Molecular Biology and Evolution

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, North Carolina State University

[2]Department of Marine Biology and Fisheries, Rosenstiel School of Marine and Atmospheric Sciences, University of Miami

***Corresponding author:** E-mail: moleksiak@rsmas.miami.edu.

Associate editor: Matthew Hahn

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 genomic approaches to identify randomly selected single nucleotide polymorphisms (SNPs) with evolutionarily significant patterns in three natural populations of *Fundulus heteroclitus* that inhabit and have adapted to highly polluted Superfund sites. Three statistical tests identified 1.4–2.5% of SNPs that were significantly different from the neutral model in each polluted population. These nonneutral patterns in populations adapted to highly polluted environments suggest that these loci or closely linked loci are evolving by natural selection. One SNP identified in all polluted populations using all tests is in the gene for the xenobiotic metabolizing enzyme, cytochrome P4501A (CYP1A), which has been identified previously as being refractory to induction in the three highly polluted populations. Extrapolating across the genome, these data suggest that rapid evolutionary change in natural populations can involve hundreds of loci, a few of which will be shared in independent events.

Key words: *Fundulus heteroclitus*, natural populations, genome scan, natural selection.

Research article

Introduction

Haldane estimated the mean rate of gene substitution as one per 300 generations and suggested that rates much larger than this would be detrimental to a species (Haldane 1957). Yet, due to human intervention, the global environment is changing rapidly (Vitousek et al. 1997; Kerr 2007) 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. Such rapid adaptation indicates a strong selective force. These three populations are exposed to some of the highest concentrations of aromatic hydrocarbon pollutants of any vertebrate species (Wirgin and Waldman 2004) and inhabit highly polluted Superfund sites (Elskus et al. 1999; Nacci et al. 1999; Meyer and Di Giulio 2002; Ownby et al. 2002), hazardous waste sites mandated for clean up by the Comprehensive Environmental Response, Compensation, and Liability Act of 1980. *Fundulus heteroclitus* from these chronically polluted sites are resistant to the aromatic hydrocarbons in their environment (e.g., embryos from the polluted sites are more than 300 times less sensitive to aromatic pollutants than embryos from clean reference sites; Nacci et al. 2010) as compared with nearby fish from relatively clean environments (Vogelbein et al. 1990; Black et al. 1998; Elskus et al. 1999; Nacci et al. 1999, 2002; Meyer et al. 2002; Ownby et al. 2002), and resistance in first and second generation embryos (Nacci et al. 2010) suggests that differential survival is due to genetic adaptation rather than physiological induction. In the research presented here, we used 354 single nucleotide polymorphisms (SNPs) to scan the genome to investigate and understand the evolutionary changes associated with rapid adaptation.

Genome wide scans have been used to explore adaptation in a variety of organisms (Grahame et al. 2006; Namroud et al. 2008; Nosil et al. 2008). Genome scans identify genetic markers that display patterns indicative of demographic effects or nonneutral divergence among populations (Luikart et al. 2003; Storz 2005; Stinchcombe and Hoekstra 2008). Yet, even when a specific SNP has a pattern of variation indicative of adaptation, the importance of this SNP, or the candidate gene, it is located on is still unknown because it may be linked to a different causative locus. Thus, further work often is needed to determine whether the SNP is causative or simply linked to a causative locus. An alternative use of genome wide scans (and the one used here) is to use unlinked SNPs to provide an estimate of the percentage of the genome involved in evolved changes.

In the research presented here, we sought to identify 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 minor allele frequencies (MAF-F_{max}). We applied these analyses to 354 randomly selected SNPs from both coding and noncoding regions in 180 individuals from nine populations.

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

http://mbe.oxfordjournals.org/

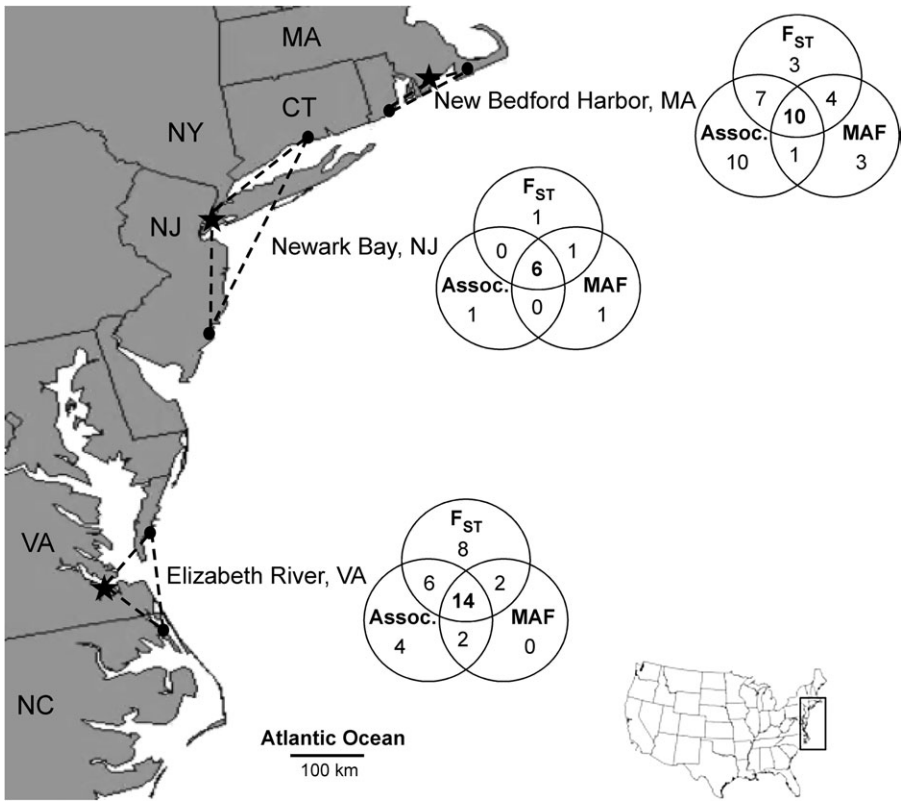

FIG. 1. *Fundulus heteroclitus* sampling sites along the East coast of the United States. Polluted sites are starred and flanked north and south by clean reference sites (circles) to form a triad. Each polluted site with its two clean reference sites is referred to as a triad of which there are three. Variation among the two reference sites captures the random/neutral variation. Thus, differences in a polluted population versus both reference populations most likely are due to evolved responses to pollution. Venn diagrams indicate the number of SNPs exhibiting nonneutral behavior using the three statistical tests: the F_{ST} modeling approach (F_{ST}), Association (Assoc.), and MAF-F_{max} (MAF).

Among three polluted populations, approximately 1.4–2.5% of loci have nonneutral, evolutionarily significant patterns of variation, where the significance reflects the power of our analysis. Considering the ecology and demography of the three polluted populations, these differences suggest that a substantial fraction of the genome is involved in the rapid adaptive response in these populations.

Materials and Methods

Fundulus 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 (fig. 1). Three of the nine collection sites were Environmental Protection Agency (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 and Martinez 1997), and DNA was resuspended in 50 µl 0.1X Tris–ethylenediaminetetraacetic acid 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 (354): SNPs that amplified in greater than 80% of *F. heteroclitus* and did not show an excess of heterozygosity. SNPs were randomly chosen from both coding (expressed sequence tags [ESTs]) and noncoding loci: 279 were coding and 75 were noncoding (supplementary table S1, Supplementary Material online).

We used Arlequin v.3.11 to calculate F_{ST} values for each SNP between populations using the analysis of molecular variance function (Excoffier et al. 2005). F_{ST} values were modeled to detect outliers using the FDIST2 program (Beaumont and Nichols 1996). Simulations were run for each pair of populations using the average heterozygosity of the empirical data with 20,000 iterations assuming ten demes, two populations, 20 individuals per sample, and a stepwise mutation model. The 99th 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 vs. both reference sites) were considered to be outliers and potentially under selection by pollution or linked to loci under selection.

We used the Expectation–Maximization (EM) algorithm (Excoffier and Slatkin 1995; Slatkin and Excoffier 1996) implemented in Arlequin 3.5 to test for statistically significant linkage disequilibrium (LD) between loci. This algorithm performs well with large sample sizes and can be used

to accurately test for LD between pairs of loci using unphased data (Slatkin and Excoffier 1996). When examining linkage among selectively important loci, we only looked at linkage in the reference populations because selection can make unlinked loci appear to be linked.

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 (i.e., are genotype frequencies different between the polluted population and both reference populations). 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 (i.e., are genotype frequencies different between the northern and southern clean, reference populations). A likelihood ratio test was used to determine which of the associations, that of the polluted population or demography between clean reference populations, 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 significant. A Bonferroni correction was applied to each triad to correct for multiple testing. Notice that the F_{ST} analyses use a model to predict the 99% confidence interval, whereas the association test uses a likelihood ratio test to determine whether association to polluted sites versus clinal divergence best explains the data. These analyses use similar data, but one compares across loci and the other compares each locus among populations. Principal component adjustment for population structure used the significant principal components for the nine populations based on an unsupervised Tracy–Widom statistic (Patterson et al. 2006).

The minor allele of the pooled populations within a triad was determined for each SNP, and MAFs were compared among the three populations within a triad to determine whether there were significant differences in the MAFs in a polluted site versus both reference sites. There is no a priori reason why only the polluted population would have a different MAF among populations within the triad. Within each population, the allele frequency of that overall minor allele was calculated for a random sampling of 15 of 20 individuals 100 times. One-hundred random samplings of 15 of 20 individuals have less than 1% probability that the same combination of individuals will be chosen more than once. A one-way analysis of variance (ANOVA) was performed on the iteratively derived values. To control for type I errors among all 100 iterations, we computed conservative, multiple test–corrected critical values for the F-statistic by the one-step adjustment method (Westfall and 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.

The ANOVA for MAF compares the polluted population to both flanking reference site populations to define significant differences in the MAF and to minimize the effect of demography. We are suggesting that a significant difference in MAF is likely due to a combination of selection and migration. This MAF test differs from the F_{ST} test in that we are asking whether there is a difference in allele frequency between polluted and reference populations without regard to all other loci. In contrast, the F_{ST} test models the distribution of F_{ST} values among loci. To determine whether MAF differences are significant, we used a F_{max} test (Westfall and Young 1993), a conservative family-wise correction for multiple comparisons which generates significant F values for each SNP by sampling the data and uses the maximum F value that occurs less than 1% of the time as the test statistic. Thus, the F_{max} test does not test a model for why there is a significant divergence in MAF but instead provides statistical support for the significance.

These three tests (F_{ST} distribution, association, and MAF) use the same SNP data and thus are not independent. However, each tests a different aspect of the data. The F_{ST} test determines whether the allele frequency is significantly greater between populations relative to within a population. For our test, we modeled these F_{ST} values to define the values that represent the outliers beyond the 99% CI. Importantly, a significant SNP had to be an outlier for two comparisons (polluted vs. both reference site populations) but could not be an outlier between the reference sites comparison. Thus, the F_{ST} test uses all alleles at a locus and tests whether they are outliers relative to all other loci. Although the association test is similar to the F_{ST} test (because loci with alleles with different frequencies will affect F_{ST} values), it identifies loci that have higher allele frequencies in polluted populations but not among reference populations. The association test uses a likelihood ratio test to determine which of the two association models best fit the data: 1) alleles significantly associated with the polluted population or 2) alleles significantly associated with demography (clinal change between reference site populations). Finally, the MAF test is a simple ANOVA to examine the frequency of minor alleles among the triad populations. Minor alleles are identified across all three populations within a triad, and thus there is no a priori reason why only the polluted population would have a different MAF among populations within the triad. Unlike the F_{ST} test, only minor alleles are examined, and the significance is based solely on a conservative statistical correction for multiple tests (and not on modeling of all other loci). Unlike the association test that uses categorical data with a chi-square statistic, the MAF test is based on the variation within versus among populations.

Results and Discussion

Demographics

The experimental design compared SNPs in each polluted population with those in two flanking, clean, reference site

populations located north and south of each polluted site (triads, fig. 1). Populations within triads show little to moderate differentiation based on amplified fragment length polymorphism (AFLP) and multilocus microsatellite estimates of F_{ST}, which range from 0.018 to 0.039 for AFLPs (Williams and Oleksiak 2008) and from 0.043 to 0.101 for microsatellites (Adams et al. 2006). Importantly, microsatellite analyses of these populations show no evidence of population bottlenecks (Adams et al. 2006). Similarly, the variation in gene expression in common gardened individuals from the polluted populations is not reduced (Fisher and Oleksiak 2007). If one assumes that some of the variation in expression is heritable (approximately 68% of individual variation is heritable in *Drosophila melanogaster*; Ayroles et al. 2009), then these microsatellite and gene expression data support the surprising conclusion that the heavily polluted Superfund populations did not suffer from a significant bottleneck. As we have suggested previously (Williams and Oleksiak 2008), this most likely reflects a moderate rate of migration from nonimpacted populations. Finally, even though one might expect that the severe selection pressure of aromatic hydrocarbon pollutants over a short time period would greatly decrease effective population sizes (Ne), effective population sizes range from 10^4 to 10^5 and are comparable among polluted and reference populations (Adams et al. 2006). Thus, the three polluted populations that have adapted to the high concentration of aromatic hydrocarbons show few signs of strong demographic effects (bottlenecks, reduced Ne). This most likely is due to the constant infusion of migrants which selection would have to overcome.

Clinal variations in allele frequency occur among *F. heteroclitus* populations, and a historical break exists between populations north and south of the Hudson River (Bernardi et al. 1993; Adams et al. 2006; Haney et al. 2009). To minimize this clinal, demographic effect, we took advantage of the population triads and statistically compared each polluted population to respective, flanking, clean reference populations (fig. 1). This 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 (Oleksiak 2010). Thus, the variation due to demography is accounted for by comparing the polluted population with the combined variation in the northern and southern reference populations. Significant divergence in a polluted population compared with both paired reference populations suggests that pollution or other environmental factors associated with the Superfund sites are affecting the change in allele frequency.

Tests of Random Divergence

The first statistical test examined the pattern of SNP frequencies using 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 (fig. 2) (Beaumont and Nichols 1996). This F_{ST} modeling approach tests whether loci are differentiated among populations in a neutral manner due to the effects of genetic drift and other random processes as modeled by the symmetrical Island Model or are significantly divergent due to natural selection. For our study, we identified loci with nonneutral patterns as outliers in each Superfund site population relative to its two reference site populations. Specifically, for each SNP, the F_{ST} value had to be improbably high for the comparison of polluted population with each reference population but not for the comparisons between the reference site populations (Williams and Oleksiak 2008). If each the comparison between polluted and reference population were independent, then the probability would be the joint probability ($P < 0.001$). It is this low joint probability (differing between each polluted and both reference populations but not between the clean reference populations) that is indicative of nonneutral changes. Twenty-four outlier loci (6.8%) were found in the New Bedford Harbor triad, eight (2.3%) in the Newark Bay triad, and 30 (8.5%) in the Elizabeth River triad. Taken at face value, these data suggest that between 2.3–8.5% of SNPs or loci linked to these markers are divergent in the Superfund polluted populations. Although 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 examined the probability of association of each SNP with a polluted population. This test calculates the strength of association of a SNP (with a chi-squared test) between sampling sites. Although similar in approach to the F_{ST} test, the association and F_{ST} tests assess significance (outlier behavior) in a different manner. The F_{ST} test uses evolutionary modeling to determine if divergence in allele frequencies is significant, whereas the association test coupled with the likelihood ratio test mathematically assesses the significance without respect to evolutionary models. To determine whether a SNP was statistically more associated with a polluted site versus both reference sites, the association test compared each SNP between a polluted site and the sum total of the flanking reference sites. To control for demographic effects independent of pollutant effects, a second association test was performed to assess whether a SNP was more associated with one particular reference site. Two calculated *P* values are displayed in figure 3: one based on the strength of association with polluted populations (red points; fig. 3) and the second based on the strength of association with reference populations (blue points; fig. 3). This test is similar to many association tests that examine the relationship between genetic markers and a phenotype. For this test, we are assuming that the fish from the Superfund populations share a common trait: the ability to survive in polluted water. This assumption is valid because fish from nearby reference sites do not survive when raised on sediments from the polluted sites (Burnett et al. 2007). 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, 28 SNPs (7.9%) were

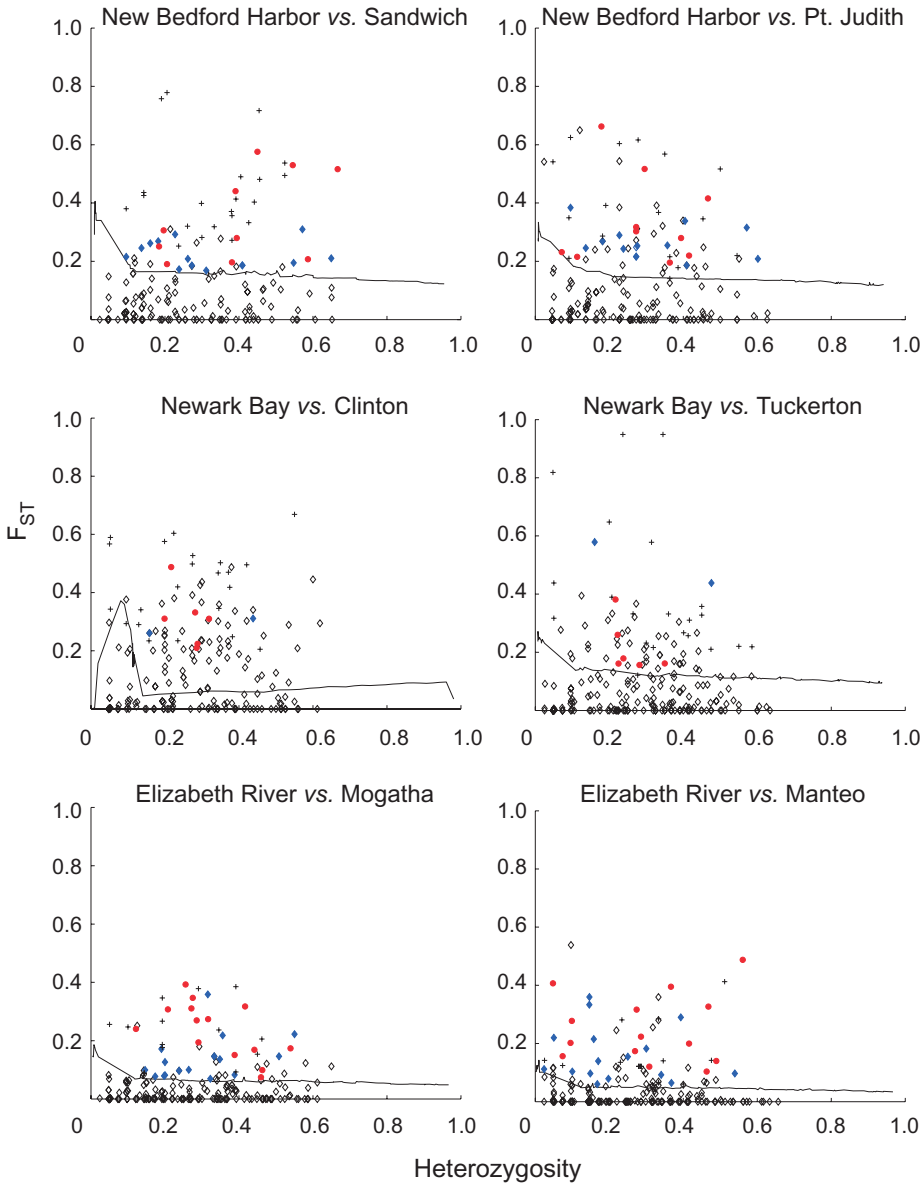

Fig. 2. F_{ST} modeling approach to detect selection. Empirical F_{ST} values are plotted against heterozygosity. The line demarks the 99th percentile estimated from a simulation model. Blue diamonds indicate SNPs that are significantly different between the polluted population and both reference populations 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 with one reference population.

significantly associated with the polluted model ($P \leq 0.01$). In the Newark Bay triad, fewer SNPs were associated with the polluted model (7; 2.0%). In the Elizabeth River triad, 26 SNPs (7.3%) were associated with the polluted model. This association test detected many demographic effects in the Newark Bay triad (blue points on the left half of the graph, fig. 3), which is located at the historical introgression zone between northern and southern *F. heteroclitus* populations (Bernardi et al. 1993; Adams et al. 2006; Haney et al. 2009); this large demographic effect may mask some of the effects of directional selection due to pollution.

To assess the ability of our experimental design to control for demographic effects, we repeated the SNP association tests using the significant principal components as covariates via the EIGENSTRAT method (Patterson et al. 2006; Price et al. 2006). Using significant principal components as covariates corrects for stratification in genome wide association (GWA) studies (Price et al. 2006). Except for two SNPs in the Elizabeth River triad, the SNPs we identified using our experimental design were a subset of those identified when the significant principal components were used as covariates, suggesting that our experimental design does control for demography in these populations.

However, correcting for demography also could remove selectively important loci if many loci are involved, and they have a strong effect on the principal component. This may be

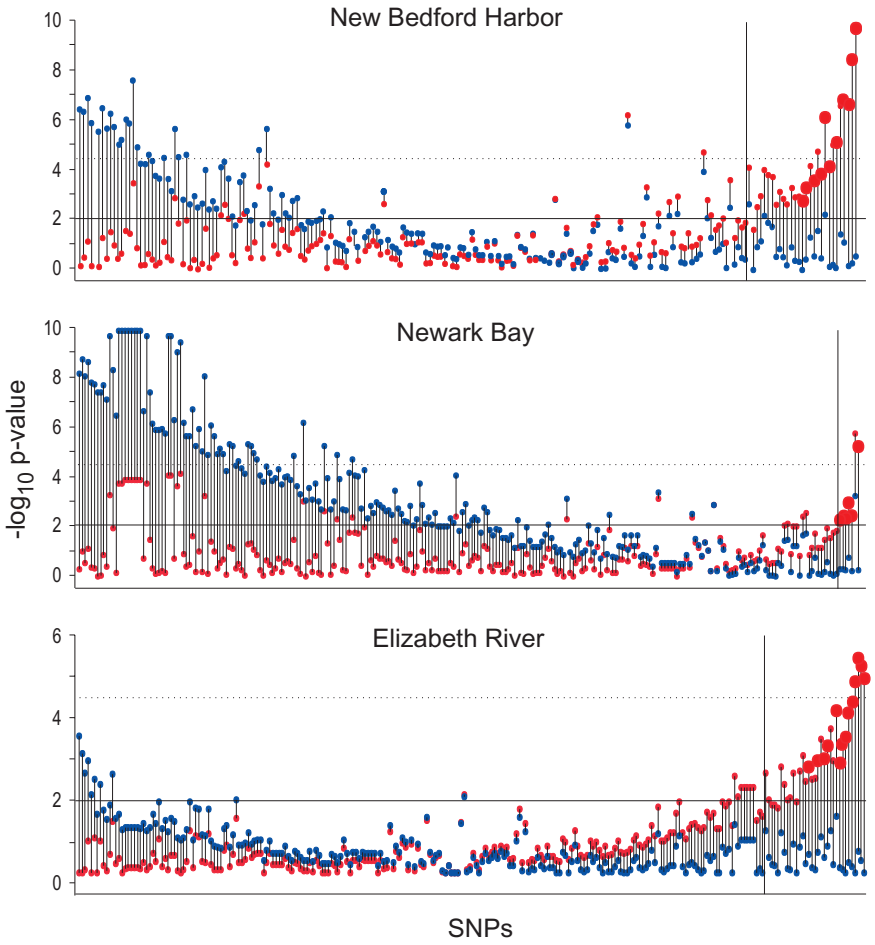

Fig. 3. Association test for detection of selection. Likelihood of association of each SNP with either the polluted site (red points) or reference sites (blue points) as a $-\log_{10} P$ value. The $-\log_{10} P$ value of 2 is marked by a black line, and the Bonferroni correction for multiple testing is marked by the dotted gray line ($-\log_{10} P$ value of 4.55). SNPs are ordered by increasing likelihood ratio test statistics. 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 $\leq$0.01 (points to the right of the vertical line on the *x* axis, supplementary table S1, Supplementary Material online). Large red dots denote SNPs also significant in the other two statistical tests.

especially important when a population experiences a selectively important, restricted environment, such as a Superfund site. For example, in panmictic populations, loci that were selected for in a subpopulation inhabiting a particular site would have the most influence on axes of variation (e.g., principal components), and the use of a principal component analysis correction would hide selection. Furthermore, tests for demography assume that any demography is defined by neutral loci and that most loci are neutral. Recent data suggest that one cannot assume that the majority of loci are neutral (Begun et al. 2007), and the positive correlation between polymorphism and recombination across many species suggests that most loci are affected by linked selection (Hahn 2008). If linked selection affects most loci, then correcting for demography will hide many of the genes undergoing adaptive natural selection (Hahn 2008).

We used a third test (MAF-F_{max} test) to determine whether the MAF of each SNP was significantly different in a polluted site versus both the reference sites. MAF were identified across all populations within a triad and thus represent a subset of alleles where there is no a priori reason why only the polluted population would have a different frequency among populations within the triad. 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. ANOVA was used on these 100 iterations to identify SNPs with MAF significantly different between the polluted and reference populations (fig.4). An F_{max} (Westfall and 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 and Young 1993). This conservative correction of multiple comparisons was used to determine whether MAF differences between the polluted and reference populations are significant (and not due to demography because the MAF of a SNP must be significantly different in the polluted site population as compared with both flanking reference site populations) but did not test a model of why they are different. In the New Bedford Harbor

Fig. 4. MAF-F_{max} test for detection of differences in SNP allele frequencies between polluted and reference sites. (A) 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. Sites left to right are as follows: (a) Sandwich, MA, USA; (b) New Bedford Harbor, MA, USA; (c) Point Judith, RI, USA; (d) Clinton, CT, USA; (e) Newark, NJ, USA; (f) Tuckerton, NJ, USA; (g) Magotha, VA, USA; (h) Elizabeth River, VA, USA; and (i) Manteo, NC, USA. (B) SNPs with allele frequencies significantly different in an ANOVA using F_{max} (Westfall and 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 (P) and both reference sites (R_1 and R_2) are plotted. In the New Bedford Harbor triad, sites left to right are (a) Sandwich, MA USA; (b) New Bedford Harbor, MA USA; and (c) Point Judith, RI, USA. In the Newark Bay triad, sites left to right are: (a) Clinton, CT, USA; (b) Newark, NJ, USA; and (c) Tuckerton, NJ, USA. In the Elizabeth River triad, sites left to right are: (a) Magotha, VA, USA; (b) Elizabeth River, VA USA; and (c) Manteo, NC, USA. Red dots denote SNPs exhibiting nonneutral behavior in all three statistical tests. The SNP exhibiting nonneutral behavior in all three triads and using all tests (CYP1A +268) is boxed.

population, 18 SNPs (5.1%) had significantly different MAF between polluted and both reference populations (fig. 4). In the Newark Bay and Elizabeth River populations, 8 and 18 SNPs (2.3% and 5.1%, respectively) had significantly different MAF between polluted and reference populations (fig. 4).

The F_{ST} modeling approach, association test, and MAF-F_{max} test all identified similar percentages of SNPs with nonneutral or nondemographic patterns in all three triads. However, the MAF-F_{max} test often identified the fewest number of SNPs (18, 8, and 18 for New Bedford Harbor, Newark Bay, and Elizabeth River triads, respectively vs. 24, 8, and 30 for the F_{ST} modeling approach and 28, 7, and 26 for the association test in these three triads). Because these three tests use the same individuals and genotyping data, they are not independent. Yet, they test different aspects of the data and identify a subset of nonoverlapping loci (fig. 1, supplementary table S2, Supplementary Material online). The MAF-F_{max} test, a mathematical bootstrapping

test of significance, asks whether a specific allele (the minor allele of the total population) occurs more frequently in the polluted population versus the flanking reference populations. The association test associates particular alleles with the polluted population. Both tests differ from the F_{ST} modeling approach, which is based on calculating differences in heterozygosities between populations and comparing them with a neutral evolutionary model. These tests use different aspects of the polymorphism spectrum measured by SNPs (MAF, frequency of each allele and heterozygosity).

SNPs identified as significant in more than one test are statistically more powerful and less likely to suffer from type I errors. These SNPs with low type I error probabilities may be causal themselves or in linkage with a causal change. Here, we explore the candidate causal changes. Within each triad, 6–14 SNPs were identified as outliers in all three tests: the New Bedford Harbor triad had ten, the Newark Bay triad had six, and the Elizabeth River triad had 14 (fig. 1, supplementary table S1, Supplementary Material online). Among all triads, 13 of these SNPs occur in coding regions. Only one of these 14 SNPs, a SNP in β_2-microglobulin, is nonsynonymous. β_2-microglobulin is vital to the immune response and is noncovalently associated with the heavy chain of the major histocompatibility complex class I antigens (Bernabeu et al. 1984). The G to A SNP in β_2-microglobulin changes an aspartic acid residue to an asparagine one predominantly in the polluted Newark Bay population. The other 12 SNPs that occur in coding regions all result in synonymous SNPs. Synonymous SNPs (as well as nonsynonymous ones) 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.

Although SNPs may just be linked to causal genes, there is some evidence that the gene closest to the SNP is causal. In a similar study comparing polymorphisms in four populations of *Arabidopsis lyrata*, two populations adapted to heavy metals in the soil and two more populations from reference sites, all of the 96 variants detected as different between the polluted and reference populations were in exons, introns, or within 1 kb of stop or start codons of 81 genes (Turner et al. 2010). The polymorphisms that are most strongly associated with soil type are enriched at heavy metal detoxification and calcium and magnesium transport loci, suggesting a causal link. So too, the loci we identified as selectively important may be causal.

If one assumes that the loci identified in all three tests are in fact undergoing natural selection, then 1.7–4.0% of the loci that we examined in the three triads is selectively important or linked to areas of the genome that are selectively important. To determine whether these loci are evolving independently (i.e., are unlinked), we used the EM algorithm (Excoffier and Slatkin 1995; Slatkin and Excoffier 1996) to test for statistically significant LD between loci. In the New Bedford Harbor triad, two of the ten selectively important loci are in significant LD with each other, in the Newark Bay triad, two of the six selectively important loci are in significant LD with each other, and in the Elizabeth River triad, 8 of the 14 selectively important loci are in significant LD with each other. Thus, there are 9 rather than 10, 5 rather than 6, and 7 rather than 14 unlinked loci or loci groups in the New Bedford Harbor, Newark Bay, and Elizabeth River triads, respectively. Considering this linkage among selectively important loci, the 1.7–4% of selectively important loci reduces to 1.4% and 2.5% (5–9 loci) assuming all nonselective loci are unlinked. However, not all nonselective loci are unlinked. On average, 48% of the SNPs (170/354) show no LD, suggesting that the percentages of selectively important loci are underestimates. Additionally, because we used 20 individuals per population, the power of our tests is not exceptionally high (Park et al. 2010). Thus, these percentages likely represent a minimum number of selectively important loci. With more statistical power, the actual number could be much higher.

Among these selectively important 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 and 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 and Oleksiak 2008), most likely due to migrants. Another explanation for the lack of fixed SNPs between polluted and reference populations is that different SNPs might be linked to the same locus, and thus different polymorphisms hitchhiked along with the locus when it became fixed. This could occur because different polymorphisms existed in the different ancestral populations.

A central question in outbred natural populations is whether similar or different solutions will evolve or be repeatedly selected from standing genetic variation in response to comparable selective forces. Three SNPs that were significant in all three tests were shared between two different triads. A SNP in fibrinogen gamma polypeptide and a SNP in peroxiredoxin 6 were 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 the 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 cytochrome P4501A (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 pathway (Hahn 1998). In all three Superfund populations, CYP1A is refractory to induction by prototypical inducers (Elskus et al. 1999; Nacci et al.

1999; Bello et al. 2001; Meyer and Di Giulio 2002), and this trait is associated with resistance to PAH, PCB, and dioxin toxicity (Nacci et al. 1999; Bello et al. 2001). 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, nonsynonymous 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 related to rapid adaptation in recent inhospitable environments (highly polluted Superfund sites). The conservative analysis suggests that at a minimum, 1.4–2.5% of loci and 1.1–2.0% of the mRNA encoding loci respond rapidly to change in the environment. SNPs from mRNA encoding loci were randomly chosen and identified from ESTs. If we assume 30,000 mRNA encoding genes, our results suggest that 330–600 of these loci have adaptively diverged in the last 50 years in each Superfund population. Relative to Haldane's expectation concerning fixation of adaptive mutations (Haldane 1957), this is a large number of loci affected by natural selection over a short time and likely represents selection acting on standing genetic variation.

Conclusions

We identified nonneutral patterns of SNP divergence in highly polluted populations. Relative to the clean reference populations, the polluted Superfund populations have low to moderate isolation and show no signs of bottleneck. In these populations, the SNPs we identified are different between polluted and both reference populations but not different between the two, clean reference populations indicating that these changes are adaptive. Thus, we conclude that at a minimum, between 1.4–2.5% of loci and 1.1–2.0% of the mRNA encoding loci are responsible for the rapid adaptive divergence to anthropogenic pollution. If this is a representative sample and there are 30,000 genes, this represents >300 loci.

Supplementary Materials

Supplementary tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org).

Acknowledgment

The Authors thank Douglas L. Crawford for helpful discussions and review of the manuscript. Partial funding for this work was received from the National Institutes of Health (5 RO1 ES011588 and ES007046) and the National Science Foundation Division of Ocean Sciences (1008542).

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.

Ayroles JF, Carbone MA, Stone EA, et al. (11 co-authors). 2009. Systems genetics of complex traits in Drosophila melanogaster. *Nat Genet*. 41:299–307.

Beaumont M, Nichols R. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond Biol Sci*. 263:1619–1626.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol*. 5:e310.

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.

Black D, Gutjahr-Gobell R, Pruell R, Bergen B, Mills L, McElroy A. 1998. Reproduction and polychlorinated biphenyls in Fundulus heteroclitus (Linnaeus) from New Bedford Harbor, Massachusetts. *Environ Toxicol Chem*. 17:1405–1414.

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

Burnett KG, Bain LJ, Baldwin WS, et al. (26 co-authors). 2007. Fundulus as the premier teleost model in environmental biology: opportunities for new insights using genomics. *Comp Biochem Physiol D-Genomics Proteomics* 2:257–286.

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? *Aquat Toxicol*. 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. *Evol Bioinform Online*. 1:47–50.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*. 12:921–927.

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

Grahame JW, Wilding CS, Butlin RK. 2006. Adaptation to a steep environmental gradient and an associated barrier to gene exchange in Littorina saxatilis. *Evolution* 60:268–278.

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

Hahn MW. 2008. Toward a selection theory of molecular evolution. *Evolution* 62:255–265.

Haldane JBS. 1957. The cost of natural selection. *J Genet*. 55:511–524.

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, Ramos EM, Mehta JP, Collins FS, Manolio TA. 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.

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.

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.

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. *Toxicol Sci*. 68:69–81.

Nacci D, Coiro L, Champlin D, Jayaraman S, McKinney R, Gleason TR, Munns WR, Specker JL, Cooper KR. 1999. Adaptations of wild populations of the estuarine fish Fundulus heteroclitus to persistent environmental contaminants. *Mar Biol*. 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.

Nacci DE, Champlin D, Jayaraman S. 2010. Adaptation of the estuarine fish Fundulus heteroclitus (Atlantic killifish) to polychlorinated biphenyls (PCBs). *Estuaries and Coasts*. 33: 853–864.

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. *Mol Ecol*. 17:3599–3613.

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.

Oleksiak MF. 2010. Genomic approaches with natural fish populations. *J Fish Biol*. 76:1067–1093.

Ownby DR, Newman MC, Mulvey M, Vogelbein WK, Unger MA, Arzayus LF. 2002. Fish (Fundulus heteroclitus) populations with different exposure histories differ in tolerance of creosote-contaminated sediments. *Environ Toxicol Chem*. 21:1897–1902.

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. 2010. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet*. 42:570–575.

Paschall JE, Oleksiak MF, VanWye JD, Roach JL, Whitehead JA, Wyckoff GJ, Kolell KJ, Crawford DL. 2004. FunnyBase: a systems level functional annotation of Fundulus ESTs for the analysis of gene expression. *BMC Genomics*. 5:96.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38:904–909.

Slatkin M, Excoffier L. 1996. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity* 76(Pt 4):377–383.

Stinchcombe JR, Hoekstra HE. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100:158–170.

Storz JF. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol*. 14:671–688.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nat Genet*. 42:260–263.

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM. 1997. Human domination of earth's ecosystems. *Science*. 277:494–499.

Vogelbein WK, Fournie JW, Van Veld PA, Huggett RJ. 1990. Hepatic neoplasms in the mummichog Fundulus heteroclitus from a creosote-contaminated site. *Cancer Res*. 50:5978–5986.

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. *Hum Ecol Risk Assess*. 8:933–953.

Westfall PH, Young SS. 1993. Resampling-based multiple testing: examples and methods for P-value adjustment. New York: Wiley.

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 Genet*. 11:32.

Williams LM, Oleksiak MF. 2008. Signatures of selection in natural populations adapted to chronic pollution. *BMC Evol Biol*. 8:282.

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