Project Focus

Discovering the genomic basis of adaptive phenotypic variation is a major research ambition and is important for understanding fundamental mechanisms of the evolutionary process. Evolved pollution tolerance in the killifish *Fundulus heteroclitus* is a compelling model system for evolutionary genomics for three primary reasons. First, the adaptive phenotype is dramatic, as individuals from tolerant populations are orders of magnitude more tolerant to extreme environmental stress than individuals from nearby reference populations and from other species. Second, this adaptive phenotype has evolved over an extraordinarily short time period (dozens of generations). Third, this dramatic tolerance has evolved independently at least four times in *F. heteroclitus*, representing multiple adaptive convergences. **We propose to discover the genomic basis of this extraordinary adaptive phenotype by anchoring multiple lines of genome-scale evidence** (QTL mapping, population genome scans, comparative transcriptome profiling) **to a complete genome sequence**. This will be accomplished through three specific aims. The first aim is to sequence, assemble, and annotate a reference genome sequence for *F. heteroclitus*. The second aim is to map multiple lines of experimental genome-scale evidence to the reference genome to identify the number and physical location of loci associated with the evolved tolerance phenotype. The third aim will further buttress and validate candidate gene identification by characterizing patterns of polymorphism in population genomic samples. **These data will be used to discover candidate genes and test our overarching hypotheses that adaptive convergence emerged from selective sweeps of a small set of ancestral polymorphisms, and that the same genes account for the adaptive phenotype among independently derived populations.**

Intellectual Merit

Uncovering the genomic basis of adaptive traits lies at the heart of evolutionary research. We seek to address important questions including "When phenotypes converge, are similar genetic changes responsible or are disparate genetic mechanisms involved?", "Did adaptation require new mutations following environmental change, or did selection just sort among pre-existing polymorphisms?", and "Are adaptive phenotypes underpinned by protein polymorphisms or by polymorphisms in *cis* regulatory regions?". The case of extreme evolved pollution tolerance in *F. heteroclitus* provides a wonderful research opportunity because the ecological significance of the evolved phenotype is clear, we have both structural (mapping) and regulatory (transcriptomic) data, and this phenotype has evolved multiple times independently. With a genome sequence, we are poised to **test cutting-edge questions related to the repeatability of evolutionary change, the role of protein *versus* regulatory variation underlying adaptation, and the relative roles of selection on standing genetic variation *versus* selection on *de novo* mutation underlying evolutionary innovation.**

Broader Impacts

As part of this project we will host a genome annotation workshop and establish a collaboration wiki, which will be **excellent training opportunities for students and post-doctoral researchers** in the analytical tools and approaches that lie at the forefront of bioinformatics and comparative genomics research. This workshop and wiki will also promote cross-disciplinary collaborations since the diverse PIs in the *Fundulus* Genomics Consortium will be brought together in a mutually engaging and intense research setting. One major product from this project will be a robust annotated genome assembly for *Fundulus heteroclitus*. Other fish for which genomes exist are excellent models for biomedical science (zebrafish, Japanese medaka), vertebrate genome evolution (pufferfish, lamprey), and morphological evolution (stickleback, cichlid), but no other teleost model exhibits as wide a range of beneficial characteristics as *F. heteroclitus*. By virtue of their diverse ecological distributions, highly plastic physiologies, and importance as a model for a large community of researchers in diverse disciplines, the *F. heteroclitus* genome sequence will **enable and accelerate research in environmental genomics, ecology, comparative physiology and biochemistry, and evolutionary biology, thereby transforming the research landscape in these fields.**

A. SPECIFIC AIMS

A.1. Project Focus

Fundulus heteroclitus is a powerful established model in physiology, ecology, toxicology, and evolutionary biology, and is an emerging model for environmental genomics research [1]. This species is resilient to many environmental stressors, but of particular note are multiple populations that have independently derived locally adaptive tolerance to diverse but mechanistically-related pollutants. Our focus is to identify the genomic basis of this dramatic, rapid, convergent evolutionary adaptation to extreme environmental stress. There are few other vertebrate systems that can offer insight into mechanisms of adaptive population divergence in response to rapid environmental change; these types of change may be particularly relevant given the pace of current human-induced environmental change.

A.2. Research Objectives and Specific Aims

Our overarching goal is to identify the genomic architecture and candidate genes underpinning adaptive convergent phenotypes, and to test hypotheses regarding the evolutionary patterns of adaptive convergence among independently adapted populations. This will be accomplished through three specific aims. The first aim is to generate a deep-coverage reference genome sequence for *F. heteroclitus*. The second aim is to map loci from multiple lines of experimental genome-scale evidence to the genome to identify genes involved in adaptation. The third aim will further buttress and validate candidate gene identification by screening patterns of polymorphism in population samples. Aims two and three will serve to **test the hypotheses that adaptive convergence emerged from selective sweeps of a small set of ancestral polymorphisms, and that the same genes account for the adaptive phenotype among independently derived populations.**

Aim 1: Sequence, assemble, and annotate a reference genome sequence. To achieve this, we will use second-generation sequencing, a novel assembly strategy with proof-of-principle in place, and a proven community-based annotation strategy which our team successfully applied to three arthropod projects.

Aim 2: Map experimental data to the reference genome to identify candidate genes involved in adaptation. To achieve this, we will map three lines of genome-scale experimental evidence to the reference genome. Data already in-hand or currently being generated include 1) QTL mapping, 2) population genome scan, and 3) comparative transcriptomics data. Once mapped, these data will be used to identify candidate regions (especially where multiple lines of evidence converge on specific genomic regions) and will allow testing of hypotheses.

Aim 3: Population re-sequencing. Pooled DNA from 30 individuals from each of four tolerant and sensitive populations will be sequenced. These sequences will provide allele frequency estimates to identify amino acid substitutions within and among populations, enable population-genomic analyses to identify regions of genomes under recent positive selection, and importantly, identify regions of genomes exhibiting signatures of recent positive selection that are shared among converged populations.

A.3. Expected Significance

With genome sequence linked to experimental genome-scale data (QTL, transcriptomics, nucleotide divergences among populations), we are poised to test questions related to the repeatability of evolutionary change, the role of protein *versus* regulatory variation underlying adaptation, and the relative roles of selection on standing genetic variation *versus* selection on *de novo* mutation. Importantly, products from this project will serve as a critical resource for accelerating the next generation of research in environmental genomics, ecology, comparative physiology and biochemistry, and evolutionary biology, and will train students and post-doctoral fellows in genomic, evolutionary, and bioinformatic sciences.

B. RESPONSE to COMMENTS from PREVIOUS PANEL

This proposal was reviewed in spring 2009, 2010, and Fall 2011. The proposal scored very well (E,E,E,VG,G from 2009; E, E,E/VG,E/VG,VG,P from 2010, E,E,E,E,E/VG,VG from 2011). The most recent panel agreed that we had effectively addressed prior criticisms: "*The previous criticisms involved a general lack of cohesion regarding the plans to integrate additional data into the genome sequence. The proposal has been revised to more explicitly describe the nature of the additional data and its relevance*."

The panel was without criticism of our broader impacts, and was also very enthusiastic about the intellectual merit of our proposal: "*With regard to Intellectual Merit, the strengths of the proposal remain the same, and they remain compelling. This is an excellent group of scientists, working on an important question and using an important model system. The sequencing of the Fundulus genome will be very valuable. The enthusiasm for this work among the panelists was very high*." The primary concern was that our timeline for sequencing and assembly was too ambitious. We address this criticism in section G.1.1 by detailing how our 6-month timeline for sequencing and assembly is conservative and feasible by outlining our hands-on experience with proposed methods. Our annotation strategy is proven (fruitfly, mouse, sea urchin, skate, etc.) and successful in our hands [2-4]. We strengthen an already strong team with the addition of John Colbourne (*Daphnia* genome project lead) as a co-PI, and by increasing the effort of Don Gilbert who helped develop Flybase [5] and wFleabase [6]. We also have formed a collaboration with the Mount Desert Island Biological Laboratory (MDIBL, see attached letter of collaboration from Ben King and Kevin Strange) to use their facilities to host the annotation workshop. The Indiana CGB group will still lead the annotation and development and implementation of bioinformatics tools, but inclusion of MDIBL as the workshop location will enhance workshop participation, broaden our research community, and broaden our impact on Marine Genomics research. In addition, it capitalizes on their existing infrastructure, which has a track record for hosting workshops and conferences and draws on their specific expertise in hosting genome annotation workshops.

C. RESULTS from PREVIOUS NSF SUPPORT

Federal support among the PIs involved in this application have produced over 54 publications that have examined gene sequence, population genetics, gene expression and physiology of *F. heteroclitus* in the last two decades [7-22] [23-60]. An award to AW (0652006) is supporting comparative genomics studies in tolerant and sensitive populations of *F. heteroclitus*, including genotyping of QTL mapping families and comparative transcriptomics studies. This work is in close collaboration with Dr. Diane Nacci (US EPA), whose lab has performed experimental crosses and dose response experiments, houses multiple generations of sensitive and tolerant populations, and has genetic samples from all experimental populations. Much of this work is ongoing, and described in section G. This grant has supported seven publications [61-67], helped train nine undergraduate students including two minority students, and data served as a teaching device for a graduate course "Ecological Genomics" taught by AW.

An award to JS as Co-PI (0221837) supported a project that explored the evolution of metal tolerance in *Daphnia pulex.* It included the direct involvement of seven senior personnel, two postdoctoral fellows, two graduate students, three technicians, and a programmer. It created genomic tools for *D. pulex*, constructed a database to disseminate the sequence information (wFleabase; http://wfleabase.org/), developed microarrays, helped establish the *Daphnia* Genomics Consortium (http://daphnia.cgb.indiana.edu/), contributed to the sequencing of the first crustacean genome, and contributed the formation of a new model system (http://www.nih.gov/science/models/). Outreach activities included training workshops, web-based conferences, and collaboration wikis to educate the community on utilizing these resources. This grant has resulted in over 15 publications of which at least six are of specific interest to this proposal [6, 68-72], and it played a significant role in attracting a special issue of *Science* scheduled for release on February 4th that will launch the *Daphnia* model system.

Two awards to DC provided the resources to initiate functional genomic analysis of *F. heteroclitus*. The first (9986602) supported the creation and annotation of 74,000 ESTs. The second (0308777) supported the use of the microarrays produced from these ESTs and was a multi-collaborator grant. This research produced twenty-four publications [31, 40, 58, 59, 73-92], many in top-tier journals, and four are being revised for submission [93-96]. These grants also supported training for three post-doctoral fellows, four Ph.D. students (two minority), one Masters student, and six undergraduates (five minority).

An ongoing RAPID to DC (1048208), AW (co-PI), and MO (co-PI) examining populations genomics of the Gulf killifish in response to the BP oil spill is helping to support two graduate students as is an ongoing EAGER to MFO (1008542), which has supported two publications to date [97, 98].

D. PRELIMINARY DATA

Preliminary data that will assist with genome assembly and annotation and be used to interrogate the final genome sequence are discussed in more detail throughout this proposal. More than 74,000 EST sequences, coupled with robust bioinformatics [80] and annotation algorithms designed specifically for *F.*

heteroclitus [99], are publicly available to assist with genome assembly, gene finding, and annotation. Among the PIs involved in this application, over 54 publications in the last two decades utilized *F. heteroclitus*. [7-22] [23-60]. Much of this research examined evolved differences among populations or taxa [7, 8, 10, 13, 14, 18, 19, 24, 26, 27, 31, 34, 35, 37, 40, 44-48, 52, 54, 56]. These evolutionary analyses have depended on the more detailed analyses of genes, gene promoters, gene expression and physiology [11, 12, 15-17, 20-22, 25, 27-30, 32, 35, 36, 38, 39, 41-43, 49-51, 53, 55, 57, 58, 60, 100].

The challenge for many genomes is the bioinformatics and annotation. The Indiana University Center for Genomics and Bioinformatics (CGB) offers extensive experience in developing the sophisticated infrastructure necessary to enable cutting edge genomics research in new model species (see sections G.1.3, G.1.4). These efforts will be enhanced through a collaboration wiki, web-based conferences, and an annotation training workshop where multiple investigators, post-doctoral fellows, graduate and undergraduate students receive training and manually curate computational annotations.

New sequencing by our group at the Washington University Genome Center (WUGC) are serving as proof-of-principle for using all next-generation sequencing for *de novo* assembly of complex eukaryotic genomes. We have produced high quality *de novo* assemblies for the chicken genome, a similar sized genome to *F. heteroclitus*, using all short-read Illumina sequence (unpublished data).

E. BACKGROUND and SIGNIFICANCE

Population responses following dramatic environmental changes could span the continuum from extinction to evolutionary adaptation. Some species will suffer severe fitness effects and will be unable to compete and persist, whereas others may have the inherent physiological capacity to tolerate the stress. In yet other species, the necessary adaptive genetic variation may increase in frequency to allow previously sensitive populations to evolve tolerance. Uncovering the genomic basis of susceptibility and tolerance to environmental stressors is fundamental to understanding the evolutionary process.

Discovering the genomic underpinnings of adaptive variation is a major goal of evolutionary research [101-105]. However, the associated challenges are not trivial. Of critical importance for top-down discovery-based approaches are genetic and physical maps. A complete genome sequence is the ultimate physical map. The recent emergence of high-throughput technologies is enabling large-scale sequencing initiatives for species of ecological importance with compelling natural histories such as *F. heteroclitus*. Importantly, the genome sequence data generated in Aim 1 provides a map for the loci exhibiting adaptive patterns of variation among populations; these loci include genes with non-neutral patterns of expression divergence among populations (data described in Aim 2), markers that segregate with tolerance in mapping families (data described in Aim 2), markers that are outliers in population genetic studies (data described in Aim 2), or genome regions exhibiting signatures of recent selective sweeps (data to be collected in Aim 3). Identifying the genomic region surrounding loci with non-neutral patterns of variation will enable us to focus our inquiries, for example by testing whether duplications are associated with gene expression divergence (e.g., [106]), whether groups of genes with similar expression divergence share transcription factor binding sites, whether acquisition or loss of specific transcription factor binding sites (such as those that mediate PCB toxicity) accompanies adaptive gene expression divergence among populations (e.g. [107]), or whether QTL or F_{ST} outlier loci bound genes with known functions that associate them with our adaptive phenotype (toxicity tolerance).

E.1. The genetic basis of adaptive variation

Identifying the genomic elements that underlie adaptive traits is a major research ambition. Although it remains a significant challenge, technological and methodological advances in recent years are accelerating the pace of discovery. Genetic changes that affect phenotype can either reside in protein coding regions affecting protein function or in non-coding gene regulatory regions affecting when, where, and how much a protein is expressed.

Though there is active debate over the relative roles of structural *versus* regulatory variation underlying adaptive traits, especially morphological traits (for example, compare contrasting views between [108] and [103]), it is clear that both types of variation can be important. For example, protein variation accounts for pigmentation variation in mice [109], birds [110], and lizards [111], cold tolerance in antarctic fishes [112], resistance to newt toxin in snakes [113], and hypoxia tolerance in mammals and birds [114, 115]. In contrast, regulatory variation plays a role in eye degeneration in cavefish [116], developmental variation in beak morphology in finches [117, 118], and pigmentation variation in fruit flies

[119, 120]. Notably, both regulatory and structural variations explain adaptive patterns of enzyme divergence among clinally distributed populations of *F. heteroclitus* (see review in [1]).

Because regulatory and structural variation may underlie the environmental pollution tolerance in *F. heteroclitus*, we have collectively accumulated evidence for both structural (SNP analyses, sequence polymorphism; Fig 1) and regulatory (transcriptomics, transcriptional factor binding domains) differences. These data will be used to interrogate the proposed genome sequence to identify the number of loci associated with tolerance, similarity of implicated genomic regions across converged populations, and association between mapping markers and genes with divergent gene expression patterns (Section G.2).

E.2. The genetic basis of convergent evolution

Convergent evolution occurs when the same phenotype evolves independently in different lineages to solve similar evolutionary challenges. The genetic changes that lead to the adaptive phenotype may involve the same genes across convergent lineages or may involve fundamentally different genes or biochemical pathways. Examples of both processes (changes in the same or in different genes) are evident in the literature. For example, mutations in the same genes underlie convergent evolution of armor plating in stickleback fish [121, 122], loss of pelvic structures in multiple stickleback species and also in manatees [123, 124], and adaptive coat color variation in beach mice [109] and other animals including lizards, birds, cats, and mammoth (see review in [101]). Intriguingly, in the case of beach mice, the *Mc1r* gene is implicated for light coat coloration in some populations but different genes are responsible in others [109]. Similarly, pigment loss and eye loss have convergently evolved through different genetic mechanisms in Mexican cavefish populations [125] as has pigment gain in rock mouse populations [126]. Clearly, both similar and different genetic mechanisms can underlie convergent phenotypes, whether or not convergent lineages are closely or distantly related [101].

The genetic variation upon which selection acts can either pre-exist in the ancestral population (as standing genetic variation) or may arise as new mutations (*de novo* mutations) in the novel environment. Selection on standing genetic variation is more likely to enable rapid evolutionary change (compared to scenarios dependent on new mutations) in part because the necessary variation is immediately available and exists starting at a higher frequency [127]. Because the tolerance phenotype has evolved so dramatically, quickly, and repeatedly in *F. heteroclitus*, we predict that selection has acted on standing variation that exists at low frequency in ancestral populations, and that similar genomic regions underlie adaptive variation in convergent lineages. If this is the case, it is possible that different mapping markers may be implicated in different tolerant population comparisons since marker variants may have been sorting differently in each tolerant population's ancestral population. A genome sequence is therefore critically necessary for mapping data from multiple lines of evidence (genotyping of QTL mapping families, genotyping of natural populations, comparative transcriptomics), and from multiple convergent population samples, to identify the genomic underpinnings of this dramatic adaptive trait.

E.3. Pollution tolerance in Fundulus

Evolved pollution tolerance in *Fundulus heteroclitus* is a compelling model for evolutionary genomics for three primary reasons. 1) The adaptive phenotype is dramatic; individuals from tolerant populations are more than three orders of magnitude more tolerant to dioxin-like compounds (DLCs) compared to other species. Yet, *F. heteroclitus* in unpolluted waters is relatively sensitive among fish species to this stress [128, 129]. In developing embryos DLCs specifically compromise cardiovascular system development [130], and these developmental effects consistently emerge at approximately three orders of magnitude lower concentrations in sensitive compared to tolerant populations [131]. 2) This adaptive phenotype has evolved over a very short time period. The polluted sites where tolerant populations reside have been contaminated with tolerance-associated chemicals (primarily polychlorinated biphenyls (PCBs) and polycyclic aromatic hydrocarbons (PAHs)) in the last fifty years [132]. 3) This extraordinary tolerance in *F. heteroclitus* has evolved independently at least four times [1, 133, 134].

These independently derived tolerant *F. heteroclitus* populations provide a powerful comparative system for determining the genomic basis of rapid adaptation, the repeatability of evolutionary change [101] and the relative roles of selection on standing genetic variation *versus* selection on *de novo* mutation [135]. We predict that pollution tolerance in *F. heteroclitus* has evolved by repeated fixation of standing variation because tolerance has evolved so quickly and repeatedly within the species. If this is supported by our data, then this has important conservation implications, insofar as species resilience to rapid environmental change may depend on the availability of sufficient standing genetic variation in

natural populations. Few other vertebrate systems offer insight into mechanisms of adaptive population divergence in response to rapid environmental change. These changes are relevant given the current pace of human-driven environmental change.

Genetic, biochemical, and physiological differences between tolerant and sensitive populations have been carefully studied over the past two decades, yet the genomic basis of the tolerant phenotype remains elusive (reviewed in [129, 132]). Nevertheless, we highlight six key features of the resistant populations that have been identified, demonstrating the potential for new insights to be facilitated by genomic studies. 1) Pollutant tolerance involves a variety of endpoints, including embryo and adult survival, teratogenicity, and altered gene expression [129]. 2) Each of the populations shows cross-tolerance to classes of compounds not abundant at the site. For example, the dioxin-resistant Newark fish are also resistant to PCBs [136]. PCB-resistant New Bedford fish are also resistant to PAHs [137, 138]. PAH-resistant Elizabeth River fish are also resistant to PCBs [139]. This cross-tolerance suggests that the mechanisms of resistance may converge on a common biochemical pathway, such as the aryl-hydrocarbon receptor pathway. 3) The tolerance appears to involve both heritable and non-heritable mechanisms [129, 137, 139-141]. 4) The tolerant populations do not show an overall loss of genetic diversity, as assessed by allozyme and molecular methods [142-145]. However, the tolerant populations are genetically distinct from nearby sensitive populations and do show evidence for selection at specific loci [144, 146-148]. 5) Tolerance evolves not just in highly contaminated sites, but at moderately contaminated sites as well, and the degree of tolerance is related to the degree of contamination [128]. Thus, pollutant tolerance is a widespread phenomenon. 6) Although the fitness costs or “trade-offs” of pollutant tolerance in *F. heteroclitus* are not yet well understood, there is evidence that such costs accompany some pollutant tolerant phenotypes [129, 149] (but see [150]).

A genome sequence will enable study of the regulatory and structural polymorphisms underlying the resistant phenotypes and may also offer insights into the mechanistic basis of fitness trade-offs following adaptation. A sequenced genome will facilitate the identification of duplicated loci and the cis-regulatory architecture of co-expressed genes, and will serve as a scaffold for mapping the physical location of divergent loci (Aim 2) and for mapping population genome scan data (Aim 3).

E.4. State of readiness for genome sequencing

Based on flow cytometry and bulk fluorometric analysis, the *F. heteroclitus* haploid genome size is 1.29 to 1.5 billion bp [151-153] which is intermediate in size between Japanese medaka and zebrafish [154]. We will sequence one male (heterogametic) individual from a stock of fish that has been inbred for ten generations to minimize heterozygosity and maximize the efficiency of reference genome assembly. Over 1.5 million EST sequences (77,000 in NCBI, remainder from Joe Shaw, unpublished data), combined with >100,000 contigs recently derived from two embryo shotgun libraries (MH, unpublished data), will aid with assembly and gene-finding (see sections G.1.3, G.1.4). A novel and highly robust gene annotation pipeline developed for *F. heteroclitus* EST collections [99] will aid in annotation of genes identified by protein and EST homology (see section G.1.4). The bioinformatics infrastructure for assembly validation and community-based manual curation of annotations has been successfully implemented for water flea, jewel wasp, and aphid genome projects and will be adapted for the *F. heteroclitus* genome [2-4]. Ongoing projects in the labs of AW, P. Schulte (UBC; see letter of collaboration), and M. Bagley (US EPA; see letter of collaboration) are generating genetic maps for *F. heteroclitus* using microsatellite, SNP, and RAD markers. Collectively, these markers and maps will be used for scaffold assembly (see section G.1.3). Finally, gene knockdown technologies have been successfully adapted to *F. heteroclitus* [155], and may enable more detailed study of adaptive variants in future studies.

F. BROADER IMPACTS

Useful and accurate annotation remains a significant challenge for any new genome, and the Indiana University CGB has established an efficient community-based platform that has proven effective for final annotation of the *Daphnia*, *Nasonia*, and aphid genomes [2-4]. For this final stage of genome annotation, the CGB will host a collaboration wiki, web-based conferences, and partner with the MDIBL to provide an annotation training workshop for members of the *Fundulus* Genomics Consortium (FGC). The partnership with the MDIBL draws on their long history of work with marine species, including *Fundulus*, which will expand the research community and ensure maximum participation in the annotation effort. These will set the stage for open, community-based manual curation of gene identities and functionalities,

focusing on those genes and gene families of particular interest to members of the FGC (see section F.5). Importantly, **this workshop will serve as an excellent training device for graduate students and post-doctoral fellows (from FGC member labs) in cutting-edge techniques at the forefront of bioinformatics and comparative genomics, and we will provide travel awards for students and post-doctoral fellows to maximize participation and training potential.** The workshop will be held at the Mount Desert Island Biological Laboratory (MDIBL; see letter of collaboration from Ben King); It will bring together PIs from the diverse *Fundulus* research community, and foster the kinds of cross-disciplinary collaborations that are most likely to emerge from communication-intensive, information-rich, community-based interactions. Skills in data mining, bioinformatics, and computational comparative biology, are among the most important for training the next generation of evolutionary, comparative, and genome biologists. We seek funding to support student and post-doctoral fellow training in these skills through travel support to the annotation training workshop and subsequent participation in the community-based annotation collaboration. This funding will support participation of up to 40 students and post-doctoral fellows in this effort, drawn from physiology, toxicology, and evolutionary biology laboratories of consortium members. Beyond the intangible benefits of exposure to cutting-edge science, participants will gain specific skills needed to curate gene model structure, functional annotation, and phylogeny/membership within gene families. Participants will learn to find gene models of interest, add missing models to the gene catalog, assess the quality of existing gene models, and manually correct errors they find. They will directly contribute to annotating the reference list of gene models, which will be submitted to GenBank. Importantly, though this workshop is short and intensive and serves to initiate training, participants will continue to contribute to the project once they return to their home institution throughout the project period and beyond through the collaboration wiki (as has been the successful model for the *Daphnia* genome consortium). Intensive training in these skills may greatly enhance the quality of student training, provide insight into their current research projects, and position them well intellectually and competitively as they initiate and expand their research careers.

Among the 40 workshop participants, we seek to **include three undergraduate students** recruited through undergraduate research programs sponsored through LSU by the **Louisiana Alliance for Minority Participation** or the Howard Hughes Medical Institute. These programs have a long history at LSU of engaging under-represented groups in science research activities. Additional funds are requested to support these three students during the academic year following the annotation training workshop to complete bioinformatics research projects associated with gene family annotation. Each student will be assigned a gene family based on their personal research interests or based on those families that are priorities for manual annotation. The three participants in this program will contribute papers in a research symposium at the end of the academic year. We also seek to include undergraduate researchers in the MDIBL Research Experience for Undergraduates Site (funded by NSF grant DBI-0453391). Workshop participation by these researchers from MDIBL's REU site, entitled *Research Experiences in Marine Molecular Physiology and Environmental Stress*, would make a highly relevant supplemental component to REU activities. A high quality genome sequence for *Fundulus heteroclitus* would **transform the research landscape for environmental, ecological, and evolutionary biology**. Other completed or ongoing fish genome sequences serve as excellent models for research in biomedical sciences (zebrafish, Japanese medaka), morphological evolution (sticklebacks, cichlids), and genome evolution (pufferfish). *F. heteroclitus*, by virtue of their diverse ecological distributions and highly plastic physiologies, will serve as an excellent model for the next generation of research in ecological, evolutionary, physiological, and environmental sciences. This is most strongly evidenced by the diversity of research programs represented by members of the FGC.

In addition to producing an important resource for the genomics community and training students and postdoctoral fellows, this project will contribute to our understanding of the genomic basis of evolved differences among taxa. Important questions related to the repeatability of evolutionary change, the role of protein *versus* regulatory variation underlying adaptation, and the relative roles of ancestral polymorphism *versus de novo* mutation underlying evolutionary innovation remain at the forefront of evolution research as highlighted by a number of recent reviews [101-105]. Adequate answers to these questions will only emerge once many case studies, representing diverse taxonomies and diverse adaptive contexts, are completed. This case of derived pollution tolerance in *F. heteroclitus* is compelling because the ecological relevance of the adaptive phenotype is clear and a dramatic level of tolerance has evolved so rapidly in so many different populations.

G. METHODS and MATERIALS

G.1. Specific Aim 1 – Genome sequencing, assembly, annotation

We propose to sequence one reference genome at 60X coverage using paired-end Illumina sequencing. Sequencing of small insert libraries (200-350 bases) will generate the bulk of shotgun sequence (50X coverage). Additional paired-end reads (10X) from mid (3kb) and large (40kb) insert libraries will contribute additional shotgun sequence, but more importantly will be used for mid and long-range contig assembly. The individual to be used for this reference genome will be selected from a stock of fish that have been inbred for ten generations to minimize heterozygosity. This stock is maintained by Dr. Richard Winn at the University of Georgia (see attached letter of collaboration).

G.1.1. DNA sequencing

Next-generation sequencing technologies have proven to be cost effective for whole genome assemblies, most recently with short reads [155-157]. The biggest challenge remains to assemble a majority of the genome in the largest contiguous blocks with high accuracy. Our strategy with the *F. heteroclitus* genome is to use the Illumina sequencing of paired reads with increasing insert size. Our current estimate includes 50x genome coverage using short (200-300 bp) paired reads plus 5x coverage with 3 kb and 8 kb each and 1x of 40 kb paired reads.

We expect that these contig lengths will be sufficient for gene predictions and post-assembly alignment-based analysis. From the recent panda whole genome *de novo* assembly study, contig and scaffold N50 values of 40 kb and 1.3 Mb, respectively, were achieved entirely from short Illumina reads [155]. Because the *F. heteroclitus* genome is smaller than the panda genome and contains fewer predicted repeats, we expect assembly contiguity to be comparable. Independently, the WU group has assembled chicken and pigeon genomes *de novo* using a similar strategy (unpublished data). For chicken, read pairs from 300 bp and 3 kb inserts assembled into N50 contigs and supercontigs of 12 and 314 kb, respectively. For pigeon, read pairs from 300 bp, 3 kb, and 8 kb inserts assembled into N50 contigs and supercontigs of 23 kb and 2,700 kb, respectively. By including paired end reads from 40 kb insert libraries we should meet or exceed these contig sizes. Libraries with 40 kb inserts (ditags) adapted to the Illumina HiSeq instrument are a relatively new option for long-range assembly contiguity, and we have experience building these libraries. In sequencing a chicken ditag library we observed the expected genomic distance between these reads by mapping to the chicken reference assembly with the CrossMatch aligner (unpublished data).

Although read base error rates for the Illumina technology are relatively low, read quality filtering is important for assembly quality. We have established a series of simple quality control filters, such as eliminating the reads of average low quality, trimming read ends of a defined number of bases, and eliminating reads with significant Ns. A draft assembly of the *F. heteroclitus* genome will be necessary and sufficient to achieve the goals of this proposed project and will be of greatest utility to the diverse *Fundulus* research community. It will also serve as an anchor for future comparative re-sequencing studies of individuals from related populations and species. In our experience, library preparation will require two weeks, the amount of sequencing we propose will require two weeks (seven Illumina flow cell lanes total), and assembly will require one month. This sums to two months, so our proposed six month timeline from isolating DNA to assembled genome is feasible and generous.

G.1.2. Genome assembly

The SOAP assembler, version 1.04, will be utilized to assemble all read types using an iterative process [155, 156]. The SOAP software resides on four 300 GB 10,000 RPM SAS hard drives, with eight 2.9GHz Quad-Core AMD Opteron Model 8389 processors, 512KB L1 Cache (32 processor cores total) and 512 GB of memory (consisting of 32 16 GB DDR2-667 ECC DIMM). Most all short read assemblers rely on the de Bruijin graphical structures, a directed graph that represents homogenous overlap between sequences (see review [158]). In brief, genome assembly will involve four principal steps that progress from forming contigs from raw sequence reads, to connecting contigs into scaffolds using paired-end sequence of large fragments, to gap filling and finally error correction. A base of smaller contigs will serve as anchor points for an iterative adding of longer range insert sizes serving to build scaffold length. Gaps that exist in the scaffolds can be filled in most cases by the use of all reads. The SOAP tool has achieved the contiguity necessary for downstream analysis, but we will consider other assemblers as they are developed, such as the recently-released ALLPATHS-LG program [159].

G.1.3. Genome assembly curation

Despite improvements in assembly algorithms, assembling genomes from millions of small sequence reads is susceptible to error. We will assess the accuracy of the assembled *F. heteroclitus* genome using methods developed by the WUGC and the Indiana University CGB. The assembly will be assessed using a machine learning approach that compares several *in silico* measures: read coverage, compression and extension statistics for unsatisfied mate-pairs, the ratio of good and bad fragments, and the maximum of absolute values of average positive and negative Z-scores. Completeness and accuracy will also be assessed using several quality metrics: read chaff rate, read depth of coverage, average quality values per contig, discordant read pairs, gene footprint coverage (as assessed by cDNA contigs) and comparative alignments to the most closely related fish, Japanese medaka. Each of these metrics reveals something unique about the assembly and defines the strengths and weaknesses of an assembly. The WUGC group has years of experience at evaluating assembly quality, and the Indiana University CGB has successfully applied these methods for the *Daphnia pulex* genome [160]. Collectively both group's assembly analyses will ensure a thorough review of quality.

To organize sequence information along chromosome boundaries we will make use of genetic linkage maps. Markers currently being used for the generation of a *F. heteroclitus* genetic map include restriction site associated DNA (RAD) sequences generated by Illumina sequencing [161] in the Whitehead laboratory from families created by collaborator Dr. Diane Nacci (US EPA), and collaborators have also generated hundreds of polymorphic microsatellite markers for genetic mapping (Dr. Mark Bagley, US EPA, see attached letter of collaboration). Following assembly, RAD tag and microsatellite primer sequences, coupled with genetic mapping data, will assist in large scaffold construction and improve order inconsistencies. Though interchromosomal rearrangements are common among species, we can take advantage of the five available fish genomes to screen alignments for regions of conserved gene order and orientation. This will help with final scaffold assembly.

A final check on assembly quality will be to screen for unexpected sources of sequence contamination. These sources typically are microbes and eukaryotic parasites. The WUGC will remove any contigs that meet our contamination criteria, which is a mega-BLAST sequence match (>97% identity) to a non-target species. The screened genome assembly will be submitted to the WGS division of Genbank for an independent contamination analysis. The final assembly will be posted on the Ensembl, the University of California Santa Cruz, and the NCBI genome browsers for public queries.

G.1.4. Gene finding, annotation & bioinformatics

First-pass gene prediction will use a modified Ensembl pipeline [162] for evidence-supported gene model building and model merging. Uniprot protein sequences from *F. heteroclitus*, *Oryzias latipes*, *Tetraodon nigroviridis*, and *Danio rerio* will be used sequentially as seeds for coding sequence prediction. In addition, cDNA sequences from *Fundulus heteroclitus* will be aligned and used to find genes and add UTR information. A portion of Ensembl's mandate is to work directly with genome sequencing projects and use custom-curated data sets (such as EST sequences and specific Uniprot data sets) to enable annotation, at no additional cost to the sequencing projects (see attached letter of collaboration from Paul Flicek, team leader for vertebrate genomics at EMBL-EBI).

Additional gene models will be predicted and improved at the CGB using in-house pipelines that include Fgenesh family models [163], Genewise family models [164] and SNAP [165]. Colleagues at the NCBI RefSeq Project Group will provide RefSeq transcript alignments [166] and Gnomon gene prediction (http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml). Finally, ESTs will be used to extend predicted gene models into fuller-length genes by adding to 5' and/or 3' UTRs. The PASA annotation pipeline [167] will be used to further refine the gene models by verifying that spliced alignments of ESTs are congruent with the predicted gene structures. The elected gene set will be given putative functional assignments by homology to annotated genes from NCBI non-redundant sets and classified according to Gene Ontology [168], eukaryotic orthologous groups [169], and KEGG metabolic pathways [170]. Correct gene annotation will be aided by performing sequential homology searches against multiple databases to verify similar BLAST hits, and integrating protein motif searches using machine learning techniques [99]. This algorithm has been tested for *F. heteroclitus* EST annotation; results indicate that if a sequence matches two species with similar annotation, the annotation is correct 99% of the time [99].

One of the most challenging elements of a genome sequencing project is functional annotation [171, 172]. To aid in this task, Indiana University's CGB will host an annotation training workshop with supporting web-conferencing, and design and implement a community-wide manual annotation project

modeled from earlier successful experiences with the waterflea *Daphnia* [2], jewel wasp [4] and pea aphid [3] genome projects, which are coordinated through collaboration wikis. The *F. heteroclitus* annotation project will employ a hybrid "jamboree" and "cottage industry" model that will bring together the *Fundulus* community with bioinformaticians and genome biologists in a five-day intensive annotation workshop that will serve to train the community and jump-start a longer-term decentralized annotation effort that will be dispersed among the community [171, 172].

The annotation workshop will be held at the MDIBL in Maine, and combined with a FGC meeting to ensure broad participation from the community. Participants will work hands-on with annotation training modules and will hear from invitees from other genome projects who will highlight the annotation process and speak from experience on ways to improve our annotation efforts. The goal is to educate and excite the community about their role in building this resource, familiarize them with the annotation software and other technical aspects, and facilitate future collaborative research efforts. The workshop will be held during the first academic break (summer or winter) following genome assembly and Ensembl annotation and will be open to all interested researchers. We will provide travel awards for students and post-doctoral researchers (see Broader Impacts, section F).

This dispersed annotation approach will initially focus on genes of interest to the community and offers the advantage of engaging the expertise of *Fundulus* biologists who will assign or validate putative function to predicted gene models in their own laboratories through experimentation (e.g., adding sequence, recording expression patterns). The major disadvantages of this approach are the potential for duplicated efforts and dissimilarities in reporting standards and data presentation, which are overcome through community organization, communications support, and data management tools [171, 173].

For our project, a *Fundulus* genome annotation steering committee will be formed to organize the community annotation project. The steering committee will adopt strict annotation guidelines that will document the manual annotation and curation of gene models and ensure common reporting standards and data presentation among research groups. The steering committee will also help coordinate the formation of annotation groups centered on themes of biological, ecological, physiological, and toxicological interests within the *Fundulus* community. Each annotation group will be facilitated by a group leader appointed from the steering committee. Progress and findings will be communicated via web-conferences and the collaboration wiki – the communication hub for the community.

The *F. heteroclitus* genome database, which will be housed in the Indiana University CGB, will be built with common Generic Model Organism Database (GMOD; [174]) components and open source software shared with other genome databases. This custom database is important for enabling continual manual annotation updating by FGC members. Genome maps will include homologies to other eukaryote proteomes, marker genes, microsatellite and EST locations, and gene predictions. The assemblies and predicted genes will be searchable by BLAST and linked to genome maps, and will also be mirrored on the Ensembl and the University of California Santa Cruz genome browsers. The computationally intense analysis we propose will benefit from the TeraGrid project (www.teragrid.org). We have used TeraGrid to annotate and validate the assembly of a *Daphnia* genome, where results included homologies to nine eukaryote proteomes, gene predictions, marker genes, and EST locations.

G.2. Specific Aim 2 – mapping experimental data to reference genome for candidate gene discovery

Adaptive variation may be underpinned by mutations in either structural or regulatory regions, yet the relative adaptive importance of variation in these regions is a matter of debate. As highlighted in Ellegren and Sheldon [102], some maintain that regulatory elements are most likely to underpin new phenotypes [108], whereas others contend that protein variation accounts for much of adaptive variation [103]. Since there is no *a priori* reason to exclude the influence of either structural or regulatory variation in extremely stress tolerant phenotypes in *F. heteroclitus*, we have surveyed both types of variation in comparative experiments. We will interrogate the genome sequence with genes and markers primarily derived from four sources in order to address our scientific goals. These sources include: 1) Markers from QTL mapping families currently created for two different tolerant populations (G.2.1), 2) Markers from population genome scans including multiple tolerant and sensitive populations (G.2.2), 3) Genes from comparative transcriptome profiling that are associated with differences in pollution stress response between tolerant and sensitive populations (G.2.3). In Specific Aim 3, we will add additional data including amino acid and promoter region substitution patterns, F_{ST} outliers, and genomic segments

indicating signatures of selective sweeps within and among population re-sequencing samples from eight experimental populations (G.2.4) to further validate and support candidate gene discovery.

Closing the gap between markers and transcription patterns and the actual locus under selection is a daunting task. Genetic and physical maps are necessary for top-down approaches, and a complete genome sequence is the ultimate physical map. With a reference genome in hand, physical associations can be made between mapping markers (G.2.1, G.2.2), between mapping markers and genes with interesting patterns of gene expression (G.2.3), and between mapping markers and substitution patterns among comparative population genome sequences (G.3) to facilitate closing of this gap (Figure 1).

G.2.1. QTL mapping data

In a recent review, Stinchcombe and Hoekstra [105] outlined the relative strengths of QTL mapping and population genome scans for identifying genes underlying adaptive traits and argue for the power of combining the two approaches. Both completed and ongoing projects in the Whitehead and Oleksiak labs have and are providing both QTL markers and genome scan markers that will be used to interrogate the genome sequence proposed here. We emphasize that some of these data are already in hand [134] and additional data are currently being generated through support from other sources; therefore we do not seek support for such data collection through this proposal. Dr. Diane Nacci (US EPA; see letter of collaboration) has generated F2 QTL mapping families from crosses of F1 hybrids derived from crosses between parents from tolerant and sensitive populations. Embryos from QTL families have been phenotyped according to PCB sensitivity, and genotyping and genetic mapping is ongoing by the US EPA (microsatellite markers) and Whitehead lab (RAD markers).

Datasets available

The Nacci lab (US EPA) has generated F2 hybrid crosses with family sizes of 200-300. Embryos from these crosses have been phenotyped for PCB-tolerance for populations New Bedford Harbor (tolerant) crossed with Block Island (sensitive) and for populations Bridgeport (tolerant) crossed with Block Island (sensitive). Offspring from these crosses are being genotyped using two different markers: RAD markers (Whitehead lab) and EST-linked microsatellites (Bagley lab, US EPA; see attached letter of collaboration). RAD markers and microsatellite markers will be mapped to the reference genome (Aim 1), and we will test for physical association between these markers and genomic regions implicated as candidates by comparative transcriptomics and population genomics data (see below).

Briefly, restriction site associated DNA (RAD) markers are 90-base Illumina-generated sequence tags that are associated with restriction sites [161, 175]. Sequence tags are of sufficient length to map to unique locations in the genome, thereby physically anchoring the markers. As proof-of-principle, these markers were used to quickly verify *Eda* as the locus responsible for the reduced armor plating morphology in stickleback fish [175], thereby validating data from laborious AFLP and microsatellite mapping [121]. Over 13,000 SNPs were identified in the stickleback studies [161], and we expect a similar number of markers as we adapt this protocol for use in *F. heteroclitus* in ongoing collaboration with the lab of Eric Johnson (University of Oregon; inventor of the RAD protocol). To date, we have genotyped the grandparents of the New Bedford Harbor by Block Island crosses, and genotyping of offspring is ongoing (unpublished data).

The Bagley Lab (US EPA) has designed PCR primers for hundreds of EST-linked microsatellite loci, and using the same hybrid cross families as described above, has genotyped hundreds of individuals. Their group will continue to map more loci throughout 2010, and they intend to collaborate with our group in mapping these markers to the genome sequence (see attached letter of collaboration from Mark Bagley). Since these microsatellite primers were designed from sequences that were linked to ESTs, they should be easily mapped to the reference genome.

Integrating data with the genome sequence

To map markers to the genome, we will conduct heuristic sequence alignments using tools such as Exonerate [176] using the RAD markers or microsatellite primer sequences as query sequences against the reference genome sequence. We will use the Indiana University CGB or Ensembl genome browsers to visualize the physical locations of markers relative to each other, relative to markers from population genome scans (section G.2.2), relative to genes with divergent expression patterns (section G.2.3), and relative to genomic regions exhibiting signatures of recent selective sweeps (section G.3). Using this approach, we can characterize the number of unique locations in the genome identified by QTL markers and test whether similar genomic regions are implicated in independently derived tolerant populations. To identify candidate genes we will test for associations between loci segregating with tolerance and 1)

genes implicated from comparative transcriptomics studies, 2) genes with non-synonymous substitutions from comparisons of population genome scans (Aim 3), and 3) genomic regions exhibiting signatures of recent selective sweeps (Aim 3). These comparisons will help direct the selection of genomic regions from which candidate mutations may ultimately be identified. Candidacy is strengthened where more than one line of evidence converges on the same genes or genomic regions, and especially when the same genomic regions are implicated across converged populations (Figure 1).

G.2.2. Population mapping data

For identifying the genomic basis of adaptive traits, population genomic data are naturally complementary to QTL mapping [102, 105]. Genome scan data are used to identify regions of the genome that show unusually low within-population variability indicating a selective sweep and regions of the genome with unusually high between-population variability indicating divergent selection [177-179]. If genome scan markers coincide with QTL-associated markers, this provides compelling evidence that genes within these regions are significantly involved in adaptive evolution. Our group will use two "population genomics" data sets (AFLP and SNP markers) to test for outlier loci between tolerant and sensitive populations implicating loci under selection. These markers represent unique sequences that can be mapped to unique locations in the genome.

Datasets available

In a data set of ~300 AFLP markers, Oleksiak's group identified 24 outlier loci in genome scans of three tolerant populations and six flanking sensitive reference populations [134]. Most outliers were specific to different tolerant populations, but four were shared among at least two of the tolerant populations. Using these same tolerant and sensitive populations, we also analyzed 367 SNPs identified in both genomic and coding regions [180] and used three independent tests to identify SNPs exhibiting non-neutral behavior (unpublished data). Among populations, 3-15% of all SNPs were outliers in a single statistical test, and 26-46% of outliers were shared among all three tests. One SNP was identified as an outlier in all tests and among all three tolerant populations. This SNP is in the proximal promoter of CYP1A; CYP1A has previously been identified as differentially expressed in the tolerant populations [22]. However, similar to the AFLP data, many outlier SNPs were unique to each tolerant population indicating that either different genomic regions are under selection in different tolerant populations, or that the genetic markers were sorting differently in separate ancestral populations but are physically linked to the same genomic region under selection in separate populations. Mapping these markers to a reference genome will resolve these alternative explanations, by indicating whether or not different outlier loci in different populations are linked and thereby implicate the same genomic region.

Population genomic data are analyzed in two broad ways. First, empirical inbreeding coefficients are calculated for all populations, and outlier loci (loci that reject neutral models of divergence) between adaptively diverged populations are identified [134, 177, 178, 181-184]. The inference is that these outlier loci are physically linked to loci under selection. A second complementary strategy is to use the same markers for selective sweep mapping [185], which exploits the fact that genome regions that are subject to strong selection tend to exhibit reduced genetic diversity since linked loci hitchhike to fixation.

Integrating data with the genome sequence

Markers implicated from the above approaches (for example, as F_{ST} outliers) will be mapped to the genome using BLAST. These population genome scan approaches will be useful for confirming whether QTL-identified loci have played a role in adaptive selection-driven divergence of populations, and to test whether the same loci are implicated in convergent tolerant populations for which QTL mapping families are not available. These data will also serve to delineate boundaries of the genome region affected by a selective sweep, thereby bounding the region within which to search for candidate genes. Using this approach, we can characterize the number of unique locations in the genome identified by outlier loci and test whether similar genomic regions are implicated in independently derived tolerant populations. To identify candidate genes we will test for physical associations between outlier loci and 1) genes implicated from comparative transcriptomics studies, 2) genes with non-synonymous substitutions from comparisons of population genome scans (Aim 3), and 3) genomic regions exhibiting signatures of recent selective sweeps (Aim 3). These comparisons will help direct the selection of genomic regions from which candidate mutations may ultimately be identified. Candidacy is strengthened where more than one line of evidence converges on the same genes or genomic regions, and especially when the same genomic regions are implicated across converged populations (Figure 1).

G.2.3. Transcriptome profiling data

Completed and ongoing projects in the Whitehead, Oleksiak, and Hahn laboratories, and other FGC members (Baldwin, Bain, Di Giulio), are generating functional genomic datasets in studies of tolerant and sensitive populations. The common purpose of these studies is to identify patterns of RNA expression that indicate genes and biochemical pathways functionally involved in producing the derived tolerant phenotype. We first describe the nature of these datasets and then outline the common plan for integrating these data with the genome sequence (Aim 1) to obtain new understanding of genes and regulatory pathways involved in adaptation to these environments.

Datasets available

1. Collaborator Diane Nacci (US EPA) has completed dose-response experiments where embryos derived from a pair of tolerant and sensitive populations (raised in a common clean environment for at least two generations) have been challenged with a range of PCB doses (six log doses). The Whitehead lab has used a custom 6,800 gene oligonucleotide microarray to assay the transcriptome response. Results show that the PCB exposure concentrations at which developmental toxicity emerged, the range of developmental abnormalities exhibited, and global as well as specific gene expression patterns were profoundly different between populations [67]. In the sensitive population PCB exposures produced dramatic, dose-dependent toxic effects, concurrent with the alterations in the expression of many genes. For example, PCB-mediated cardiovascular system failure was associated with the altered expression of cardiomyocyte genes, consistent with sarcomere mis-assembly. In contrast, genome-wide expression was comparatively refractory to PCB induction in the tolerant population. Tolerance was associated with the global blockade of the aryl hydrocarbon receptor (AHR) signaling pathway, the key mediator of PCB toxicity, in contrast to the strong dose-dependent up-regulation of AHR pathway elements observed in the sensitive population. Altered regulation of signaling pathways that cross-talk with AHR was implicated as one candidate mechanism for the adaptive AHR signaling repression and the pollution tolerance that it affords. Having the genome sequence will enable us to test whether F_{ST} outlier loci or QTL mapping markers physically coincide with any of these genes with adaptive patterns of gene expression. Also, if any of the genes with adaptive patterns of expression fall within genomic regions with signatures of a selective sweep (Aim 3), this would be strong evidence for candidacy. Similar experiments were completed in the summer 2009 by collaborator Nacci that include additional tolerant populations and paired sensitive reference populations, to test whether tolerant populations have converged on similar compensatory responses to PCBs, and comparative transcriptomics data collection is nearly complete.

2. The Oleksiak lab has completed a two year time-course experiment in *Fundulus* populations from three separate, genetically distinct, tolerant populations and paired sensitive populations (9 populations total). These experiments were designed to differentiate both evolved and physiologically induced changes in gene expression in tolerant *versus* sensitive populations. Initial experiments with a targeted metabolic array have found that evolved differences among polluted and references affects the expression of 17% of metabolic genes [54]. The Oleksiak group also has completed dose-response experiments in collaboration with Rich DiGiulio (Duke University) to explore effects of polluted sediment as well as representative PAH-type CYP1A inducers and inhibitors on morphology, physiology and gene expression in developing fish embryos from both sensitive and tolerant populations. What is needed is to determine whether the promoters for genes with non-neutral patterns of expression have similar regulatory regions, to define primers to assay for allelic-specific expression to distinguish cis *versus* trans regulatory changes, and to determine whether there are any adjacent loci that share the same patterns of expression. That is, a genome provides the necessary information to address the evolutionary importance of these non-neutral patterns of expression and the molecular tools to define them.

3. The Hahn group has conducted a series of studies comparing targeted gene expression and aryl hydrocarbon receptor (AHR) polymorphisms in killifish from the New Bedford Harbor (NBH; MA) Superfund site and a reference site (Scorton Creek (SC), MA) [138, 147, 186-188]. In an ongoing collaboration with Oleksiak, Hahn's group exposed F1 embryos from the two sites to a dioxin-like PCB and sampled at three different stages of development. Analysis of the embryo RNA samples using a 7,000 gene cDNA array revealed a sizeable set of PCB-responsive genes as well as dramatic population-specific differences in responsiveness to the PCB at all stages. Additional analysis using 454-based sequencing of 3'-anchored cDNA libraries from these fish confirmed the array data and identified additional population-specific responses. The results suggest that entire response pathways are mis-regulated in the tolerant population. Clearly, a genome would provide the data on the regulatory regions

shared by "mis-regulated" genes in polluted populations and the DNA sequences to design primers to survey populations to define if some or many polymorphic nucleotides have adaptive patterns.

Integrating data with the genome sequence

Together, these results describing population-specific differences in basal and induced gene expression provide a rich set of data whose analysis and interpretation will be greatly enhanced by availability of the *Fundulus* genome. Gene expression and genome information will be integrated in several ways: 1) We will test for physical association between genes with evolved differences in expression and mapped QTL, SNP, RAD, and AFLP loci (G.2.1 and G.2.2), and regions of the genome indicative of recent selective sweeps (G.3) (Figure 1). Convergence of these independent lines of evidence on specific loci or genomic regions would provide powerful insights into the mechanisms of adaptation. 2) We will determine if the architecture of genes with evolved differences in gene expression is different from genes that are not differentially expressed. For example, are these genes more or less likely to lack TATAA sites, contain GC-boxes, or have a higher density of known transcription factor binding sites? 3) We will determine if genes that exhibit population-specific expression patterns share binding sites for specific transcription factors known to be involved in responsiveness to environmental stressors. For example, do all genes that are responsive to PCB or PAH in sensitive populations contain consensus regulatory elements for AHR (the PCB and PAH receptor)? Have genes with adaptive patterns of expression acquired or lost *cis*-regulatory binding sites for AHR or other transcription factors that are involved in the PCB or PAH response? Putative transcription factor binding sites will be identified using FootPrinter [189] and MULAN [190, 191] and sites confirmed by comparison to TRANSFAC [192-195]. Fisher's exact test will be used to determine TFBS that are conserved in statistically significant numbers within and between subunits and populations.

G.2.4. Experimental data integration

Where more than one line of evidence implicates common genomic regions, these regions will earn highest priority as candidates. QTL mapping data and population genome scan data will indicate regions of the genome associated with phenotype or test for non-neutral patterns of population genetic variation, respectively. Once these markers are mapped to the genome, we will test for associations between datasets (Figure 1). For example, if markers implicated as QTLs also appear as population genetic outliers, then such markers strongly implicate the genomic region to which they map as being associated with adaptation. Implicated genome regions may be large and include many functional elements. We will further distill our set of candidates using additional gene expression data, comparative population genome sequence data, and what we know about functionally relevant genes from the literature. For example, if the genomic region bounded by mapping markers includes genes with divergent patterns of gene expression between tolerant and sensitive populations, this would strengthen candidacy. Having a genome sequence is critically important for mapping and testing for physical association between markers and genes from our collective experimental datasets. Importantly, population genome scan data (Aim 3) will be critically useful for further ruling out or supporting candidacy; that is, if some candidate loci (QTL markers, mapping markers, genes with divergent expression) fall within regions exhibiting a signature of selective sweep, whereas other candidate loci do not, then this evidence would help prioritize candidacy (Figure 1). Where candidate regions appear shared only among tolerant populations, this would be the strongest evidence for candidacy, because it suggests that the same genomic regions are responsible for derived tolerance in all tolerant populations. This hypothesis will be further tested by screening for genomic regions indicating signatures of recent, and especially convergent, selective sweeps in Aim 3 (section G.3).

G.3. Specific Aim 3 – Population-level Resequencing (Genome Scans)

To screen the genomes of populations for genomic regions containing convergent genotypes, we will sequence pools of DNA obtained from 30 individuals of each of eight sampled populations for a total sampling of 240 individuals. These population genome scans will represent each of the four well-studied tolerant populations, and for each of those, a nearby paired sensitive reference population. These include populations from New Bedford Harbor, MA (Tolerant) and Block Island, RI (sensitive reference nearby New Bedford Harbor); Bridgeport, CT (Tolerant) and Flax pond, NY (sensitive reference nearby Bridgeport); Newark Bay, NJ (Tolerant) and Sandy Hook, NJ (sensitive reference nearby Newark); and Elizabeth River, VA (Tolerant) and Kings Creek, VA (sensitive reference nearby Elizabeth River). These

re-sequencing data will offer far greater resolution of genomic intervals exhibiting non-neutral patterns of evolution compared to the current AFLP and SNP datasets (Aim 2).

With the sequence reads from populations mapped to the reference genome assembly, we can quantify polymorphism and associate regions of the genome that share similar patterns of diversity between tolerant and sensitive populations. For instance, loci that have experienced recent, independent selective sweeps due to similar ecological pressures (pollution exposure) will share the signature of reduced variation over the genomic region surrounding the genomic targets of selection, while members of divergent sensitive-tolerant pairs of populations will not exhibit a shared signal at that site. When phylogenetic independence of ecotypes is established, candidate gene-environment relationships can be identified by scans of variation and allele frequency correlations. Shared alleles that are observed at high frequency in independently evolved lineages that share ecological characteristics (highly polluted) will be detected through our proposed plan to scan for nucleotide diversity across entire genomes.

F_{ST} is a measure of genetic connectivity between populations and is sensitive to demographic parameters such as population size fluctuations that affect the entire genome. Locus-by-locus measures of F_{ST} can help directly identify regions where local evolutionary pressures such as positive natural selection have adjusted allele frequencies rapidly in recent evolutionary history. We propose to calculate locus-specific estimates of F_{ST} across the genome for our tolerant and sensitive populations, with outlier loci (having high F_{ST}) being reasonable candidates for regions associated with diversifying selection among populations [196]. A locus with a relatively high F_{ST} value signifies a region of the genome where local genetic diversity is not easily explained by random mechanisms, but is instead best explained by divergent, rapid fixation of allelic variants. Using the mapped genomic reads for each sampled population, we will calculate F_{ST} values for each defined window (~500 bp) along the genome, comparing within-population diversity estimates with among-population data. While each population may have a set of unique adaptive events, only the outlier F_{ST} values found in common when comparing pooled tolerant and sensitive types will be considered candidates for ecologically relevant adaptation.

Two problems in the analysis of high-throughput sequence data are the relatively high incidences of sequence errors (which create false variation) and the disproportionate sampling of alleles (which results in loss of information when by chance alleles are either absent or greatly under-represented). Our colleague M. Lynch has recently developed maximum-likelihood techniques for extracting key population-genetic parameters in an unbiased fashion [197]. A substantial advantage of these methods is that they do not require extrinsic estimates of the sequence error rate, but instead use the data themselves to estimate this nuisance parameter and remove its influence from population parameter estimates. It also automatically accounts for the bias resulting from allele sampling. As a result, these methods are capable of yielding unbiased estimates of nucleotide variation within restricted chromosomal regions of pooled individuals, patterns of heterozygosity disequilibrium and allele frequencies and LD at individual nucleotide sites in population-wide samples (and hence, the spectrum of allele and haplotype frequencies), even when the error rate exceeds the true site-specific heterozygosity.

From our genome scans of sequences from natural populations, we propose to search for intervals that have: 1) reduced levels of within-population polymorphism, and 2) elevated levels of population subdivision (F_{ST}). For populations and species with occasional gene flow, we expect to find marker intervals that exhibit high F_{ST} values. In principle, such DNA markers are signatures of genomic regions affected by divergent selection [198, 199]. This approach has been used to find gene regions involved in speciation with gene flow [200], and to examine the genetic basis of local adaptation (e.g., [181, 184]). However, for methodological reasons and because of inherent limitations for most study systems, few investigations have successfully narrowed the gap between the F_{ST} outlying markers and the nearest QTL, thus failing to pinpoint the genes of interest (but see [201]). In part, the low resolution is the result of working with species without a complete reference genome sequence. The ability to scan whole genomes from multiple genotypes and populations places us in a position to overcome these difficulties in *Fundulus*, mainly because a complete genome sequence will be in place as a reference (Aim 1).

Finally, we will test for coincidence of genomic regions exhibiting signatures of selective sweeps from these population genome scan data with genomic regions identified as candidates from our experimental data (Aim 2). Where these multiple lines of evidence implicate the same genes, especially if these coincidences are consistent among converged tolerant populations, this will provide the strongest evidence of loci functionally relevant to the derived tolerant phenotype (Figure 1).

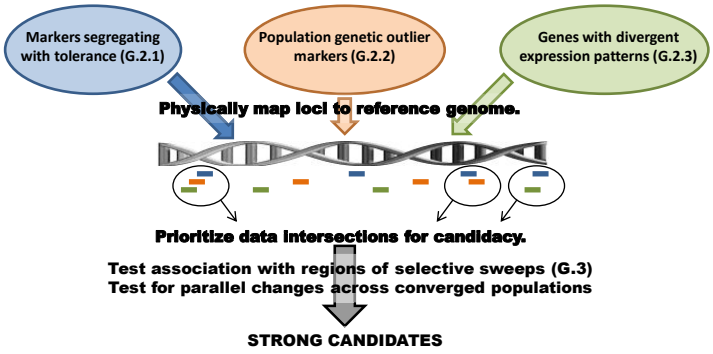
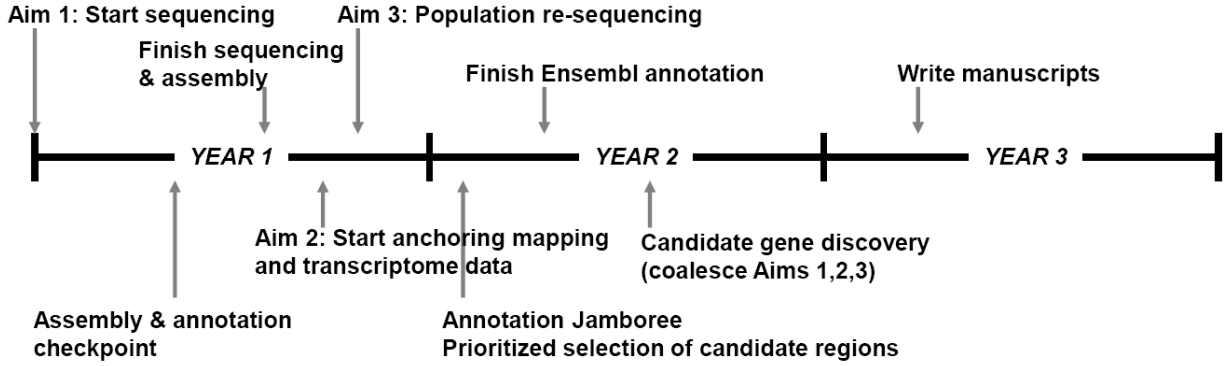


Figure 1: Flowchart of workflow

H. FUTURE STUDIES

Future studies could include gene knockdown experiments and functional enzyme and promoter experiments to verify candidates and experiments necessary to verify the specific nucleotide changes that enabled pollution adaptation. We are experienced in promoter function and enzyme function assays [10, 13, 18, 73, 76, 138, 186, 202-212] and morpholino knockdown in *Fundulus* [213]. Additionally, these data provide a foundation for future research using RNA-Seq that can identify allele-specific and alternative splicing patterns of gene expression, ChIP-seq assays to define all binding sites of specific transcription factors, 3' nucleotide and exon-specific sequences for more robust microarrays, and proteomic studies. In summary, a genome provides answers to our current questions and allows the *Fundulus* community to address the evolution of the molecular mechanisms associated with *F. heteroclitus'* success in many environments.

I. TIMELINE



J. MANAGEMENT PLAN

AW will oversee selection of genome to be sequenced. WW will oversee genome sequencing and assembly (Aim 1). JS will oversee gene finding, annotation, and bioinformatics (Aim 1), and DC will contribute approaches for gene annotation (Aim 1). AW, JS, and JC will organize the annotation workshop, and everyone will participate. AW, WW, JS, JC, MO, DC, and MH will form the annotation steering committee. AW, MH, and MO will contribute experimental data, map their respective mapping and transcriptomics data sets to the genome, and assemble a prioritized list of candidate genes (Aim 2). AW, JC will screen population samples for signatures of natural selection (Aim 3). AW will mentor undergraduate students. AW will coordinate the merging of data sets and writing of papers. Everyone will contribute to analyzing and interpreting data, and writing papers.

REFERENCES:

1. Burnett, K.G., et al., *Fundulus* as the premier teleost model in environmental biology: Opportunities for new insights using genomics. Comparative Biochemistry and Physiology, Part D, 2007. **2**: p. 257-286.
2. Colbourne, J.K., et al., The ecoresponsive genome of *Daphnia pulex.* Science, in press.
3. Richards, S., et al., Genome Sequence of the Pea Aphid *Acyrthosiphon pisum.* PLoS Biology, 2010. **8**(2): p. e1000313.
4. Werren, J.H., et al., Functional and Evolutionary Insights from the Genomes of Three Parasitoid *Nasonia* Species. Science, 2010. **327**(5963): p. 343-348.
5. Tweedie, S., et al., FlyBase: enhancing *Drosophila* Gene Ontology annotations. Nucleic Acids Research, 2009. **37**: p. D555-D559.
6. Colbourne, J.K., V.R. Singan, and D.G. Gilbert, wFleaBase: The *Daphnia* genome database. BMC Bioinformatics, 2005. **6**: p. 45.
7. Crawford, D.L. and D.A. Powers, 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, 1989. **86**(23): p. 9365-9.
8. Crawford, D.L. and D.A. Powers, Evolutionary adaptation to different thermal environments via transcriptional regulation. Molecular Biology & Evolution, 1992. **9**(5): p. 806-813.
9. Oleksiak, M.F. and J.J. Stegeman, Cloning and partial sequence analysis of multiple P450 genes from the marine fish, *Fundulus heteroclitus.* Marine Environmental Research, 1996. **42**(1-4): p. 25.
10. Pierce, V.A. and D.L. Crawford, Variation in the glycolytic pathway: The role of evolutionary and physiological processes. Physiological Zoology, 1996. **69**(3): p. 489-508.
11. Segal, J.A., et al., Descriptive and functional characterization of variation in the *Fundulus heteroclitus* Ldh-B proximal promoter. Journal of Experimental Zoology, 1996. **275**(5): p. 355-364.
12. Hahn, M.E., et al., Molecular evolution of two vertebrate aryl hydrocarbon (dioxin) receptors (AHR1 and AHR2) and the PAS family. Proc Natl Acad Sci U S A, 1997. **94**(25): p. 13743-8.
13. Pierce, V.A. and D.L. Crawford, Phylogenetic analysis of thermal acclimation of the glycolytic enzymes in the genus *Fundulus.* Physiological Zoology, 1997. **70**(6): p. 597-609.
14. Pierce, V.A. and D.L. Crawford, Phylogenetic analysis of glycolytic enzyme expression. Science, 1997. **276**(5310): p. 256-9.
15. Stegeman, J., J., et al., Cytochromes P450 (CYP) in tropical fishes: Catalytic activities, expression of multiple CYP proteins and high levels of microsomal P450 in liver of fishes from Bermuda. Comparative Biochemistry & Physiology. C, Comparative Pharmacology & Toxicology, 1997. **116**(1): p. 61-75.
16. Besselink, H.T., et al., Low inducibility of CYP1A activity by polychlorinated biphenyls (PCBs) in flounder (*Platichthys flesus*): characterization of the Ah receptor and the role of CYP1A inhibition. Toxicol Sci, 1998. **43**(2): p. 161-71.
17. Oleksiak, M.F., et al., Characterization of members of the novel cytochrome P450 subfamilies CYP2N and CYP2P from the fish *Fundulus heteroclitus.* Marine Environmental Research, 1998. **46**(1-5): p. 125-126.
18. Crawford, D.L., V.A. Pierce, and J.A. Segal, Evolutionary physiology of closely related taxa: Analyses of enzyme expression. American Zoologist, 1999. **39**(2): p. 389-400.
19. Crawford, D.L., J.A. Segal, and J.L. Barnett, Evolutionary analysis of TATA-less proximal promoter function. Molecular Biology & Evolution, 1999. **16**(2): p. 194-207.
20. Karchner, S.I., W.H. Powell, and M.E. Hahn, 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. J Biol Chem, 1999. **274**(47): p. 33814-24.
21. Powell, W.H., et al., Functional diversity of vertebrate ARNT proteins: identification of ARNT2 as the predominant form of ARNT in the marine teleost, Fundulus heteroclitus. Arch Biochem Biophys, 1999. **361**(1): p. 156-63.
22. Segal, J.A., J.L. Barnett, and D.L. Crawford, Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. Journal of Molecular Evolution, 1999. **49**: p. 736-749.

23. Oleksiak, M.F., et al., Identification, functional characterization, and regulation of a new cytochrome P450 subfamily, the CYP2Ns. Journal of Biological Chemistry., 2000. **275**(4): p. 2312-21.
24. Podrabsky, J., E., et al., Intraspecific variation in aerobic metabolism and glycolytic enzyme expression in heart ventricles. American Journal of Physiology., 2000. **279**(6 Part 2): p. R2344-R2348.
25. Powell, W.H., et al., Developmental and tissue-specific expression of AHR1, AHR2, and ARNT2 in dioxin-sensitive and -resistant populations of the marine fish Fundulus heteroclitus. Toxicol Sci, 2000. **57**(2): p. 229-39.
26. Powell, W.H. and M.E. Hahn, The evolution of aryl hydrocarbon signaling proteins: diversity of ARNT isoforms among fish species. Mar Environ Res, 2000. **50**(1-5): p. 39-44.
27. Bello, S.M., et al., 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, 2001. **60**(1): p. 77-91.
28. Oleksiak, M.F., K. Kolell, and D.L. Crawford, The utility of natural populations for microarray analyses: isolation of genes necessary for functional genomic studies. Marine Biotechnology, 2001. **3**: p. S203-S211.
29. Toomey, B.H., et al., 2,3,7,8-Tetrachlorodibenzo-p-dioxin induces apoptotic cell death and cytochrome P4501A expression in developing *Fundulus heteroclitus* embryos. Aquat Toxicol, 2001. **53**(2): p. 127-38.
30. Bard, S.M., et al., Expression of P-glycoprotein in killifish (*Fundulus heteroclitus*) exposed to environmental xenobiotics. Aquat Toxicol, 2002. **59**(3-4): p. 237-51.
31. Crawford, D.L., Evolution of physiological adapatation, in Cell and Molecular Responses to Stress, K.B. Storey and J.M. Storey, Editors. 2002, Elsevier Publishing: NY.
32. Karchner, S.I., et al., Regulatory interactions among three members of the vertebrate aryl hydrocarbon receptor family: AHR repressor, AHR1, and AHR2. J Biol Chem, 2002. **277**(9): p. 6949-59.
33. Kolell, K.J. and D.L. Crawford, Evolution of Sp transcription factors. Molecular Biology & Evolution., 2002. **19**(3): p. 216-22.
34. Oleksiak, M.F., G.A. Churchill, and D.L. Crawford, Variation in gene expression within and among natural populations. Nature Genetics, 2002. **32**(2): p. 261-6.
35. Powell, W.H. and M.E. Hahn, Identification and functional characterization of hypoxia-inducible factor 2alpha from the estuarine teleost, *Fundulus heteroclitus*: interaction of HIF-2alpha with two ARNT2 splice variants. J Exp Zool, 2002. **294**(1): p. 17-29.
36. Meyer, J.N., et al., Expression and inducibility of aryl hydrocarbon receptor pathway genes in wild-caught killifish (*Fundulus heteroclitus*) with different contaminant-exposure histories. Environ Toxicol Chem, 2003. **22**(10): p. 2337-43.
37. Hahn, M.E., et al., Aryl hydrocarbon receptor polymorphisms and dioxin resistance in Atlantic killifish (*Fundulus heteroclitus*). Pharmacogenetics, 2004. **14**(2): p. 131-43.
38. Paschall, J.E., et al., FunnyBase: a Systems Level Functional Annotation of Fundulus ESTs for the Analysis of Gene Expression. BMC Genomics, 2004. **5**: p. 96.
39. Powell, W.H., et al., Cloning and analysis of the CYP1A promoter from the atlantic killifish (*Fundulus heteroclitus*). Mar Environ Res, 2004. **58**(2-5): p. 119-24.
40. Oleksiak, M.F., J.L. Roach, and D.L. Crawford, Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus.* Nature Genetics, 2005. **37**(1): p. 67-72.
41. Whitehead, A. and D. Crawford, Variation in tissue-specific gene expression among natural populations. Genome Biology, 2005. **6**(2): p. R13.1-13.14.
42. Stanton, C.R., et al., Arsenic inhibits CFTR-mediated chloride secretion by killifish (*Fundulus heteroclitus*) opercular membrane. Cell Physiol Biochem, 2006. **17**(5-6): p. 269-78.
43. Tarrant, A.M., et al., Estrogen receptor-related receptors in the killifish *Fundulus heteroclitus*: diversity, expression, and estrogen responsiveness. J Mol Endocrinol, 2006. **37**(1): p. 105-20.
44. Whitehead, A. and D.L. Crawford, Variation within and among species in gene expression: raw material for evolution. Mol Ecol, 2006. **15**(5): p. 1197-211.
45. Whitehead, A. and D.L. Crawford, Neutral and adaptive variation in gene expression. Proc Natl Acad Sci U S A, 2006. **103**(14): p. 5425-30.

46. Burnett, K.G., et al., *Fundulus* as the premier teleost model in environmental biology: opportunities for new insights using genomics. Comp Biochem Physiol Part D Genomics Proteomics, 2007. **2**(4): p. 257-86.
47. Crawford, D.L. and M.F. Oleksiak, The biological importance of measuring individual variation. J Exp Biol, 2007. **210**(9): p. 1613-1621.
48. Fisher, M.A. and M.F. Oleksiak, Convergence and divergence in gene expression among natural populations exposed to pollution. BMC Genomics, 2007. **8**: p. 108.
49. Miller, D.S., et al., MRP2 and acquired tolerance to inorganic arsenic in the kidney of killifish (*Fundulus heteroclitus*). Toxicol Sci, 2007. **97**(1): p. 103-10.
50. Shaw, J.R., et al., Role of glucocorticoid receptor in acclimation of killifish (*Fundulus heteroclitus*) to seawater and effects of arsenic. Am J Physiol Regul Integr Comp Physiol, 2007. **292**(2): p. R1052-60.
51. Shaw, J.R., et al., The influence of exposure history on arsenic accumulation and toxicity in the killifish, *Fundulus heteroclitus.* Environ Toxicol Chem, 2007. **26**(12): p. 2704-9.
52. Duvernell, D.D., et al., Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, *Fundulus heteroclitus.* Mol Ecol, 2008. **17**(5): p. 1344-60.
53. Matson, C.W., et al., Development of the morpholino gene knockdown technique in *Fundulus heteroclitus*: a tool for studying molecular mechanisms in an established environmental model. Aquat Toxicol, 2008. **87**(4): p. 289-95.
54. Oleksiak, M.F., Changes in gene expression due to chronic exposure to environmental pollutants. Aquat Toxicol, 2008. **90**(3): p. 161-71.
55. Shaw, J.R., et al., The role of SGK and CFTR in acute adaptation to seawater in Fundulus heteroclitus. Cell Physiol Biochem, 2008. **22**(1-4): p. 69-78.
56. Williams, L.M. and M.F. Oleksiak, Signatures of selection in natural populations adapted to chronic pollution. BMC Evol Biol, 2008. **8**: p. 282.
57. Merson, R.R., S.I. Karchner, and M.E. Hahn, Interaction of fish aryl hydrocarbon receptor paralogs (AHR1 and AHR2) with the retinoblastoma protein. Aquat Toxicol, 2009. **94**(1): p. 47-55.
58. Scott, C.P., et al., Technical analysis of cDNA microarrays. PLoS ONE, 2009. **4**(2): p. e4486.
59. Scott, C.P., D.A. Williams, and L. Crawford Douglas, The effect of genetic and environmental variation on gene expression Molecular Ecology, 2009. **18**: p. 2832-2843.
60. Whitehead, A., Comparative mitochondrial genomics within and among species of killifish. BMC Evol Biol, 2009. **9**: p. 11.
61. Cheviron, Z.A., A. Whitehead, and R.T. Brumfield, Transcriptomic variation and plasticity in rufous-collared sparrows (*Zonotrichia capensis*) along an altitudinal gradient. Molecular Ecology, 2008. **17**(20): p. 4556-4569.
62. Duvernell, D.D., et al., Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, *Fundulus heteroclitus.* Molecular Ecology, 2008. **17**(5): p. 1344-1360.
63. Triant, D.A. and A. Whitehead, Simultaneous extraction of high-quality RNA and DNA from small tissue samples. J Hered, 2009. **100**(2): p. 246-250.
64. Whitehead, A., Comparative mitochondrial genomics within and among species of killifish. BMC Evolutionary Biology, 2009. **9**: p. 11.
65. Whitehead, A., The evolutionary radiation of diverse osmotolerant physiologies in killifish (*Fundulus* sp.). Evolution, 2010. **64**(7): p. 2070-2085.
66. Whitehead, A., et al., Functional genomics of physiological plasticity and local adaptation in killifish. Journal of Heredity, 2010. **available online in advance of publication**.
67. Whitehead, A., et al., Comparative transcriptomics implicates mechanisms of evolved pollution tolerance in a killifish population. Molecular Ecology, 2010. **19**: p. 5186-5203.
68. Colbourne, J.K., et al., Sampling *Daphnia's* expressed genes: preservation, expansion and invention of crustacean genes with reference to insect genomes. BMC Genomics, 2007. **8**: p. 217.
69. Cook, J.C., et al., "Omic" approaches in the context of environmental toxicology, in Genomic Approaches for Cross-Species Extrapolation in Toxicology, W.H. Benson and R. DiGulio, Editors. 2006, CRC Press: Boca Raton, FL.

70. Denslow, N.D., et al., Selection of surrogate animal species for comparative toxicogenomics, in Genomic Approaches for Cross-Species Extrapolation in Toxicology, W.H. Benson and R. DiGulio, Editors. 2006, CRC Press: Boca Raton, FL.
71. Shaw, J., et al., Gene response profiles for *Daphnia pulex* exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC Genomics, 2007. **8**(1): p. 477.
72. Shaw, J.R., et al., *Daphnia* as an emerging model for toxicological genomics, in Advances in Experimental Biology: Comparative Toxicogenomics, C. Hogstrand and P. Kille, Editors. 2008, Elsevier Science: London.
73. Podrabsky, J.E., et al., Intraspecific variation in aerobic metabolism and glycolytic enzyme expression in heart ventricles. American Journal of Physiology-Regulatory Integrative and Comparative Physiology, 2000. **279**(6): p. R2344-R2348.
74. Crawford, D.L., Functional genomics does not have to be limited to a few select organisms. Genome Biology., 2001. **2**(1): p. interactions1001.1–interactions1001.2.
75. Oleksiak, M.F., K.J. Kolell, and D.L. Crawford, Utility of natural populations for microarray analyses: Isolation of genes necessary for functional genomic studies. Marine Biotechnology, 2001. **3**: p. S203-S211.
76. Kolell, K.J. and D.L. Crawford, Evolution of Sp transcription factors. Molecular Biology and Evolution, 2002. **19**(3): p. 216-222.
77. Oleksiak, M.F., G.A. Churchill, and D.L. Crawford, Variation in gene expression within and among natural populations. Nature Genetics, 2002. **32**(2): p. 261-266.
78. Oleksiak, M.F. and D.L. Crawford, 5' genomic structure of human Sp3. Molecular Biology & Evolution, 2002. **19**(11): p. 2026-9.
79. Clark, M.S., L. Crawford Douglas, and A. Cossins, Worldwide genomic resources for non-model fish species. Comparative and Functional Genomics, 2003. **4**: p. 502-508.
80. Paschall, J.E., et al., FunnyBase: a systems level functional annotation of *Fundulus* ESTs for the analysis of gene expression. BMC Genomics, 2004. **5**(1): p. 96.
81. Cossins, A.R. and D.L. Crawford, Fish as models for environmental genomics. Nature Reviews Genetics, 2005. **6**(4): p. 324-33.
82. Whitehead, A. and D.L. Crawford, Variation in tissue-specific gene expression among natural populations. Genome Biology, 2005. **6**(2): p. -.
83. Oleksiak, M., F. and L. Crawford Douglas, Functional Genomics in Fishes, Insights into Physiological Complexity, in The Physiology of Fishes, D. Evan and J. Claiborne, Editors. 2006, CRC Press: Boca Raton. p. 523-550.
84. Richardson, D.E., et al., High-throughput species identification: from DNA isolation to bioinformatics. Molecular Ecology, 2006. **7**: p. 199-207.
85. Whitehead, A. and D.L. Crawford, Neutral and adaptive variation in gene expression. Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(14): p. 5425-5430.
86. Whitehead, A. and D.L. Crawford, Variation within and among species in gene expression: raw material for evolution. Molecular Ecology, 2006. **15**(5): p. 1197-1211.
87. Burnett, K.G., et al., *Fundulus* as the premier teleost model in environmental biology: Opportunities for new insights using genomics. Comparative Biochemistry and Physiology, Part D, 2007. **2**(4): p. 257-286.
88. Crawford, D.L., Human reference sequence makes sense of names. Nature, 2007. **447**(7141): p. 142.
89. Crawford, D.L. and M.F. Oleksiak, The biological importance of measuring individual variation. Journal of Experimental Biology, 2007. **210**(9): p. 1613-1621.
90. Vera, J.C., et al., Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology, 2008. **17**(7): p. 1636-1647.
91. Williams, D.A., S.D. Brown, and D.L. Crawford, Contemporary and historical influences on the genetic structure of the estuarine-dependent Gulf killifish Fundulus grandis. Marine Ecology-Progress Series, 2008. **373**: p. 111-121.
92. Everett, M.V. and D.L. Crawford, Adaptation versus allometry: Population and body mass effects on hypoxic metabolism in *Fundulus grandis.* Physiological and Biochemical Zoology, in press.
93. Everett, M.V. and D.L. Crawford, Time course for hypoxic induction of mRNA expression in preparation.

94. Oleksiak, M., F. and D.L. Crawford, Inter and intra-specific variation in cardiac metabolism using three different substrates. in review.
95. Rees, B.B. and L. Crawford Douglas, Individual variation in protein expression: a protemic analysis of gene expression in preparation.
96. Young, S. and D. Crawford, L., Improved Annotations with FuzzyMatch and Incremental p-values. in review.
97. Williams, L., et al., SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genetics, 2010. **11**(1): p. 32.
98. Williams, L.M. and M.F. Oleksiak, Ecologically and evolutionarily important SNPs identified in natural populations. Molecular Biology & Evolution, in press.
99. Young, S., Improved EST Annotation Using Multiple External Databases, in Bioinformatics. 2008, North Carolina State University: Raleigh.
100. Triant, D.A. and A. Whitehead, Simultaneous extraction of high-quality RNA and DNA from small tissue samples. J Hered, 2009. **100**(2): p. 246-50.
101. Arendt, J. and D. Reznick, Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends in Ecology & Evolution, 2008. **23**(1): p. 26-32.
102. Ellegren, H. and B.C. Sheldon, Genetic basis of fitness differences in natural populations. Nature, 2008. **452**(7184): p. 169-175.
103. Hoekstra, H.E. and J.A. Coyne, The locus of evolution: Evo devo and the genetics of adaptation. Evolution, 2007. **61**(5): p. 995-1016.
104. Hoffmann, A.A. and Y. Willi, Detecting genetic responses to environmental change. Nat Rev Genet, 2008. **9**(6): p. 421-432.
105. Stinchcombe, J.R. and H.E. Hoekstra, Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity, 2008. **100**(2): p. 158-170.
106. Hanikenne, M., et al., Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. Nature, 2008. **453**(7193): p. 391-U44.
107. Gompel, N., et al., Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature, 2005. **433**(7025): p. 481-487.
108. Carroll, S.B., Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom. 1st ed. 2005, New York: Norton. xi, 350 p.
109. Hoekstra, H.E., et al., A single amino acid mutation contributes to adaptive beach mouse color pattern. Science, 2006. **313**(5783): p. 101-104.
110. Mundy, N.I., et al., Conserved genetic basis of a quantitative plumage trait involved in mate choice. Science, 2004. **303**(5665): p. 1870-1873.
111. Rosenblum, E.B., H.E. Hoekstra, and M.W. Nachman, Adaptive reptile color variation and the evolution of the Mc1r gene. Evolution, 2004. **58**(8): p. 1794-1808.
112. Chen, L.B., A.L. DeVries, and C.H.C. Cheng, Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(8): p. 3817-3822.
113. Geffeney, S.L., et al., Evolutionary diversification of TTX-resistant sodium channels in a predator-prey interaction. Nature, 2005. **434**(7034): p. 759-763.
114. Jessen, T.H., et al., Adaptation of Bird Hemoglobins to High-Altitudes - Demonstration of Molecular Mechanism by Protein Engineering. Proceedings of the National Academy of Sciences of the United States of America, 1991. **88**(15): p. 6519-6522.
115. Storz, J.F., et al., The Molecular Basis of High-Altitude Adaptation in Deer Mice. PLoS Genetics, 2007. **3**(3): p. e45.
116. Yamamoto, Y., D.W. Stock, and W.R. Jeffery, Hedgehog signalling controls eye degeneration in blind cavefish. Nature, 2004. **431**(7010): p. 844-847.
117. Abzhanov, A., et al., The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. Nature, 2006. **442**(7102): p. 563-567.
118. Abzhanov, A., et al., Bmp4 and morphological variation of beaks in Darwin's finches. Science, 2004. **305**(5689): p. 1462-1465.
119. Jeong, S., et al., The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell, 2008. **132**(5): p. 783-793.
120. Rebeiz, M., et al., Stepwise modification of a modular enhancer underlies adaptation in a *Drosophila* population. Science, 2009. **326**(5960): p. 1663-1667.

121. Colosimo, P.F., et al., Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science, 2005. **307**(5717): p. 1928-1933.
122. Colosimo, P.F., et al., The genetic architecture of parallel armor plate reduction in threespine sticklebacks. Plos Biology, 2004. **2**(5): p. 635-641.
123. Cresko, W.A., et al., Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(16): p. 6050-6055.
124. Shapiro, M.D., M.A. Bell, and D.M. Kingsley, Parallel genetic origins of pelvic reduction in vertebrates. Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(37): p. 13753-13758.
125. Wilkens, H. and U. Strecker, Convergent evolution of the cavefish *Astyanax* (Characidae, Teleostei): genetic evidence from reduced eye-size and pigmentation. Biological Journal of the Linnean Society, 2003. **80**(4): p. 545-554.
126. Hoekstra, H.E. and M.W. Nachman, Different genes underlie adaptive melanism in different populations of rock pocket mice. Molecular Ecology, 2003. **12**(5): p. 1185-1194.
127. Barrett, R.D.H. and D. Schluter, Adaptation from standing genetic variation. Trends in Ecology & Evolution, 2008. **23**(1): p. 38-44.
128. Nacci, D.E., et al., Predicting the occurrence of genetic adaptation to dioxinlike compounds in populations of the estuarine fish *Fundulus heteroclitus.* Environmental Toxicology and Chemistry, 2002. **21**(7): p. 1525-1532.
129. Van Veld, P.A. and D.E. Nacci, Toxicity resistance, in The Toxicology of Fishes, R.T. Di Giulio and D.E. Hinton, Editors. 2008, Taylor and Francis: Boca Raton, FL.
130. Antkiewicz, D.S., et al., Heart malformation is an early response to TCDD in embryonic zebrafish. Toxicological Sciences, 2005. **84**(2): p. 368-377.
131. Nacci, D., D. Champlin, and S. Jayaraman, Adaptation of the estuarine fish *Fundulus heteroclitus* (Atlantic Killifish) to polychlorinated biphenyls (PCBs). Estuaries and Coasts, 2010. **33**: p. 853-864.
132. Nacci, D.E., et al., Effects of environmental stressors on wildlife populations, in Coastal and Estuarine Risk Assessment: Risk on the Edge, M.C. Newman, Editor. 2002, CRC Press/Lewis Publishers: Washington, DC.
133. Adams, S.M., J.B. Lindmeier, and D.D. Duvernell, Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus.* Molecular Ecology, 2006. **15**(4): p. 1109-1123.
134. Williams, L.M. and M.F. Oleksiak, Signatures of selection in natural populations adapted to chronic pollution. BMC Evolutionary Biology, 2008. **8**: p. 282.
135. Dean, A.M. and J.W. Thornton, Mechanistic approaches to the study of evolution: the functional synthesis. Nature Reviews Genetics, 2007. **8**(9): p. 675-688.
136. Elskus, A.A., et al., Altered CYP1A expression in *Fundulus heteroclitus* adults and larvae: a sign of pollutant resistance? Aquatic Toxicology, 1999. **45**(2-3): p. 99-113.
137. Nacci, D., et al., Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. Marine Biology, 1999. **134**(1): p. 9-17.
138. Bello, S.M., et al., Acquired resistance to aryl hydrocarbon receptor agonists in a population of *Fundulus heteroclitus* from a marine Superfund site: In vivo and in vitro studies on the induction of xenobiotic-metabolizing enzymes. Toxicological Sciences, 2001. **60**(1): p. 77-91.
139. Meyer, J. and R. Di Giulio, 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, 2002. **54**(3-5): p. 621-626.
140. Meyer, J.N., D.E. Nacci, and R.T. Di Giulio, Cytochrome P4501A (CYP1A) in killifish (*Fundulus heteroclitus*): Heritability of altered expression and relationship to survival in contaminated sediments. Toxicological Sciences, 2002. **68**(1): p. 69-81.
141. Ownby, D.R., et al., Fish (*Fundulus heteroclitus*) populations with different exposure histories differ in tolerance of creosote-contaminated sediments. Environmental Toxicology and Chemistry, 2002. **21**(9): p. 1897-1902.

142. Roark, S.A., et al., Population genetic structure of a nonmigratory estuarine fish (*Fundulus heteroclitus*) across a strong gradient of polychlorinated biphenyl contamination. Environ Toxicol Chem, 2005. **24**(3): p. 717-25.
143. McMillan, A.M., et al., Genetic diversity and structure of an estuarine fish (*Fundulus heteroclitus*) indigenous to sites associated with a highly contaminated urban harbor. Ecotoxicology, 2006. **15**(6): p. 539-548.
144. Mulvey, M., et al., Genetic structure of *Fundulus heteroclitus* from PAH-contaminated and neighboring sites in the Elizabeth and York Rivers. Aquatic toxicology, 2002. **61**: p. 195-209.
145. Mulvey, M., et al., Genetic structure and mtDNA diversity of *Fundulus heteroclitus* populations from polycyclic aromatic hydrocarbon-contaminated sites. Environ Toxicol Chem, 2003. **22**(3): p. 671-7.
146. Cohen, S., Strong positive selection and habitat-specific amino acid substitution patterns in MHC from an estuarine fish under intense pollution stress. Mol Biol Evol, 2002. **19**(11): p. 1870-1880.
147. Hahn, M.E., et al., Aryl hydrocarbon receptor polymorphisms and dioxin resistance in Atlantic killifish (*Fundulus heteroclitus*). Pharmacogenetics, 2004. **14**: p. 131-143.
148. Hahn, M.E., et al., Mechanism of PCB- and Dioxin-Resistance in Fish in the Hudson River Estuary: Role of Receptor Polymorphisms. Final Report, Hudson River Foundation Grant 004/02A. 2005.
149. Meyer, J.N. and R.T. Di Giulio, Heritable adaptation and fitness costs in killifish (*Fundulus heteroclitus*) inhabiting a polluted estuary. Ecological Applications, 2002. **13**(2): p. 490-503.
150. Nacci, D., D. Champlin, and S. Jayaraman, Adaptation of the estuarine fish *Fundulus heteroclitus* (Atlantic Killifish) to polychlorinated biphenyls (PCBs). Estuaries and coasts, 2010.
151. Dawley, R.M., Clonal hybrids of the common laboratory fish *Fundulus heteroclitus*. Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(6): p. 2485-2488.
152. Hardie, D.C. and P.D.N. Hebert, Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences, 2004. **61**(9): p. 1636-1646.
153. Hinegardner, R., Evolution of cellular DNA content in teleost fishes. American Naturalist, 1968. **102**: p. 517-523.
154. Gregory, T.R. Animal Genome Size Database. 2001; Available from: http://www.genomesize.com/fish.htm.
155. Li, R., et al., The sequence and *de novo* assembly of the giant panda genome. Nature, 2010. **463**(311-317).
156. Li, R., et al., *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Research, 2010(20): p. 265-272.
157. Simpson, J.T., et al., ABySS: A parallel assembler for short read sequence data. Genome Research, 2009. **19**(6): p. 1117-1123.
158. Flicek, P. and E. Birney, Sense from sequence reads: methods for alignment and assembly. Nat Meth, 2009. **6**(11s): p. S6-S12.
159. Gnerre, S., et al., High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences, 2011. **epub ahead of print**.
160. Choi, J.H., et al., A machine-learning approach to combined evidence validation of genome assemblies. Bioinformatics, 2008. **24**(6): p. 744-750.
161. Baird, N.A., et al., Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE, 2008. **3**(10): p. e3376.
162. Curwen, V., et al., The Ensembl automatic gene annotation system. Genome Research, 2004. **14**(5): p. 942-950.
163. Salamov, A.A. and V.V. Solovyev, *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Research, 2000. **10**(4): p. 516-522.
164. Birney, E. and R. Durbin, Using GeneWise in the *Drosophila* annotation experiment. Genome Research, 2000. **10**(4): p. 547-548.
165. Korf, I., Gene finding in novel genomes. BMC Bioinformatics, 2004. **5**: p. 59.
166. Pruitt, K.D., T. Tatusova, and D.R. Maglott, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 2005. **33**: p. D501-D504.
167. Haas, B.J., et al., Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research, 2003. **31**(19): p. 5654-5666.

168. Harris, M.A., et al., The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 2004. **32**: p. D258-D261.
169. Koonin, E.V., et al., A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology, 2004. **5**(2): p. -.
170. Kanehisa, M., et al., The KEGG resource for deciphering the genome. Nucleic Acids Research, 2004. **32**: p. D277-D280.
171. Elsik, C.G., et al., Community annotation: Procedures, protocols, and supporting tools. Genome Research, 2006. **16**(11): p. 1329-1333.
172. Stein, L., Genome annotation: From sequence to biology. Nature Reviews Genetics, 2001. **2**(7): p. 493-503.
173. Cameron, R.A., et al., SpBase: the sea urchin genome database and web site. Nucl. Acids Res., 2009. **37**(suppl_1): p. D750-754.
174. Stein, L.D., et al., The Generic Genome Browser: A building block for a model organism system database. Genome Research, 2002. **12**(10): p. 1599-1610.
175. Miller, M.R., et al., Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research, 2007. **17**(2): p. 240-248.
176. Slater, G.S. and E. Birney, Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 2005. **6**: p. 31.
177. Black, W.C., et al., Population genomics: Genome-wide sampling of insect populations. Annual Review of Entomology, 2001. **46**: p. 441-469.
178. Luikart, G., et al., The power and promise of population genomics: From genotyping to genome typing. Nature Reviews Genetics, 2003. **4**(12): p. 981-994.
179. Schlotterer, C., Hitchhiking mapping - functional genomics from the population genetics perspective. Trends in Genetics, 2003. **19**(1): p. 32-38.
180. Williams, L.M., et al., SNP identification, verification, and utility for population genetics in a non-model genus. BMC Genet, 2010. **11**: p. 32.
181. Beaumont, M.A., Adaptation and speciation: what can F-st tell us? Trends in Ecology & Evolution, 2005. **20**(8): p. 435-440.
182. Bonin, A., et al., Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). Molecular Biology and Evolution, 2006. **23**(4): p. 773-783.
183. Campbell, D. and L. Bernatchez, Generic 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, 2004. **21**(5): p. 945-956.
184. Storz, J.F., Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular Ecology, 2005. **14**(3): p. 671-688.
185. Williamson, S.H., et al., Localizing recent adaptive evolution in the human genome. PLoS Genetics, 2007. **3**(6): p. 901-915.
186. Karchner, S.I., et al., Regulatory interactions among three members of the vertebrate aryl hydrocarbon receptor family: AHR repressor, AHR1, and AHR2. Journal of Biological Chemistry, 2002. **277**(9): p. 6949-6959.
187. Meyer, J.N., et al., Expression and inducibility of aryl hydrocarbon receptor (AHR) pathway genes in wild-caught killifish (*Fundulus heteroclitus*) with different contaminant exposure histories. Environmental Toxicology and Chemistry, 2003. **22**(10): p. 2337-2343.
188. Powell, W.H., et al., Developmental and tissue-specific expression of AHR1, AHR2, and ARNT2 in dioxin-sensitive and -resistant populations of the marine fish, *Fundulus heteroclitus.* Toxicological Sciences, 2000. **57**: p. 229-239.
189. Blanchette, M. and M. Tompa, FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res, 2003. **31**(13): p. 3840-2.
190. Ovcharenko, I., et al., Mulan: multiple-sequence local alignment and visualization for studying function and evolution. Genome Research, 2005. **15**(1): p. 184-94.
191. Ovcharenko, I., et al., Evolution and functional classification of vertebrate gene deserts. Genome Research, 2005. **15**(1): p. 137-45.
192. Fu, Y. and Z. Weng, Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences. Genome Inform, 2005. **16**(1): p. 68-72.

193. Fogel, G.B., et al., A statistical analysis of the TRANSFAC database. Biosystems, 2005. **81**(2): p. 137-54.
194. Matys, V., et al., TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res, 2003. **31**(1): p. 374-8.
195. Wingender, E., et al., The TRANSFAC system on gene expression regulation. Nucleic Acids Res, 2001. **29**(1): p. 281-3.
196. Holsinger, K.E. and B.S. Weir, FUNDAMENTAL CONCEPTS IN GENETICS Genetics in geographically structured populations: defining, estimating and interpreting F-ST. Nature Reviews Genetics, 2009. **10**(9): p. 639-650.
197. Lynch, M., Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects. Molecular Biology and Evolution, 2008. **25**(11): p. 2409-2419.
198. Lewontin, R.C. and J. Krakauer, Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms. Genetics, 1973. **74**(1): p. 175-195.
199. Via, S., Natural selection in action during speciation. Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**: p. 9939-9946.
200. Butlin, R.K., Population genomics and speciation. Genetica, 2010. **138**(4): p. 409-418.
201. Rogers, S.M. and L. Bernatchez, Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (Coregonus clupeaformis). Molecular Ecology, 2005. **14**(2): p. 351-361.
202. Crawford, D.L., J.A. Segal, and J.L. Barnett, Evolutionary analysis of TATA-less proximal promoter function. Molecular Biology and Evolution, 1999. **16**(2): p. 194-207.
203. Evans, B.R., et al., Repression of aryl hydrocarbon receptor (AHR) signaling by AHR repressor: role of DNA binding and competition for AHR nuclear translocator. Mol Pharmacol, 2008. **73**(2): p. 387-98.
204. Karchner, S.I., et al., The active form of human aryl hydrocarbon receptor repressor lacks exon 8 and its Pro185 and Ala185 variants repress both AHR and HIF. Mol Cell Biol, 2009. **29**(13): p. 3465-3477.
205. Oleksiak, M.F., et al., Identification, functional characterization, and regulation of a new cytochrome P450 subfamily, the CYP2Ns. Journal of Biological Chemistry, 2000. **275**(4): p. 2312-2321.
206. Oleksiak, M.F., et al., Identification and regulation of a new vertebrate cytochrome P450 subfamily, the CYP2Ps, and functional characterization of CYP2P3, a conserved arachidonic acid epoxygenase/19-hydroxylase. Archives of Biochemistry and Biophysics, 2003. **411**(2): p. 223-234.
207. Pierce, V.A. and D.L. Crawford, Rapid enzyme assays investigating the variation in the glycolytic pathway in field-caught populations of *Fundulus heteroclitus.* Biochemical Genetics, 1994. **32**(9-10): p. 315-330.
208. Pierce, V.A. and D.L. Crawford, Phylogenetic analysis of glycolytic enzyme expression. Science, 1997. **276**(5310): p. 256-259.
209. Powell, W.H., et al., Cloning and analysis of the CYP1A promoter from the Atlantic killifish (*Fundulus heteroclitus*). Marine Environmental Research, 2004. **58**: p. 119-124.
210. Segal, J.A., J.L. Barnett, and D.L. Crawford, Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. Journal of Molecular Evolution, 1999. **49**(6): p. 736-749.
211. Stegeman, J.J., et al., Cytochromes p450 (CYP) in tropical fishes: Catalytic activities, expression of multiple CYP proteins and high levels of microsomal p450 in liver of fishes from Bermuda. Comparative Biochemistry and Physiology C-Pharmacology Toxicology & Endocrinology, 1997. **116**(1): p. 61-75.
212. Yang, X., et al., The aryl hydrocarbon receptor constitutively represses c-myc transcription in human mammary tumor cells. Oncogene, 2005. **24**(53): p. 7869-81.
213. Matson, C.W., et al., Development of the morpholino gene knockdown technique in *Fundulus heteroclitus*: A tool for studying molecular mechanisms in an established environmental model. Aquatic Toxicology, 2008. **87**(4): p. 289-295.