTCTACA
PJ 1            -CTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PJ 2            CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PJ 3            -CTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PJ 4            -CTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PJ 5            -CTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PJ 6            CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PJ 19           CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PJ 26           CCTGGTAGAATAAAATGCTCAAATGAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
PO Grand 13     CCTGGTAGAATAAAATGCTCAAATAAGCA--GTCTTTTTCATGATTGTATTGGTCATGGCTATAATAAACTTCTATA
                 *** ** *** ************ ****  ******** ****************** ******** ******* *
                    P      P           P                                P       P       P

                2222222222222222222222222222222222222222222222222222222222222222222222333333333
                3333333344444444445555555555566666666667777777777888888888899999999990000000000
Taxon           2345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
-------------------------------------------------------------------------------------------------------
Sandwich 6      AGCAATC-ATTC-TTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
Sandwich 7      AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
Sandwich 8      AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
Sandwich 11     AGCAATC-ATTCCTTATCGTTCCTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
Sandwich 12     AGCAATCCATTC-TTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTTTGTGATTTGCTTGAAATCTA
Sandwich 14     AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTCTGAGTT-GTGAT---------------
Sandwich 27     AGCAATC-AATCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
Sandwich 32     AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
NBH 3           AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
NBH 4           AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
NBH 6           AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
NBH 7           AGCAATC-AATCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
NBH 9           AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
NBH 12          AGCAATC-ATTC-TTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
NBH 13          AGCAATC-ATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTGTGAGTT-GTGAT---------------
NBH 307         AGCAATCCATTC-TTATTGTTCTTAAGAAATATAATTTTCGTAAACAGTTTTTAANTTGTGATTTGCTTGAAATCTA
PJ 1            AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PJ 2            AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PJ 3            AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PJ 4            AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PJ 5            AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PJ 6            AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PJ 19           AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PJ 26           AGCAATCCATTCCTTATCGTTCTTAAGAAATATCATTTTTGTAAACAGTTTTGAGTT-GTGAT---------------
PO Grand 13     AGCAATCCATTC-TTATCGTTCTTAAGAAATATAATTTTTGTAAACAGTTTTGAGTT-GTGATTTGCTTGAAATCTA
                ******* * ** **** **** ********** ***** ********          *****
                        P       P    P          P     P          P P P
```

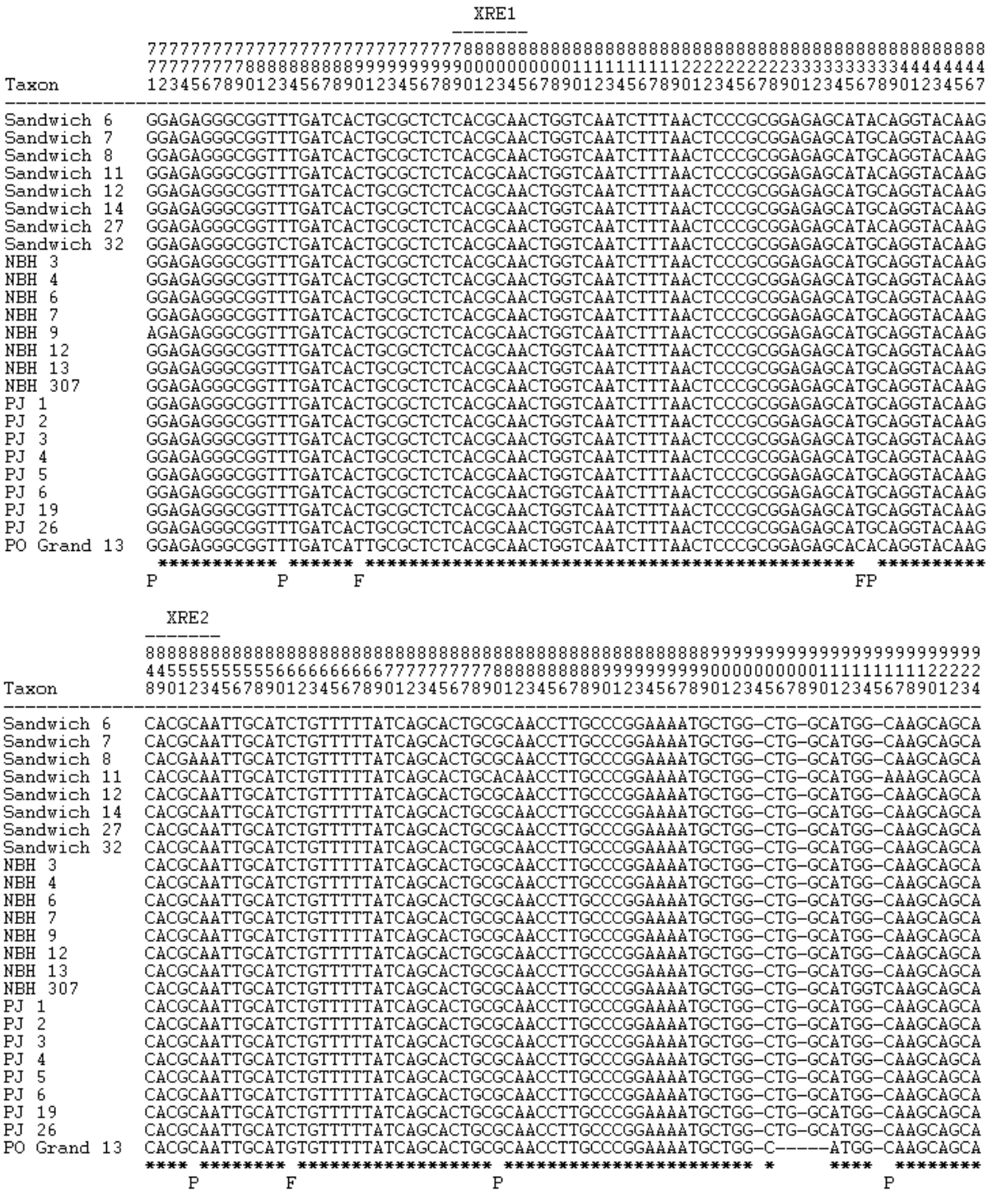

Taxon
Sandwich 6
Sandwich 7
Sandwich 8
Sandwich 11
Sandwich 12
Sandwich 14
Sandwich 27
Sandwich 32
NBH 3
NBH 4
NBH 6
NBH 7
NBH 9
NBH 12
NBH 13
NBH 307
PJ 1
PJ 2
PJ 3
PJ 4
PJ 5
PJ 6
PJ 19
PJ 26
PO Grand 13
GRE 1/2 site

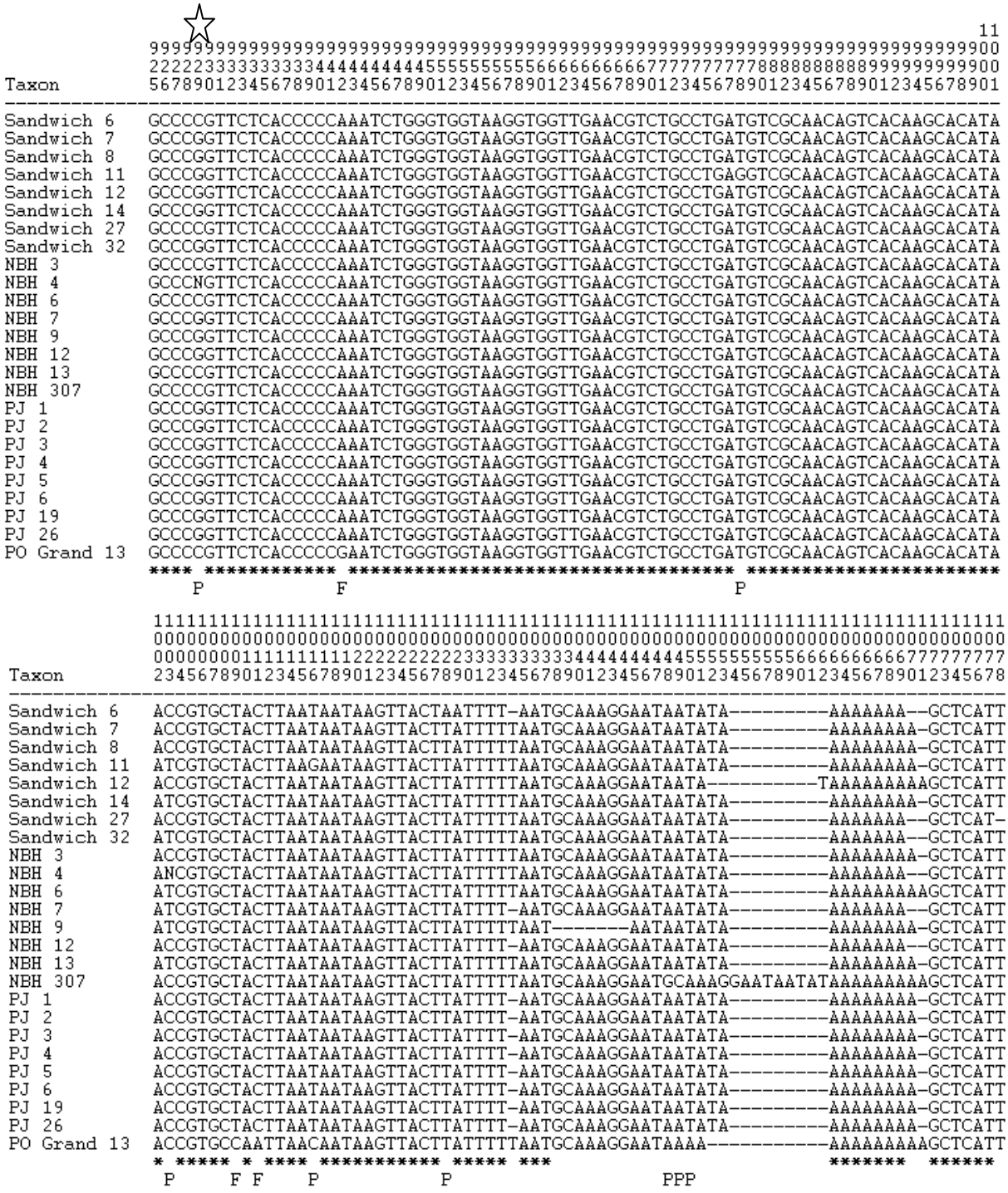
```
            4444444444444444444444444444444444444455555555555555555555555555555555555555555
            6666666777777777788888888889999999990000000000111111111122222222233333333333
Taxon       3456789012345678901234567890123456789012345678901234567890123456789012345678901
-------------------------------------------------------------------------------------------
Sandwich 6  GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCGG
Sandwich 7  GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
Sandwich 8  GGATTCAACAGGTTTTTGTTT------TTATTGTGTACTGTTACACCACATATCGGTTAAAGGAATAAATATTTCAG
Sandwich 11 GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
Sandwich 12 GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
Sandwich 14 GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
Sandwich 27 GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
Sandwich 32 GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
NBH 3       GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
NBH 4       GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
NBH 6       GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
NBH 7       GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
NBH 9       GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
NBH 12      GGATTAAAGAGGTTTTTGTTT------TTATTGTGTACTGTTACACCACATATCGGTTAAAGGAATAAATATTTCAG
NBH 13      GGATTCAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
NBH 307     GGATTTAACAGATTTTTTTTTTTATTATTATTGTATAATGTTACACCACATATCGGTTAAAGGAAGAAATATTTCAG
PJ 1        GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PJ 2        GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PJ 3        GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PJ 4        GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PJ 5        GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PJ 6        GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PJ 19       GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PJ 26       GGATTAAACAGATTTTTGTTT------TTATTGTATAATGTTACACCACATATCGGTAAAAGGAATAAATATTTCAG
PO Grand 13 GGATTTAACAGATTTTTGTTT------TTATTGTGTAATGTTACACCACATATCGGTTAAAGGGAGAAATATTTCAG
            ***** ** ** ***** ***      ******* ** ****************** ***** * ********* *
                 P  P  P     P              P  P                   P     F P         P
```

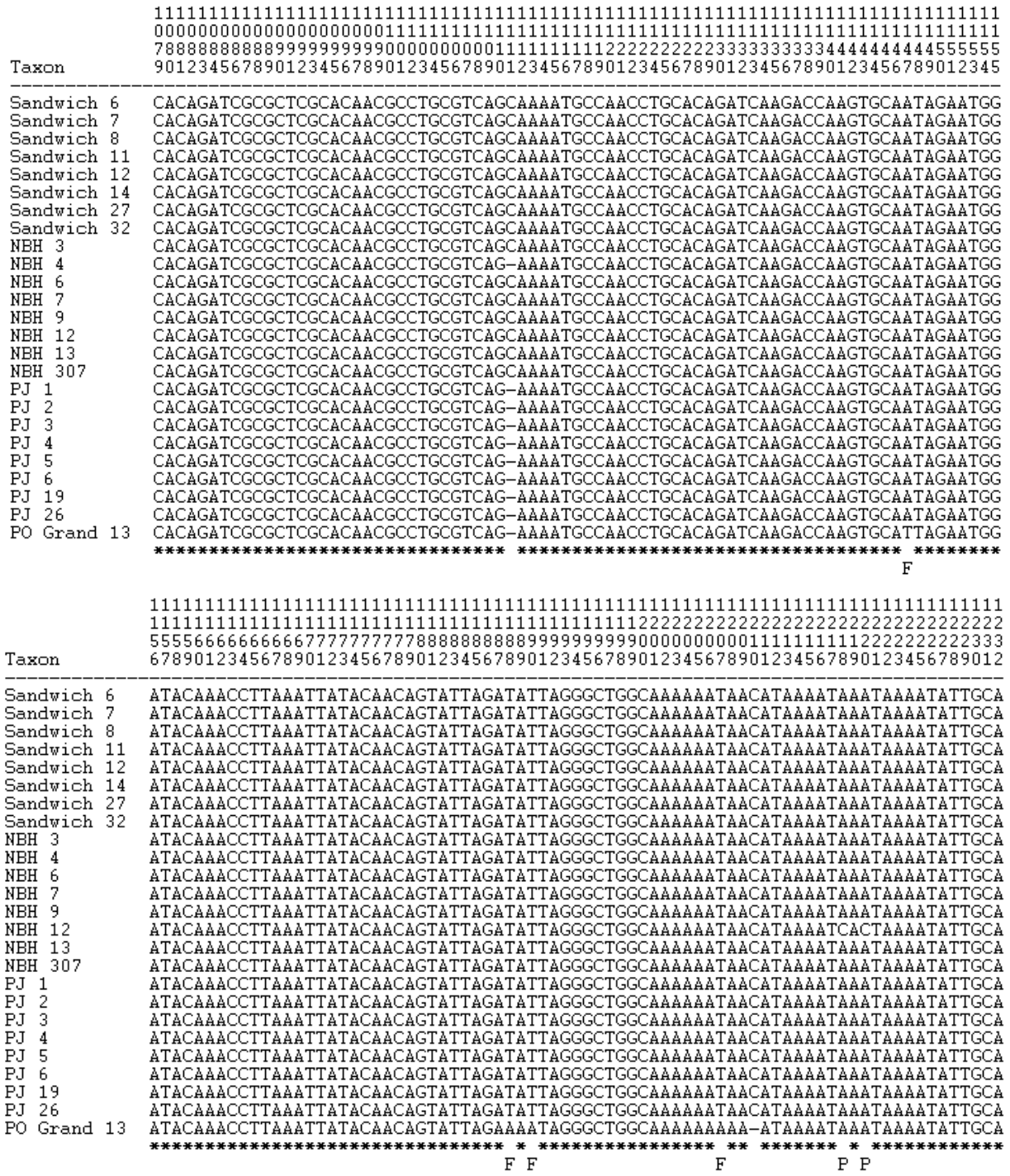
```
            5555555555555555555555555555555555555555555555555555555555555555555566666666666666666
            4444444445555555555566666666667777777777888888888899999999990000000000111111
Taxon       0123456789012345678901234567890123456789012345678901234567890123456789012345678
-------------------------------------------------------------------------------------------
Sandwich 6  GTTAACTAAAAGCTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCA-T-TTTTTTTTCCATTTACGC
Sandwich 7  GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATTCTTTTTTTTCCATTTACGT
Sandwich 8  GTTAACTAAAAGCTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCC-T-TTTTTTCCCCATTTACGT
Sandwich 11 GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCA-T-TTTTTTTTCCATTTACGT
Sandwich 12 GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATTTTTTTTTTTTCCATTTACGT
Sandwich 14 GTTAACTAAAAGCTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATTTTTTTTTTTTCCATTTACGT
Sandwich 27 GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCAAT-TTTTTTTTCCATTTACGT
Sandwich 32 GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATTTTTTTTCCCCCATTTACGT
NBH 3       GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
NBH 4       GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT-TTTTTTTTCCATTTACGT
NBH 6       GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
NBH 7       GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT-TTTTTTTTCCACTTACGT
NBH 9       GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCA-T-TTTTTTTTCCATTTACGT
NBH 12      GTTAACTAAAAGCTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCA-T-TTTTTTTTCCATTTACGT
NBH 13      GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
NBH 307     GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATTTTTTTTTTT-CCATTTACAT
PJ 1        GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PJ 2        GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PJ 3        GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PJ 4        GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PJ 5        GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PJ 6        GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PJ 19       GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PJ 26       GTTAACTAAAAACTAAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTCCCATT--TTTTTTTCCATTTACGT
PO Grand 13 GTTAACTAAAAACTGAAGTATAGCTGAAAAACAGTTTGGTGCGCTCATTGTTGCCATTTTTTTTCC--CCATTTACAT
            *********** ** ************************************* **           ** ****
                      P  F                                     F  P       FP     P    PP
```

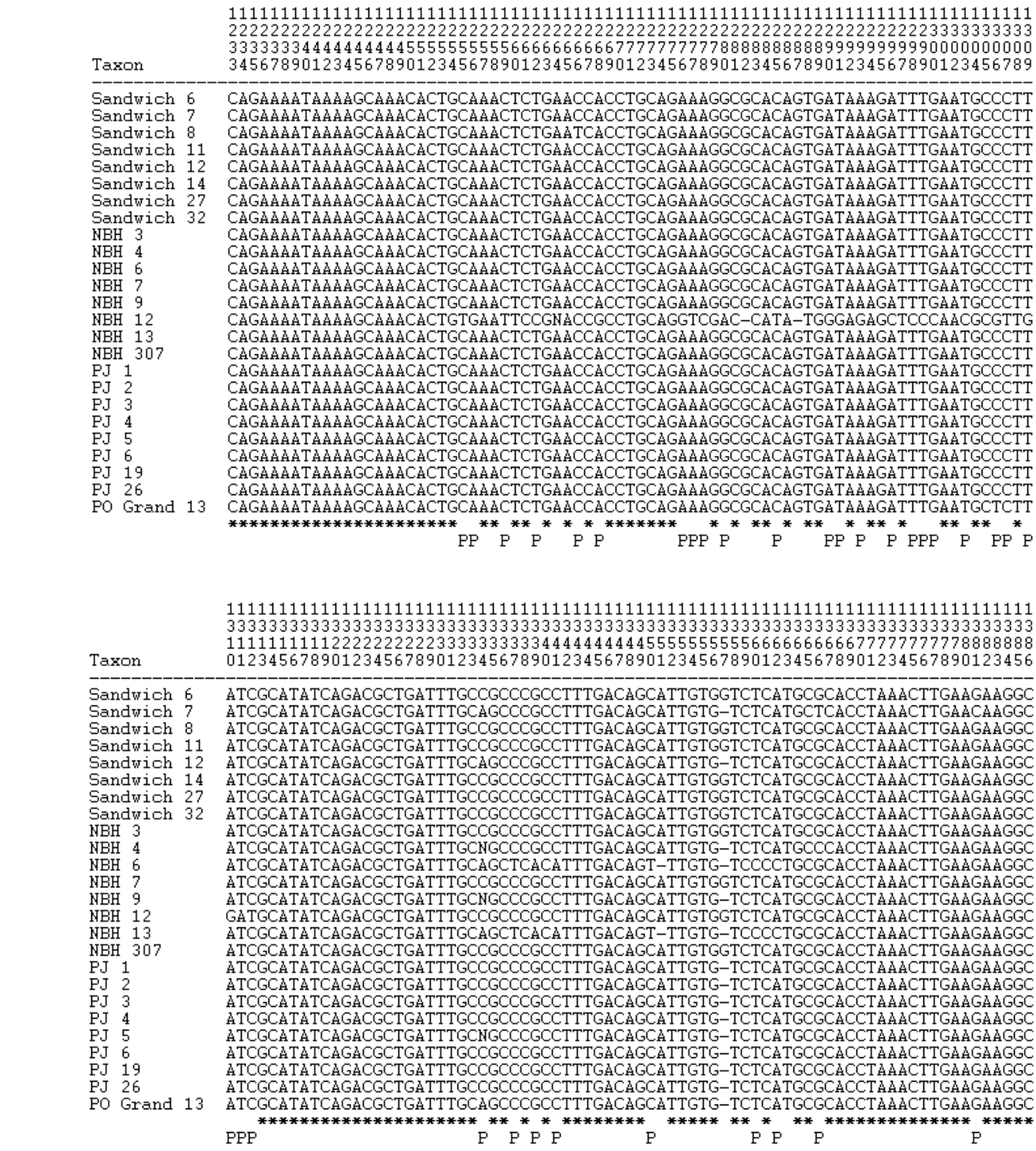
```
                666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666
                111222222222233333333334444444444555555555566666666667777777777888888888899999
Taxon           789012345678901234567890123456789012345678901234567890123456789012345678901234567890123
--------------------------------------------------------------------------------------------------------------
Sandwich 6      GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
Sandwich 7      GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
Sandwich 8      GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCAGTGCGCAGACTTTAATAAACGTAATGA
Sandwich 11     GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCAGTGCGCAGACTTTAATAAACGTAATGA
Sandwich 12     GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
Sandwich 14     GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
Sandwich 27     GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
Sandwich 32     GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 3           GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 4           GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 6           GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 7           GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 9           GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 12          GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 13          GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
NBH 307         GTTTAAAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 1            GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 2            GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 3            GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 4            GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 5            GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 6            GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 19           GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PJ 26           GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCGTTGCGCAGCCTTTAATAAACGTAATGA
PO Grand 13     GTTTAGAGCCGTTTATCCAAGTTGTAATGCTGCATTATTTAAGCCTCCCATTGCGCAGCCTTTAATAAACGTAATGA
                ***** ****************************************  ******* ****************
                     P                                          PP       P

                666666777777777777777777777777777777777777777777777777777777777777777777777777777777777
                999999000000000011111111112222222222333333333344444444445555555555666666666667
Taxon           456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
--------------------------------------------------------------------------------------------------------------
Sandwich 6      ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
Sandwich 7      ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
Sandwich 8      ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
Sandwich 11     ACTCGTCTGCCCTTGCGCTCCGACAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCATGCATCCTCCTCGAAGG
Sandwich 12     ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
Sandwich 14     ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
Sandwich 27     ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
Sandwich 32     ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
NBH 3           ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
NBH 4           ACTCGTCTGCCCTTGCGCTCCGCGAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCATGCATCCTCCTCGAAGG
NBH 6           ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
NBH 7           ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCATGCATCCTCCTCGAAGG
NBH 9           ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
NBH 12          ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCACGCATCCTCCTCGAAGG
NBH 13          ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
NBH 307         ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
PJ 1            ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGA
PJ 2            ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGA
PJ 3            ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
PJ 4            ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
PJ 5            ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGA
PJ 6            ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
PJ 19           ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
PJ 26           ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTGGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
PO Grand 13     ACTCGTCTGCCCTTGCGCTCCGGCAGCAGTAGCGCCTCCTCCCCTCCCTGTCCGTGCCCA-GCATCCTCCTCGAAGG
                ********************  ****** ***************************** ***************
                                     PP      F                                          P
```


XRE1
Taxon
Sandwich 6
Sandwich 7
Sandwich 8
Sandwich 11
Sandwich 12
Sandwich 14
Sandwich 27
Sandwich 32
NBH 3
NBH 4
NBH 6
NBH 7
NBH 9
NBH 12
NBH 13
NBH 307
PJ 1
PJ 2
PJ 3
PJ 4
PJ 5
PJ 6
PJ 19
PJ 26
PO Grand 13
P
P
F
FP
XRE2
Taxon
Sandwich 6
Sandwich 7
Sandwich 8
Sandwich 11
Sandwich 12
Sandwich 14
Sandwich 27
Sandwich 32
NBH 3
NBH 4
NBH 6
NBH 7
NBH 9
NBH 12
NBH 13
NBH 307
PJ 1
PJ 2
PJ 3
PJ 4
PJ 5
PJ 6
PJ 19
PJ 26
PO Grand 13
P
F
P
P

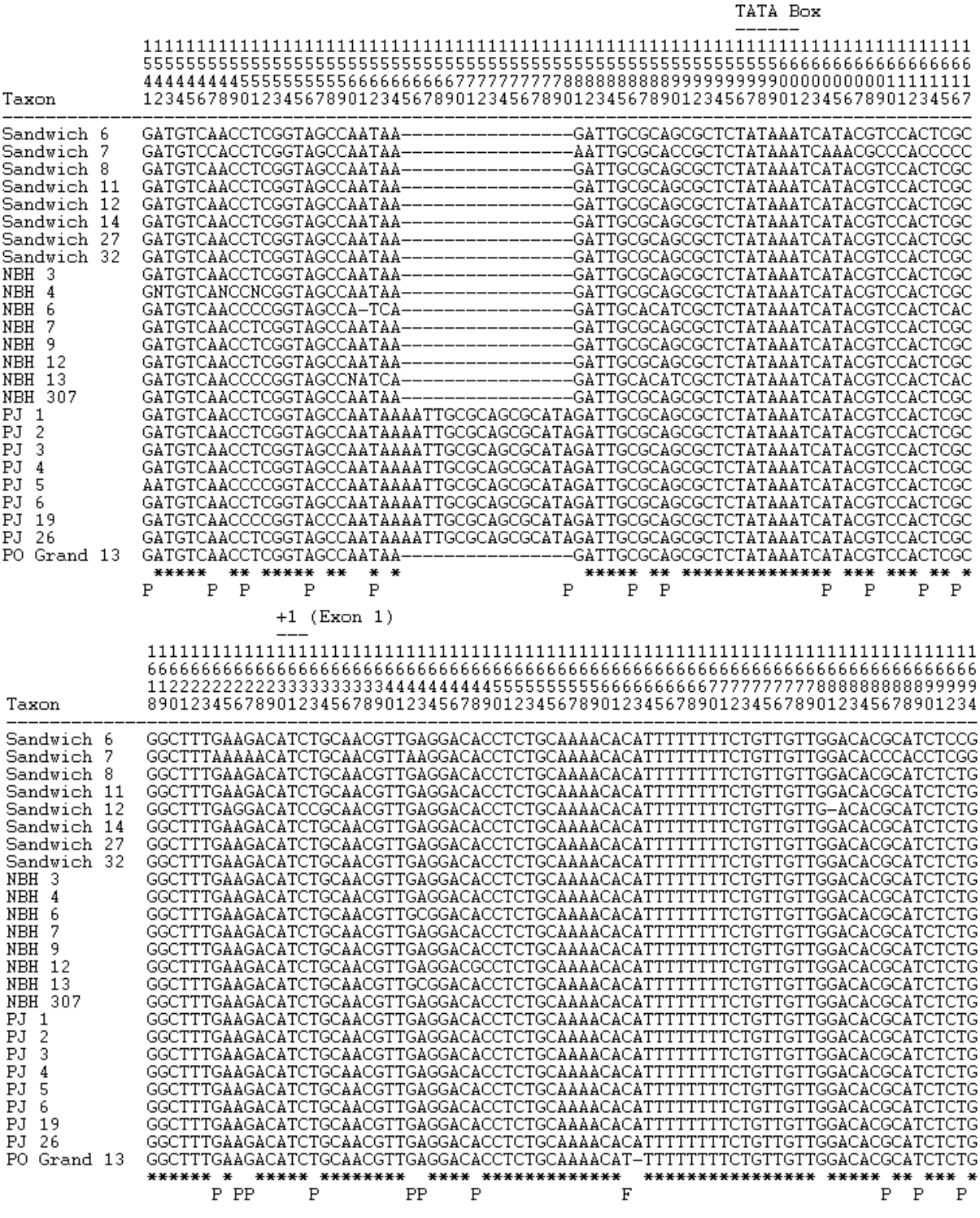

Taxon
Sandwich 6
Sandwich 7
Sandwich 8
Sandwich 11
Sandwich 12
Sandwich 14
Sandwich 27
Sandwich 32
NBH 3
NBH 4
NBH 6
NBH 7
NBH 9
NBH 12
NBH 13
NBH 307
PJ 1
PJ 2
PJ 3
PJ 4
PJ 5
PJ 6
PJ 19
PJ 26
PO Grand 13


Taxon
Sandwich 6
Sandwich 7
Sandwich 8
Sandwich 11
Sandwich 12
Sandwich 14
Sandwich 27
Sandwich 32
NBH 3
NBH 4
NBH 6
NBH 7
NBH 9
NBH 12
NBH 13
NBH 307
PJ 1
PJ 2
PJ 3
PJ 4
PJ 5
PJ 6
PJ 19
PJ 26
PO Grand 13


Taxon
Sandwich 6
Sandwich 7
Sandwich 8
Sandwich 11
Sandwich 12
Sandwich 14
Sandwich 27
Sandwich 32
NBH 3
NBH 4
NBH 6
NBH 7
NBH 9
NBH 12
NBH 13
NBH 307
PJ 1
PJ 2
PJ 3
PJ 4
PJ 5
PJ 6
PJ 19
PJ 26
PO Grand 13

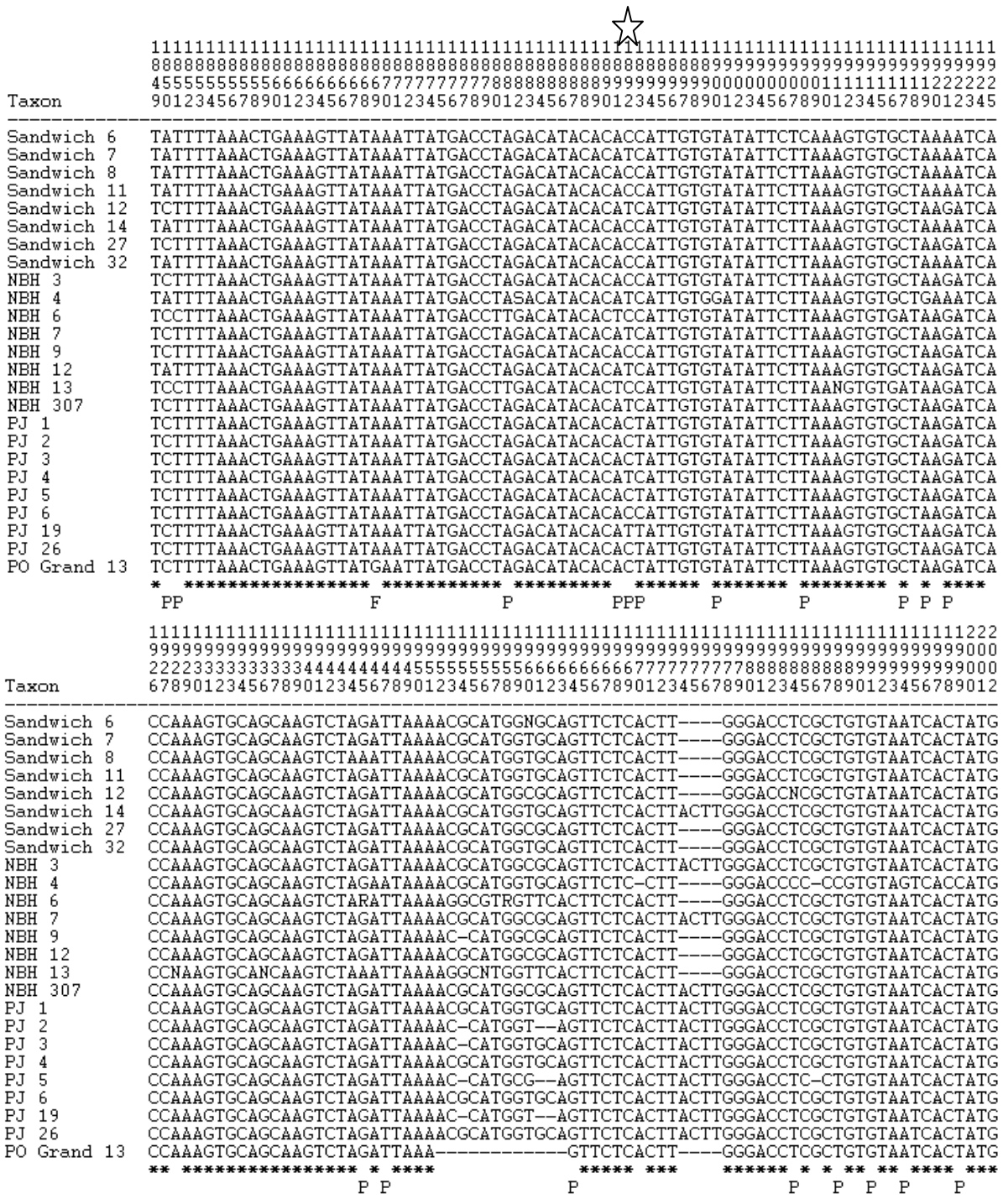
```
                                XRE3
                               ------
            1111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
            3333333333333444444444444444444444444444444444444444444444444444444444444444444444444444444444
            8889999999999000000000011111111112222222222333333333344444444445555555556666
Taxon       7890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123
-------------------------------------------------------------------------------------------------------------
Sandwich 6  GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
Sandwich 7  GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
Sandwich 8  GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
Sandwich 11 GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
Sandwich 12 GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
Sandwich 14 GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
Sandwich 27 GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
Sandwich 32 GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
NBH 3       GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
NBH 4       GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTTACCAACTT
NBH 6       GGTTG-CACTTTGTAATGCTTGCGAATTGCGTAGCGCCCGGACCACCCCCCCGAC-CACCTTCTTTCTTT---------
NBH 7       GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
NBH 9       GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
NBH 12      GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
NBH 13      GGTTG-CACTTTGTAATGCTTGCGAATTGCGTAGCGCCCGGACCACCCCCCCGAC-CACCTTCTTTCTTT---------
NBH 307     GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 1        GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 2        GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 3        GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 4        GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 5        GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 6        GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 19       GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PJ 26       GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
PO Grand 13 GGTAGACACTTTGTAATGCACGCGAATTGTGTACCGCCAGGACCACACACAGACACACCCACCAACTTT---------
            *** * *************  ******** *** **** ******* * * *** ****  *   ****
               P               PP        P   P    P       P P P        PP PPP

            1111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
            4444444444444444444444444444444444444445555555555555555555555555555555555555555555555555555555
            6666667777777777888888888899999999990000000000111111111122222222223333333333
Taxon       4567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
-------------------------------------------------------------------------------------------------------------
Sandwich 6  -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
Sandwich 7  ---TTTTTTTCTACGGCCCCAAACTTCTTTCAGGGCAGGAAATTAAAAACGGGCCCTCGAATGGAGGAGGGGAAAG
Sandwich 8  -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
Sandwich 11 -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
Sandwich 12 -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
Sandwich 14 -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
Sandwich 27 -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
Sandwich 32 -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
NBH 3       -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
NBH 4       TTTTTTTTTTTCTNNNGCNCCAAACTTCNTTCNTGGNNGGGAATTAAANANNGGNNCNCANANGGNGGNGGGGANAT
NBH 6       --TTTCTTTTTCTACTGCTCCCCACTTCATTCATGGCAGGGAATGAAAAACAGGCACTGAGATGGAAGAAGGGAGAT
NBH 7       -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
NBH 9       -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
NBH 12      -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
NBH 13      --TTTCTTTTTCTACTGCTCCCCACTTCATTCATGGCAGGGAATGAAAAACAGGCACTGAGATGGAAGAAGGGAGAT
NBH 307     -TTTTTTTTTTCTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
PJ 1        -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGARAT
PJ 2        -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGAGAT
PJ 3        -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGAGAT
PJ 4        -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGAGAT
PJ 5        -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGAAAT
PJ 6        -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGAGAT
PJ 19       -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGAAAT
PJ 26       -TTTTTTTTTTCTACTGCTCCAAACTTCATTCAGGGCAGGGAATTAAAAACAGGCACTCAAATGGAGGAGGGGAGAT
PO Grand 13 -TTTTTT-----CTACTGCTCCAAACTTCATTCATGGCAGGGAATTAAAGACAGGCACTCAGATGGAGGAGGGGAGAT
              *** *    **    * **  ***** ***  **  ** *** *** *  **       * **  *  **** *
                 P        P  P  PP    P    P     P   P   P  P   P  PPP    P  P    P P
```


TATA Box
Taxon
Sandwich 6 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
Sandwich 7 GATGTCCACCTCGGTAGCCAATAA----------AATTGCGCACCGCTCTATAAATCAAACGCCCACCCCC
Sandwich 8 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
Sandwich 11 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
Sandwich 12 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
Sandwich 14 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
Sandwich 27 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
Sandwich 32 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
NBH 3 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
NBH 4 GNTGTCANCCNCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
NBH 6 GATGTCAACCCCGGTAGCCA-TCA----------GATTGCACATCGCTCTATAAATCATACGTCCACTCAC
NBH 7 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
NBH 9 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
NBH 12 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
NBH 13 GATGTCAACCCCGGTAGCCNATCA----------GATTGCACATCGCTCTATAAATCATACGTCCACTCAC
NBH 307 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 1 GATGTCAACCTCGGTAGCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 2 GATGTCAACCTCGGTAGCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 3 GATGTCAACCTCGGTAGCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 4 GATGTCAACCTCGGTAGCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 5 AATGTCAACCCCGGTACCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 6 GATGTCAACCTCGGTAGCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 19 GATGTCAACCCCGGTACCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PJ 26 GATGTCAACCTCGGTAGCCAATAAAATTGCGCAGCGCATAGATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
PO Grand 13 GATGTCAACCTCGGTAGCCAATAA----------GATTGCGCAGCGCTCTATAAATCATACGTCCACTCGC
+1 (Exon 1)
Taxon
Sandwich 6 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTCTGTTGTTGGACACGCATCTCCG
Sandwich 7 GGCTTTAAAAACATCTGCAACGTTAAGGACACCTCTGCAAAACACATTTTTTTTCTGTTGTTGGACACCCACCTCGG
Sandwich 8 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
Sandwich 11 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
Sandwich 12 GGCTTTGAGGACATCCGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTG-ACACGCATCTCTG
Sandwich 14 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
Sandwich 27 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
Sandwich 32 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 3 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 4 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 6 GGCTTTGAAGACATCTGCAACGTTGCGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 7 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 9 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 12 GGCTTTGAAGACATCTGCAACGTTGAGGACGCCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 13 GGCTTTGAAGACATCTGCAACGTTGCGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
NBH 307 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 1 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 2 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 3 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 4 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 5 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 6 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 19 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PJ 26 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACACATTTTTTTTTCTGTTGTTGGACACGCATCTCTG
PO Grand 13 GGCTTTGAAGACATCTGCAACGTTGAGGACACCTCTGCAAAACAT-TTTTTTTTCTGTTGTTGGACACGCATCTCTG

```
                                                                  Intron 1
                                                                  ------
            1111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
            6666677777777777777777777777777777777777777777777777777777777777777777777777777777777777777777777
            9999900000000001111111111222222222233333333334444444444555555555566666666667
Taxon       5678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901
------------------------------------------------------------------------------------------------------------------
Sandwich 6  GAATTAGAGTGTTCGTTTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
Sandwich 7  GAATTAGAGGGTTCGCCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
Sandwich 8  GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
Sandwich 11 GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
Sandwich 12 GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATTACCTGCATGATAAAAGTTATTTCT
Sandwich 14 GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
Sandwich 27 GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
Sandwich 32 GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
NBH 3       GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATTACCTGCATGATAAAAGTTATTTCT
NBH 4       GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
NBH 6       GAATTACAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATTACCTGCATGATAAAAGTTATTTCT
NBH 7       GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
NBH 9       GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
NBH 12      GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATTACCTGCATGATAAAAGTTATTTCT
NBH 13      GAATTACAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATTACCTGCATGATAAAAGTTATTTCT
NBH 307     GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATTACCTGCATGATAAAAGTTATTTCT
PJ 1        GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PJ 2        GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PJ 3        GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PJ 4        GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PJ 5        GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PJ 6        GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PJ 19       GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PJ 26       GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATAACCTGCATGATAAAAGTTATTTCT
PO Grand 13 GAATTAGAGTGTTCGTCTTCTTTTTTTATCATTCAGCTAAAGGTAAGGTGATTACCTGCATGATAAAAGTTATTTCT
            ****** ** *****  ************************************ ************************
                  P  P     PP                                        P

            1111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
            7777777777777777777777777777778888888888888888888888888888888888888888888888888888888888888888888
            7777777788888888899999999990000000000111111111122222222223333333333444444444
Taxon       2345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
------------------------------------------------------------------------------------------------------------------
Sandwich 6  AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATAAGTCATATGTACAGTACATGCTATATGCAA
Sandwich 7  AAGGAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATAAGTCATATGTACAGTACATGCTATATGCAA
Sandwich 8  AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATAAGTCATATGTACAGTACATNCTATATGCAA
Sandwich 11 AAGGAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATAAGTCATATGTACAGTACATGCTATATGCAA
Sandwich 12 ATGAAAGGATTGAGAGGATTATTCAAGACGTAGATTTACTGTTAATCAGTCATATGTACAGTACATACTATATGCAA
Sandwich 14 AAGRAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATAAGTCATATGTACAGTACATGCTATATGCAA
Sandwich 27 AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTTAATCAGTCATATGTACAGTACATGCTATATGCAA
Sandwich 32 AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATAAGTCATATGTACAGTACATGCTATATGCAA
NBH 3       AAGAAAGGATTGAGAGGATTATTTAAGACGTAGATTTACTGTTAATCAGTCATATGTACAGTACATACTATATGCAA
NBH 4       AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTTAATCAGTCATATGTACAGTACATGCTATATGCTA
NBH 6       AAGAAAGGATTGAGAGGATTATTCAAGACGTAGATTTACTGTTAATCAGTCATATGTACAGAACATACGATATGCAA
NBH 7       AAGAAAGGATTGAGAGGATTGTTTAGGATGTAGATTTACTGTTAATCAGTCATATGTACAGAACATACTATATGCAA
NBH 9       AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTTAATCAGTCATATGTACA-----TACTATATGCAA
NBH 12      AAGAAAGGATTGAGAGGAATATTTAAGACGTAGATTTACTGTGAATAAGTCATATGTACAGTACATNCTATATGCAA
NBH 13      AAGAAAGGATTGAGAGGATTATTCAAGACGTAGATTTACTGTTAATCAGTCATATGTACAGTACATACGATATGCAA
NBH 307     AAGAAAGGATTGAGAGGATTATTTAAGACGTAGATTTACTGTTAATCAGTCATATGTACAGTACATACTATATGCAA
PJ 1        AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATMAGTCATATGTACA-----TACTATATGCAA
PJ 2        AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATCAGTCATATGTACA-----TACTATATGCAA
PJ 3        AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATCAGTCATATGTACA-----TACTATATGCAA
PJ 4        AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATCAGTCATATGTACA-----TACTATATGCAA
PJ 5        AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTTAATCAGTCATATGTACA-----TACTATATGCAA
PJ 6        AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATCAGTCATATGTACA-----TACTATATGCAA
PJ 19       AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATCAGTCATATGTACA-----TACTATATGCAA
PJ 26       AAGAAAGGATTGAGAGGATTGTTTAGGACGTAGATTTACTGTGAATCAGTCATATGTACA-----TACTATATGCAA
PO Grand 13 AAGAAAGGATTGAGATGATTATTTAAGACGTAGATTTACTGTTAATCAGTCATATGTACA-----TACTATATGCAA
            * * *********** ** * ** * ** ************* *** *************     * * ****** *
             P P          F  P P  P P  P           P   P                  P P     P
```


Taxon
Sandwich 6 TATTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTCAAAGTGTGCTAAAATCA
Sandwich 7 TATTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACATCATTGTGTATATTCTTAAAGTGTGCTAAAATCA
Sandwich 8 TATTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAAATCA
Sandwich 11 TATTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAAATCA
Sandwich 12 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACATCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
Sandwich 14 TATTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAAATCA
Sandwich 27 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
Sandwich 32 TATTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAAATCA
NBH 3 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
NBH 4 TATTTTAAACTGAAAGTTATAAATTATGACCTASACATACACATCATTGTGGATATTCTTAAAGTGTGCTGAAATCA
NBH 6 TCCTTTAAACTGAAAGTTATAAATTATGACCTTGACATACACTCCATTGTGTATATTCTTAAAGTGTGATAAGATCA
NBH 7 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACATCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
NBH 9 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
NBH 12 TATTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACATCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
NBH 13 TCCTTTAAACTGAAAGTTATAAATTATGACCTTGACATACACTCCATTGTGTATATTCTTAANGTGTGATAAGATCA
NBH 307 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACATCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 1 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACTATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 2 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACTATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 3 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACTATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 4 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACATCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 5 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACTATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 6 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACCATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 19 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACATTATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PJ 26 TCTTTTAAACTGAAAGTTATAAATTATGACCTAGACATACACACTATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PO Grand 13 TCTTTTAAACTGAAAGTTATGAATTATGACCTAGACATACACACTATTGTGTATATTCTTAAAGTGTGCTAAGATCA
PP F P PPP P P P P P
Taxon
Sandwich 6 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGNGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
Sandwich 7 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
Sandwich 8 CCAAAGTGCAGCAAGTCTAAATTAAAACGCATGGTGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
Sandwich 11 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
Sandwich 12 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGCGCAGTTCTCACTT----GGGACCNCGCTGTATAATCACTATG
Sandwich 14 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
Sandwich 27 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGCGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
Sandwich 32 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
NBH 3 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGCGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
NBH 4 CCAAAGTGCAGCAAGTCTAGAATAAAACGCATGGTGCAGTTCTC-CTT----GGGACCCC-CCGTGTAGTCACCATG
NBH 6 CCAAAGTGCAGCAAGTCTARATTAAAAGGCGTRGTTCACTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
NBH 7 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGCGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
NBH 9 CCAAAGTGCAGCAAGTCTAGATTAAAAC-CATGGCGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
NBH 12 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGCGCAGTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
NBH 13 CCNAAGTGCANCAAGTCTAAATTAAAAGGCNTGGTTCACTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
NBH 307 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGCGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PJ 1 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PJ 2 CCAAAGTGCAGCAAGTCTAGATTAAAAC-CATGGT--AGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PJ 3 CCAAAGTGCAGCAAGTCTAGATTAAAAC-CATGGTGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PJ 4 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PJ 5 CCAAAGTGCAGCAAGTCTAGATTAAAAC-CATGCG--AGTTCTCACTTACTTGGGACCTC-CTGTGTAATCACTATG
PJ 6 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PJ 19 CCAAAGTGCAGCAAGTCTAGATTAAAAC-CATGGT--AGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PJ 26 CCAAAGTGCAGCAAGTCTAGATTAAAACGCATGGTGCAGTTCTCACTTACTTGGGACCTCGCTGTGTAATCACTATG
PO Grand 13 CCAAAGTGCAGCAAGTCTAGATTAAA----------GTTCTCACTT----GGGACCTCGCTGTGTAATCACTATG
P P P P P P P P P

Taxon	22222222222 00000000000 00000001111 34567890123
Sandwich 6	TTGCACAACAC
Sandwich 7	TTGCACAACAC
Sandwich 8	TTGCACAACAC
Sandwich 11	TTGCACAACAC
Sandwich 12	TTGCACAACAC
Sandwich 14	TTGCACAACAC
Sandwich 27	TTGCACAACAC
Sandwich 32	TTGCACAACAC
NBH 3	TTGCACAACAC
NBH 4	TTGCACAACAC
NBH 6	TTGCACAACAC
NBH 7	TTGCACAACAC
NBH 9	TTGCACAACAC
NBH 12	TTGCACAACAC
NBH 13	TTGCACAACAC
NBH 307	TTGCACAACAC
PJ 1	TTGCACAACAC
PJ 2	TTGCACAACAC
PJ 3	TTGCACAACAC
PJ 4	TTGCACAACAC
PJ 5	TTGCACAACAC
PJ 6	TTGCACAACAC
PJ 19	TTGCACAACAC
PJ 26	TTGCACAACAC
PO Grand 13	TTGCACAACAC
	**********

SUMMARY AND DISCUSSION

This thesis sought to understand whether adaptation in *F. heteroclitus* to anthropogenic pollution has genetic basis in three independent populations. Prior to this thesis, there was very little knowledge of the overall genetic variation in populations of *F. heteroclitus* populations living in polluted areas as compared to clean, reference sites. This thesis used genome-wide scans to establish whether there may be genetic adaptation through the process of natural selection acting on these populations and if so, to identify those signatures of selection.

Summary of Chapters

Chapter one focused on establishing and carrying out a molecular technique to accurately capture genetic variation in nine populations of *F. heteroclitus*. The *F. heteroclitus* genome is not sequenced, and so a technique called amplified fragment length polymorphisms (AFLP) was employed. In total, 296 loci were scored among *F. heteroclitus* from New Bedford Harbor and its reference sites, 336 loci in *F. heteroclitus* from Newark and its two reference sites, and 299 loci in *F. heteroclitus* from Elizabeth River and its two reference sites. Among all populations, 450 distinct loci were scored. F_{ST} values were calculated for each locus among each pairwise comparison and compared against simulated values (Beaumont, Nichols, 1996). When empirical F_{ST} values were compared to a simulated distribution of F_{ST} values, 24 distinct outlier loci were identified among pairwise comparisons of pollutant impacted *F. heteroclitus* populations and both surrounding

reference populations. Two outlier loci were shared between New Bedford Harbor and Elizabeth River populations, and two different loci were shared between Newark Bay and Elizabeth River populations. In sum, 1% to 6% of loci are implicated as being under selection or linked to areas of the genome under selection in three *F. heteroclitus* populations that reside in polluted estuaries. Shared loci among polluted sites indicate that selection may be acting on multiple loci involved in adaptation, and loci shared between polluted sites potentially are involved in a generalized adaptive response. This chapter illustrated the utility of AFLP analyses to detect selection in natural *F. heteroclitus* populations.

Due to the anonymous nature of AFLPs, they are difficult to sequence using traditional methods (*e.g.*, gel extraction, reconstitution and sequencing). Chapter two sought to employ a new, high-throughput sequencing method to establish its utility in sequencing AFLP fragments to identify single nucleotide polymorphisms (SNPs) for population genetic studies in *F. heteroclitus* and related species *F. grandis* and *F. similis* (Gulf of Mexico) and *F. majalis* (Atlantic coast). High-throughput genotyping methods were also used to verify previously identified SNPs in expressed sequence tags (ESTs) among populations and species of the *Fundulus* genus. Using the 454 FLX pyrosequencing system, 1,464 distinct contigs were obtained. Two-hundred and sixty-one SNPs were identified in 96 of these contigs using a new statistical framework developed by the Bustamante lab at Cornell University. We further verified 81 of these SNPs on the MassARRAY genotyping platform. SNPs showed latitudinal clinal variation separating northern and southern populations and established isolation by distance in *F. heteroclitus* populations. In *F. grandis*, the SNPs isolated from *F. heteroclitus* were less polymorphic (*i.e.*, 74% were monomorphic versus

11.2 in *F. heteroclitus*) but still established isolation by distance. Markers differentiated species and populations. These approaches were used to quickly determine differences within the *Fundulus* genome and provide markers for population genetic studies. The process by which this data was collected also provided a framework with which researchers in the future could sample and carry out population genetic studies on non-model species. This is especially important, because a majority of species do not have a sequenced genome, since it is still cost prohibitive. Thus establishing a method which can generate hundreds of markers in a non-sequenced genome accurately is necessary. While our method was not perfect, given several changes such as making reduced representation libraries from many pooled individuals for SNP discovery followed by individual genotyping, laboratories studying adaptation in any number of species would benefit from the method which could provide valuable information with respect to how genetics and molecular mechanisms affect gene frequency.

SNP data collected in chapter two was further analyzed in chapter three to detect signatures of selection in polluted populations of *F. heteroclitus*. Three statistical tests, which tested different polymorphic parameters of the loci (minor allele frequency, frequency of each allele and heterozygosity), were used to detect outlier behavior among 367 SNPs across nine populations (3 polluted and 6 reference). The first test, an F_{ST} test, identified 30 outlier loci (8.2%) in the New Bedford Harbor triad, 9 (2.5%) in the Newark Bay triad, and 43 (11.7%) in the Elizabeth River triad. To further test the robustness of our outliers, an association test, which calculates p-values based on the strength of association with polluted populations was used. The association test identified 30 SNP (8.2%) which were

significantly associated with the polluted New Bedford Harbor model ($p \leq 0.01$), 7 (1.9%) in the Newark Bay triad, and 29 (7.9%) in the Elizabeth River triad. Differences between minor allele frequencies (MAF) also were tested between polluted and reference sites. In the New Bedford Harbor population, 22 SNPs (6.0%) had significantly different MAF between polluted and both reference populations. In the Newark Bay and Elizabeth River populations, 9 and 18 SNPs (2.5% and 4.9%, respectively) had significantly different MAF between polluted and reference populations. Within each triad, 6-15 SNPs were identified as outliers in all three tests: the New Bedford Harbor triad had 12, the Newark Bay triad had 6, and the Elizabeth River triad had 15. Among all triads, 15 of these SNPs occur in coding regions. Only one of these 15 SNPs, a SNP in β_2-microglobulin, is non-synonymous. One SNP was significant in all three tests and across all three triads: a SNP in the first intron of the phase I xenobiotic metabolizing enzyme CYP1A. In total, 1.6% to 4.1% of the loci that we examined in the three triads is selectively important or linked to areas of the genome that are selectively important. Extrapolating across the genome, 110-300 genes are involved in adaptation in natural populations in a short time. This is the first estimate that has been made for *F. heteroclitus* populations residing in polluted areas, and also has implications for understanding adaptation in other species adapted to complex environments.

Few studies on natural populations have shown that a single gene is responsible for a given phenotype; rather many phenotypes in natural populations are the composite of genotypes across many loci. In order to address this, many researchers are now using quantitative trait loci (QTL) studies to identify multiple markers that may contribute to a particular phenotype. However, a prerequisite for a QTL study is a linkage map, which

assists in the determination of associate phenotypes with specific identifiable regions of the genome (Stinchcombe, Hoekstra, 2007). There are several non-model species where linkage maps have been made by maintaining species in captivity and breeding crosses in the laboratory including butterflies (*Heliconius, Bicyclus*), sticklebacks (*Gasterous*), deermice (*Peromyscus*), monkeyflowers (*Mimulus*), and columbines (*Aqueligia*). An example where QTL studies have identified candidate genes underlying a phenotype in a natural population is in the three-spine stickleback, where the QTL study identified a 10 Mb region containing a large effect contributing to adaptive variation in pelvic morphology between oceanic and lake populations, and the gene *Pitx1* was identified as the underlying genetic cause for this difference (Shapiro *et al.*, 2004). However, QTL studies often only focus on the large effect regions and do not characterize small effect regions, which may be contributing to a phenotype. In the case of *F. heteroclitus*, where the phenotype is resistance to complex mixtures of contaminants, a QTL study may identify large effects (such as CYP1A transcription or protein activity), but will most likely fail to investigate small effects (if they are identified at all). The study conducted in chapter three provides evidence that the phenotype may be composed of many small genotype effects across the genome, making it harder to point to one responsible, underlying adaptive locus. That said, chapter three did identify several SNPs in the CYP1A promoter as being selectively important. The last chapter of this thesis sought to understand the role of these SNPs in a known phenotype for *F. heteroclitus* residing in polluted populations.

In an attempt to link an outlier SNP to a functionally significant phenotype, the role of the CYP1A SNP found to be under selection in all three triads was explored in the fourth

data chapter. In all three Superfund populations, CYP1A is refractory to induction by prototypical inducers (Bello *et al.*, 2001; Elskus *et al.*, 1999; Meyer, Di Giulio, 2002; Nacci *et al.*, 1999), and this trait is associated with resistance to PAH, PCB and dioxin toxicity (Bello *et al.*, 2001; Nacci *et al.*, 1999). Potentially, the SNP in the first intron of CYP1A affects transcription or is linked to SNPs affecting transcription. Thus, data chapter 4 sought to quantify the amount of transcription *in vitro* between populations containing different CYP1A promoter haplotypes linked to the selectively important SNP in the New Bedford Harbor triad. Haplotypes were determined by sequencing 1.5 kb of CYP1A promoter and an additional 500 basepairs containing the first exon and first intron for 8 individuals in the New Bedford polluted population and flanking reference sites, Point Judith and Sandwich. Representative promoters from each population were cloned into a luciferase-containing pGL3-basic vector and transfected into PLHC-1 cells. Luciferase activity was quantified under control conditions (vehicle) or after dosage with a prototypic CYP1A inducer, 3-methylcholanthrene (3-MC).

The study found significant variation in the CYP1A promoter. The vast majority of this variation fell in regions that have no described function in binding transcription factors. While sequence variation distinguished population divergence and established genetic distance by geographical distance, when analyses for evolutionary relationships were conducted on the entire CYP1A promoter through phylogenetic analysis, there was no pattern. Point Judith was the only population that formed its own monophyletic group. This population is physically isolated (with a man-made seawall). As a result, there was little variation in the promoter, most likely due to inbreeding in the population and a lack of

migrants into the population. This study, aside from describing the variation within the promoter, also sought to determine if the SNPs detected as being under selection in chapter three were in linkage disequilibrium with other SNPs along the promoter. Selective sweeps often act upon regions of a genome where a selectively important SNP causes hitchhiking effects. Thus, while our SNPs may not cause functional effects, they may be linked to SNPs or regions that have an underlying functional effect. This study, however, did not find any strong LD between these selectively important SNPs and the rest of the promoter, the first exon or the first intron. LD may in fact be with other SNPs or regions up or downstream of these SNPs. Tests for natural selection on the CYP1A promoter also failed to reject the null hypothesis that the promoter was evolving by random genetic forces.

In addition to understanding the overall genetic variation and evolutionary forces acting upon the promoter, this chapter sought to determine if there was a difference in the inducibility of the promoter to a prototypic PAH. CYP1A has been found to be refractory to induction in *F. heteroclitus* residing in polluted estuaries. If the promoter had a part in this phenotype, the expectation is that the inducibility would be less for the promoters from polluted individuals as compared to clean, reference individuals. A dose response curve on two individuals (one from Sandwich and one from New Bedford Harbor) showed that each of the promoter constructs was able to significantly induce luciferase expression over control when treated with doses of 3-MC higher than 0.05 μM for New Bedford and 0.1 μM for Sandwich. The promoter constructs also induced luciferase expression in a dose-dependent manner, and hit an asymptote at a dose of 1 μM. This 1 μM dose was then used to determine if there were population specific differences in the promoter constructs to induce luciferase

expression. There were significant population differences found: New Bedford Harbor had higher average fold induction over both its reference sites. This result does not explain why CYP1A is refractory to induction in the New Bedford Harbor population. It is possible that there are repressors not present in the cell line in which the transfection assays were carried out, or transcription factor binding sites well upstream of the 1.5kb that was cloned, which are necessary for the refractory phenotype. It is, however, interesting that the portion of the CYP1A promoter that was cloned from New Bedford Harbor was more sensitive to induction by 3-MC as compared to the reference sites. This sensitivity may reflect changes in the ability of the promoter to bind the heterodimer, AHR-ARNT, or may reflect a change to other transcription factor binding sites that were outside the scope of this study.

This thesis demonstrated the utility of high-throughput molecular biology and population genetics and genomics to examine selective forces in multiple natural populations. Anthropogenic selective forces have shaped the genomes of *F. heteroclitus* along the east coast of the United States by changing allelic frequencies of 1-7% of the genome in individuals exposed to constant high levels of pollution as compared to reference populations. Since such a large number of loci are under selection, it is hard to imagine that only one to a few of the selectively important loci are contributing to the resistance phenotype, but that it is the sum of the interactions between loci that allow *F. heteroclitus* to adapt to a very hostile environment.

Future Directions

AFLP Study (Chapter 1)

The AFLP study identified 24 outlying loci among pairwise comparisons of pollutant impacted *F. heteroclitus* populations and both surrounding reference populations. These loci should be sequenced and characterized in order to understand their overall function in the *F. heteroclitus* adaptive phenotype to pollution.

Sequencing: *Loci* were initially amplified through a PCR reaction of ligated DNA with primers containing a selective extension of three base pairs. A fourth and fifth selective nucleotide could be added to the selective extension to further decrease the total number of products generated and selectively amplify the candidate loci. With each additional selective nucleotide, the total number of amplified fragments should decrease by 75%. For example, if 70 *loci* are generated after the first selective extension (*primer* +3 selective nucleotides), a reaction using a primer with four selective nucleotides will generate approximately 18 fragments. If those 18 fragments are further amplified with a primer with five selective nucleotides, approximately four fragments will remain. Since each outlying locus is characterized by the primer set from which it was amplified as well as its size, it can be easily identified and isolated from other loci. Products could be resolved on a high percentage agarose gel to allow for isolation and purification. Loci could then be cloned into a TA vector, PCR-amplified using vector specific primers, and sequenced in both directions.

Identifying outlying loci: Sequences initially could be compared against GenBank using the Basic Local Alignment Search Tool (http://www.ncbi.nlm.nih.gov/BLAST) to determine similarity to sequences in the database. Loci sequences with no match will be

mapped to a *F. heteroclitus* bacterial artificial chromosome (BAC) library to identify proximal genes due to the absence of a sequenced genome.

Characterizing outlying loci: All identified loci represent a polymorphism among populations in the genomic DNA by virtue of the way they were identified, that being outlier distribution based on the presence or absence of a PCR product. Loci may represent one of three areas of genomic DNA: coding regions, regions proximal to coding regions, and non-coding (distal) regions.

Coding Regions: If a locus is within the coding region of a gene, it should be determined if the locus results in a synonymous or non-synonymous change to a codon. If the polymorphism changes the deduced, primary, amino acid sequence of the protein, the protein could be expressed using a bacterial or baculovirus expression system such as the VariFlex™ system (commercially available through Stratagene) to determine whether the altered amino acid results in an altered protein function. This will be more or less feasible depending on the protein and whether it has a defined function. Appropriate assays could then be developed to test the (altered) function of the protein. For example, ligand binding assays might be done with receptors while DNA binding assays might be run with transcription factors.

If the polymorphism is synonymous, the gene's expression should be determined. For genes with altered expression between populations, the contribution of the gene to a pollutant response could be tested with dose-response experiments using a mixture of pollutants found in the particular Superfund site and quantitative, real-time PCR.

Proximal Regions: Constructs could be made in sequences from the polluted population to assess changes in promoter strength (reporter constructs expressing luciferase activity) as was done in chapter four.

Distal Regions: Sequences could be compared against known, functional binding sites using TRANSFAC software as was completed in chapter four. If binding sites are determined, column affinity chromatography with site specific oligonucleotides could be used to purify transcription factors, which could be further characterized. To determine if the binding is biologically relevant, an *in vivo* footprinting assay could be run. These assays would determine if distal regions bound distinct transcription factors in polluted populations *versus* reference populations which could account for some of the resistance phenotype.

SNP Characterization (Chapters two and three)

For chapter two, several additional resources would have been useful for SNP identification and characterization. First, a fully sequenced genome would have provided a backbone through which the SNPs could have been mapped and polymorphisms within and between populations more easily identified. This also would have provided a context through which linkage and linkage disequilibrium could have been determined between SNPs. A recent study in the mosquito (*Anopheles gamniae*) used measures of LD (LD-based haplotype diversity analysis) to determine selection acting upon insecticide resistance mutations in several populations (Lynd *et al.*, 2010). If a genome was available for *F. heteroclitus*, a similar study could have been conducted with the SNP data collected in chapters two and three, providing additional statistical tests for selection. This may have also

identified larger genomic regions that were under selection, making those regions targets for future studies. Because not all of the SNPs identified in our study are directly under selection, but rather are the result of hitchhiking effects, a sequenced genome would allow for the calculation of LD between our SNPs and the rest of the genome in order to identify other loci that may be directly under selection by pollution. It is the hope of the *Fundulus* consortium to get a sequenced genome in the near future.

The SNPs identified as being under selection are rich targets to explore the resistance phenotype in the three independent polluted populations. In the Newark population, a SNP in β2-microglobulin changes an aspartic acid residue to an asparagine. β2-microglobulin is a component of MHC class I molecules. Future studies could look at the functional consequence of this non-synonymous change and whether it affects the function of the protein by modeling the folding of the gene, determining if the change affects the way the protein interacts with others in the MHCI complex or establishing whether there is a difference in the immune function of wild type or mutant β2-microglobulin *F. heteroclitus* to pathogen challenges. The other 14 SNPs that occur in coding regions all result in synonymous changes. These SNPs may be optimal or suboptimal changes to the codon because of codon bias and could be determined with statistical tests (Comeron, Aguade, 1998). It is thought that optimal codons help to achieve faster translation rates and high accuracy (Shields *et al.*, 1988; Sorensen *et al.*, 1989). The rate of translation could also be tested between wild and mutant-types through a system developed by Björnsson and Isaksson (Bjornsson, Isaksson, 1988). This test may elucidate a potential role for these changes in

translational efficiency. These SNPs could also be linked to other selectively important SNPs, but without a sequenced genome determining these linkages is not possible.

CYP1A Promoter (chapter 4)

For the CYP1A promoter, the functional activity of each of the XRE's has yet to be tested. In order to do so, deletion constructs containing none, one, and two XREs could be used to test if the inducibility of the promoter changes in the presence or absence of any one of the XREs. An *in vivo* footprinting study would also be useful to identify functional cis-acting elements along the promoter between populations. Furthermore, the genetic variation and promoter inducibility by xenobiotics needs to be tested in the other two triads. The SNP in the first intron was under selection in all three triads, so its LD in the other two polluted populations must be evaluated for the same region sequenced in chapter four. With a sequenced genome, the LD of these SNPs with more distal regions could also be evaluated. Additionally, potential enhancer and repressor regions much further upstream of what was sequenced could be investigated to provide insight into the high inducibility of the CYP1A promoter found in the New Bedford Harbor population *versus* its reference sites.

Overall Conclusion

This thesis established that *F. heteroclitus* living in polluted areas along the east coast of the United States are adapting to contamination through natural selection. Natural selection seems to be acting on many loci along the genome, and several of these areas are shared between polluted sites indicating a conserved evolutionary response. There are also

selectively important loci specific to each polluted site, indicating a specific selective force given a particular contaminant mixture. In total, without *a priori* sequence information, genome scans proved useful in identifying loci that have been selected for or are linked to areas that are selectively important and may be contributing to the resistance phenotype.

References

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences* **263**, 1619-1626.

Bello SM, Franks DG, Stegeman JJ, Hahn ME (2001) Acquired resistance to Ah receptor agonists in a population of Atlantic killifish (*Fundulus heteroclitus*) inhabiting a marine superfund site: in vivo and in vitro studies on the inducibility of xenobiotic metabolizing enzymes. *Toxicol Sci* **60**, 77-91.

Bjornsson A, Isaksson LA (1988) Test System for Measurement of Translational Activity in Vivo. *Nucleosides and Nucleotides* **7**, 565 - 569.

Comeron JM, Aguade M (1998) An evaluation of measures of synonymous codon usage bias. *Journal of Molecular Evolution* **47**, 268-274.

Elskus AA, Monosson E, McElroy AE, Stegeman JJ, Woltering DS (1999) Altered CYP1A expression in *Fundulus heteroclitus* adults and larvae: a sign of pollutant resistance? *Aquatic Toxicology* **45**, 99-113.

Lynd A, Weetman D, Barbosa S, *et al.* (2010) Field, Genetic, and Modeling Approaches Show Strong Positive Selection Acting upon an Insecticide Resistance Mutation in *Anopheles gambiae* s.s. *Molecular Biology and Evolution* **27**, 1117-1125.

Meyer J, Di Giulio R (2002) Patterns of heritability of decreased EROD activity and resistance to PCB 126-induced teratogenesis in laboratory-reared offspring of killifish (*Fundulus heteroclitus*) from a creosote-contaminated site in the Elizabeth River, VA, USA. *Mar Environ Res* **54**, 621-626.

Nacci D, Coiro L, Champlin D, *et al.* (1999a) Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. *Marine Biology* **134**, 9-17.

Shapiro MD, Marks ME, Peichel CL, *et al.* (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717-723.

Shields DC, Sharp PM, Higgins DG, Wright F (1988) Silent Sites in Drosophila Genes Are Not Neutral - Evidence of Selection among Synonymous Codons. *Molecular Biology and Evolution* **5**, 704-716.

Sorensen MA, Kurland CG, Pedersen S (1989) Codon Usage Determines Translation Rate in *Escherichia coli*. *Journal of Molecular Biology* **207**, 365-377.

Stinchcombe JR, Hoekstra HE (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**, 158-170.