Collaborative Research: The genomic basis of dramatic, rapid, convergent evolution in the killifish *Fundulus heteroclitus*

PI: Andrew Whitehead (Louisiana State University)

Collaborator: Wesley Warren (Washington University)

Collaborator: Joseph Shaw (Indiana University)

Collaborator: Marjorie Oleksiak, Douglas Crawford (University of Miami)

Collaborator: Mark Hahn (Woods Hole Oceanographic Institution)

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*, and additional low-coverage genomes from experimental populations. The second aim is to map multiple lines of experimental evidence to identify the number and physical location of loci associated with the evolved tolerance phenotype. The third aim is to validate candidates and test whether the same loci are implicated in each population where tolerance has convergently evolved. **These data will be used to 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 global climate 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*, as well as additional low-coverage sequences from wild populations that vary in their pollution tolerance (hereafter referred to as "experimental populations"). 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 include validation of candidate genes by screening patterns of polymorphism in experimental population samples and mapping families. 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. Additional genomes from experimental populations will be generated at lower sequence coverage and used as additional evidence to accelerate and support candidate gene discovery.

Aim 2: Map experimental data to the reference genome to identify candidate genes involved in adaptation. To achieve this, we will map four lines of genome-scale 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. Finally, comparisons of genome sequences from experimental populations (Aim 1) will buttress evidence from data sources 1-3 and accelerate the pace of discovery. 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 preliminary testing of hypotheses.

Aim 3: Validation of candidate genes and test hypotheses. From a prioritized list of candidate genes identified in Aim 2, we will screen patterns of polymorphism in multiple tolerant and sensitive populations and in mapping families to further test for their role in adaptation. These data will more explicitly test hypotheses related to the ancestral versus *de novo* origin of adaptive variants.

A.3. Expected Significance

With genome sequence linked to experimental genome-scale data (QTL, transcriptomics, nucleotide divergences among populations), 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. 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 cutting-edge genomic, evolutionary, and bioinformatic sciences.

B. RESPONSE to COMMENTS from PREVIOUS PANEL

The original version of this proposal was reviewed in spring 2009. The proposal scored very well (E,E,E,VG,G), and comments from reviewers were very helpful. Our overall research objectives were considered highly meritorious (*e.g.*: "The questions raised by the PIs are important and entirely appropriate") and our group considered competent and able (*e.g.*: "The authors make a compelling case that they will be able to successfully carry out the research"). The main criticism was that the proposal was too focused on our sequencing, assembly, and annotation strategy, at the expense of detailed description of hypothesis testing. We thought that the three pages originally spent to outline our sequencing, assembly, and annotation strategy was necessary to convince reviewers that our team could successfully generate this important research tool in order to enable our scientific objectives. Based on comments, reviewers were convinced of our likelihood for success here, and we therefore scale back our description of sequencing and assembly in this revised proposal and more explicitly highlight our scientific objectives (Aims 2 and 3). Although our hypotheses and overall goals remain the same, there are a few important revisions of our research activities. Based on significant advances in *de novo* genome sequencing strategies in the past year, our sequencing strategy is revised (Section G.1), resulting in significantly improved efficiency. These improvements enable additional genomes to be generated from wild experimental populations (Section G.1) and expansion of candidate validation (section G.3) to leverage more rigorous testing of hypotheses. This will considerably accelerate the pace of (and confidence in) candidate gene discovery (G.2.4). Specific aims are revised to more clearly outline our hierarchical approach to candidate gene discovery, candidate validation, and hypothesis testing.

Regarding broader impacts, though reviewers recognized that the genome would serve as an important resource to a broad research community, they felt that development of this section could be considerably improved, especially by highlighting how this study would aid in student training. We therefore scale back our description of the diverse benefits to the research community in favor of expanding description of student training. Inclusion in the annotation workshop, and in the community annotation effort, will be a valuable training opportunity for graduate students and postdoctoral fellows. We expand on this training in our "Broader Impacts" section. Furthermore, we seek support for inclusion of three undergraduate students in workshops and community annotations, who would leverage this training to initiate independent research projects. Undergraduate research programs at LSU sponsored by the Louisiana Alliance for Minority Participation will be used to recruit these students.

C. RESULTS from PREVIOUS NSF SUPPORT

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. All of 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 collected genetic samples from all experimental populations. Much of this work is ongoing, and described in section G. This grant has supported four publications [2-5], a fifth in press [6], a sixth in review [7], and others in preparation.

An award to JRS as Co-PI (0221837) supported a project that coupled lab and field studies to explore the evolution of metal tolerance in the micro-crustacean sentinel, *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 *D. pulex,* which are now community resources including sequencing its transcriptome, constructing a database to disseminate the sequence information (wFleabase; http://wfleabase.org/), and developing micorarrays for functional genomics investigations of this species complex. In addition to research findings, this grant helped establish the *Daphnia* Genomics Consortium (http://daphnia.cgb.indiana.edu/), contributed to the sequencing of the first crustacean genome and accompanying database, 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 [8-13], and it played a significant role in attracting a special issue of *Science* scheduled for release in 2010 that will launch the *Daphnia* model system and will be supported by the concordant publication of over 40 companion papers in BioMed Central.

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 [1,14-36], many in top-tier journals, and four are being revised for submission [37-40]. These grants also supported training for three post-doctoral fellows, four Ph.D. students (two minority), one Masters student, and six undergraduates (five minority).

D. PRELIMINARY DATA

Preliminary data that will assist with genome assembly and annotation and used to interrogate the final genome sequence are discussed in more detail throughout this proposal but are briefly summarized here. Extensive EST sequence data, coupled with robust bioinformatics [22] and annotation algorithms designed specifically for *F. heteroclitus* [41], are available to assist with genome assembly, gene finding, and annotation. Importantly, 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).

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 next-generation sequence (unpublished data).

Much of the data necessary to interrogate the genome to identify candidate genes are already in hand or are currently being produced through support from other sources by the PIs and collaborators. These data include QTL mapping data, population genome scan data, and comparative transcriptomics data and are described in more detail in section G.2. These multiple lines of experimental genome-scale data will be mapped to the reference genome in order to test hypotheses 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* mutations.

E. BACKGROUND and SIGNIFICANCE

Population responses following dramatic environmental changes could span the continuum from extirpation to evolutionary adaptation. Species will vary in their ability to persist. 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 goal that lies at the forefront of evolutionary research, and has been the subject of a number of recent high-profile reviews [42-46]. 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 will serve as an anchor for transcriptomic and mapping data to identify the genomic architecture and candidate genes underlying the convergent evolution of extreme stress tolerance (Aims 2 and 3)**. These data will be used to 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.**

E.1. The genetic basis of adaptive variation

Identifying the genomic elements that underlie adaptive traits is a major research ambition and remains a significant challenge though 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 [47] and [44]), it is clear that both types of variation can be important. For example, protein variation accounts for pigmentation variation in mice [48], birds [49], and lizards [50], cold tolerance in antarctic fishes [51], resistance to newt toxin in snakes [52], and hypoxia tolerance in mammals and birds [53,54]. In contrast, regulatory variation plays a role in eye degeneration in cavefish [55], developmental variation in beak morphology in finches [56,57], and pigmentation variation in fruit flies [58,59]. Notably, both regulatory and structural variations explain adaptive patterns of enzyme divergence among clinally distributed populations of *F. heteroclitus* (see review in [1]).

Since regulatory and structural variation may underlie the environmental pollution tolerance in *F. heteroclitus*, we have collectively accumulated evidence for both structural (QTL and population mapping) and regulatory (mapping and transcriptomics) differences. These mapping and functional 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.

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 or even the same mutations 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 [60,61], loss of pelvic structures in multiple stickleback species and also in manatees [62,63], and adaptive coat color variation in beach mice [48] and other animals including lizards, birds, cats, and mammoth (see review in [42]). 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 [48]. Similarly, pigment loss and eye loss have convergently evolved through different genetic mechanisms in Mexican cavefish populations [64] as has pigment gain in rock mouse populations [65]. Clearly, both similar and different genetic mechanisms can underlie convergent phenotypes, whether or not convergent lineages are closely or distantly related [42].

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 [66]. This is the likely mechanism whereby convergent morphological types have evolved in stickleback fish [60]. 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 important for mapping data from multiple lines of evidence (QTL mapping, population genome mapping, comparative transcriptomics), and from multiple convergent populations, 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 system 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 compared to other species. Yet, *F. heteroclitus* in unpolluted waters is relatively sensitive among fish species to this stress [67,68].

2. This adaptive phenotype has evolved over an extraordinarily short time period. The polluted sites where tolerant populations reside have been contaminated with tolerance-associated chemicals (primarily PCBs and PAHs) for only dozens of generations [69].
3. This extraordinary tolerance in *F. heteroclitus* has evolved independently at least four times [1,70,71].

These independent *F. heteroclitus* populations provide a powerful comparative system for determining the genomic basis of rapid adaptation, the repeatability of evolutionary change [42] and the relative roles of selection on standing genetic variation *versus* selection on *de novo* mutation [72]. Few other vertebrate systems offer insight into mechanisms of adaptive population divergence in response to rapid environmental change. These changes are relevant given the pace of current climate change.

F. heteroclitus has long been known to be highly adaptable to changes in environmental conditions [73-75]. Evolved tolerance or resistance to pollutants was demonstrated first for methylmercury [76,77] and later for several structurally related aromatic hydrocarbons, including chlorinated dioxins (Newark, NJ [78,79]), PCBs (New Bedford Harbor, MA [80,81]), and PAHs (from creosote; Elizabeth River, VA [68,82,83]).

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 [68,69]). Nevertheless, several key features of the resistant populations have been identified, demonstrating the potential for new insights to be facilitated by genomic studies. (i) Pollutant tolerance involves a variety of endpoints, including embryo and adult survival, teratogenicity, and altered gene expression [68]. (ii) 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 [84]. PCB-resistant New Bedford fish are also resistant to PAHs [80,81]. PAH-resistant Elizabeth River fish are also resistant to PCBs [85]. This cross-tolerance suggests that the mechanisms of resistance may converge on a common biochemical pathway, such as the aryl-hydrocarbon receptor pathway. (iii) The tolerance appears to involve both heritable and non-heritable mechanisms [68,80,83,85,86]. (iv) The tolerant populations do not show an overall loss of genetic diversity, as assessed by allozyme and molecular methods [87-90]. However, the tolerant populations are genetically distinct from nearby sensitive populations and do show evidence for selection at specific loci [89,91-93]. (v) 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 [67]. Thus, pollutant tolerance is a widespread phenomenon. (vi) 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 [68,94] (but see [95]). A genome sequence will facilitate 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.

E.4. State of readiness for genome sequencing

Based on flow cytometry and bulk fluorometric analysis, the *F. heteroclitus* haploid genome size is between 1.29 and 1.5 billion bp [96-98] which is intermediate in size between Japanese medaka and zebrafish [99]. We will sequence one individual from a stock of fish that has been inbred for ten generations to minimize heterozygosity. This will maximize the efficiency of reference genome assembly. Additional lower-coverage genomes from experimental outbred populations will be useful for SNP discovery. Other genome projects have demonstrated that successful assembly does not depend on sequencing from a purely homozygous inbred line [100-104]. Over 1.5 million EST sequences (77,000 in NCBI, remainder from Joe Shaw, unpublished data) will aid with assembly and gene-finding and be used for adding UTR information to genes identified through protein homology (see sections G.1.3, G.1.4). A novel and highly robust gene annotation pipeline has been developed for *F. heteroclitus* EST collections [41], and will serve to correctly annotate 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 and aphid genome projects and will be adapted for the *F. heteroclitus* genome (see attached letter of collaboration from John Colbourne). Ongoing projects in the labs of A. Whitehead, P. Schulte (University of British Columbia; see letter of collaboration), and M. Bagley (US EPA) are generating genetic maps from *F. heteroclitus* QTL mapping experiments 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* [105], and may enable more detailed study of adaptive variants in future studies.

F. BROADER IMPACTS

At the final stage of genome annotation, the Indiana University CGB will host a collaboration wiki, web-based conferences, and an annotation training workshop for members of the *Fundulus* Genomics Consortium (FGC). 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.** It also 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.

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 (see attached letters of support).

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 [42-46]. 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 (10kb) 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. An additional eight genomes from experimental populations will be sequenced at 25X genome coverage using short insert paired-end reads. These additional genomes will be of fish representing 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). Selection of individuals from these experimental populations for sequencing will include experimental verification that individuals are indeed tolerant or sensitive (and therefore representative of the population in terms of tolerance) using published methods (adapted from [106]). These experiments will be performed by collaborator Dr. Diane Nacci (US EPA – see attached letter of collaboration). Dr. Nacci currently has appropriate facilities and biological resources and will characterize these individuals for sequencing as part of ongoing studies.

G.1.1. DNA sequencing

Next-generation sequencing technologies have proven to be cost effective for whole genome assemblies, most recently with short reads [107-109]. 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 base pair) paired reads plus 5x coverage with 3kb and 10kb paired reads each. Although shorter contigs are used relative to the traditional Sanger-based assemblies, we expect 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 contig N50 values of 40kb and 3Mb, respectively, were achieved from short Illumina reads [107]. Since the *Fundulus heteroclitus* genome is smaller and contains fewer predicted repeats we expect assembly contiguity to be sufficient for accurate gene predictions. The mean length of unspliced genes downloaded from the medaka database at Ensembl is ~10 kb. Our goal will be to meet or exceed this contig size defined by mean gene length. 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.

G.1.2. Genome assembly

The SOAP assembler, version 1.04, will be utilized to assemble all read types using an iterative process [107,108]. 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 by Flicek and Birney [110]). 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. Other assemblers are available for short reads, but only the SOAP tool has achieved the contiguity necessary for downstream analysis at this point.

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 [111]. 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 [112] 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, personal communication). Following assembly, RAD tag and microsatellite primer sequences, coupled with genetic mapping data, will assist in large scaffold construction and improve order inconsistencies. For scaffolds that cannot be joined by any of the above procedures, we will screen multi-species genome alignments (including puffer fish, zebrafish, stickleback, cichlid and Japanese medaka) for conserved gene orders and orientations to enable final genome 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 [113] 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 [114], Genewise family models [115] and SNAP [116]. Colleagues at the NCBI RefSeq Project Group will provide RefSeq transcript alignments [117] 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 [118] 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 [119], eukaryotic orthologous groups [120], and KEGG metabolic pathways [121].

One of the most challenging elements of a genome sequencing project is functional annotation, which is essential for extracting biological significance from the vast amounts of newly acquired sequence information [122,123]. 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 experiences with the waterflea (*Daphnia;* (http://conferences.cgb.indiana.edu/daphnia2007/index.html), jewel wasp and pea aphid (https://dgc.cgb.indiana.edu/display/aphid/Workshop+I) 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 [122,123].

The annotation workshop will be 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 [122,124].

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; [125]) components and open source software shared with other genome databases. The genome access tools of GMOD (GBrowse [125], BioMart [126] and BLAST [127]) will be available for searching the *F. heteroclitus* genome. Another aid to integrating and mining these data is GMOD Lucegene (www.gmod.org/lucegene/) that forms a core component for rapid data retrieval by attributes, GBrowse data retrieval, and databank partitioning for Grid analyses. 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 genome database and access tools will be available for annotation of the *F. heteroclitus* genome through the application of the Apollo annotation viewer and editor, which was developed for annotating the *Drosophila melanogaster* genome and is the annotation workbench adopted by GMOD [128]. Correct gene annotation will be aided by implementing a novel filtering algorithm. This routine performs sequential homology searches against multiple databases to verify similar BLAST hits, and integrates protein motif searches using machine learning techniques to further direct accurate functional annotations [41]. 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 [41].

The computationally intense analysis we propose will benefit from the TeraGrid project (www.teragrid.org), which is part of a shared cyber infrastructure for sciences, funded primarily by NSF. 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 [43], some maintain that regulatory elements are most likely to underpin new phenotypes [47], whereas others contend that protein variation accounts for much of adaptive variation [44]. 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).
4. Amino acid and promoter region substitution patterns among genomes from eight experimental populations (G.2.4).

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 genome sequences (G.2.4) to facilitate closing of this gap.

G.2.1. QTL mapping data

In a recent review, Stinchcombe and Hoekstra [46] outline 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 [71] 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 dioxin sensitivity, and genotyping and genetic mapping is ongoing by the US EPA (microsatellite markers) and Whitehead lab (RAD markers).

We are using a novel ultra-high throughput method to generate thousands of markers for genome-wide genotyping. Briefly, restriction site associated DNA (RAD) markers are 45-base Illumina-generated sequence tags that are associated with restriction sites [112,129]. 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 [129], thereby validating data from laborious AFLP and microsatellite mapping [60]. Over 13,000 SNPs were identified in the stickleback studies [112], 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). Using this approach, we can characterize the number of unique locations in the genome identified by QTL markers, test whether similar genomic regions are implicated in independently derived tolerant populations, test for associations between QTL loci and genes implicated from comparative transcriptomics studies, and test for associations between QTL loci and genes with non-synonymous substitutions from comparisons of

experimental genomes, and thereby help direct the selection of genomic regions from which candidate mutations may ultimately be identified.

To anchor markers to the genome, we will conduct basic local alignment search tool (BLAST) searches [130] using the RAD markers as query sequences and the genome as the database sequence. We will use a threshold of at least $e < 10^{-10}$ to identify significant hits in the genome, and 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 genes with non-neutral patterns of variation among genomes from experimental populations (section G.2.4).

G.2.2. Population mapping data

For identifying the genomic basis of adaptive traits, population genomic data are naturally complementary to QTL mapping [43,46]. 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 [131-133]. 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 three "population genomics" data sets 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.

Markers from three population data sets (AFLP, SNP and RAD markers) will be mapped to the genome. 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 [71]. 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 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 between all three tests. Only one SNP was identified as an outlier in all tests and among all three tolerant populations. These data indicate 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. To test this, we will anchor relevant markers to the genome sequence and identify marker associations to each other and to QTL loci and candidate genes.

A third population genomics study is using the same RAD markers as those used for QTL analysis. These experiments are being conducted in collaboration with Diane Nacci (US EPA) whose group has collected appropriate genetic samples including DNA from 60 field-caught individuals from each of four tolerant populations (two of which are subjects for QTL studies) and eight sensitive reference populations. Sixty individuals from each population are planned for genotyping at thousands of Illumina-sequenced RAD tag loci. We have already genotyped 24 individuals (unpublished data) and are currently optimizing the library preparation protocol before further genotyping. Since QTL analyses are limited to a relatively small number of meioses from F2 families, these population samples will allow for greater scope for inference for QTL-identified loci. Furthermore, we can screen many more populations than is feasible for QTL experiments, since the generation of F2 families requires at least four years of husbandry.

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 [71,131,132,134-137]. 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 [138], 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.

Again, markers implicated from the above approaches will be mapped to the genome using BLAST. These population genome scan approaches will be useful for confirming that 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.

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.

Collaborator Diane Nacci (US EPA) has completed dose-response experiments where embryos derived from tolerant and sensitive populations (raised in a common clean environment for at least two generations) have been challenged with a range of PCB doses. The Whitehead lab has extracted RNA from these embryos and hybridized it to a 7,500 gene oligonucleotide microarray. These comparative data are in the midst of analysis, but preliminary results indicate intriguing contrasts between sensitive and tolerant populations, for example, in regulatory control of the *Wnt* signaling pathway which is known to cross-talk with the Ah receptor pathway [139], and regulatory control of haemopoietic and heart morphogenesis pathways, and gene expression appears highly correlated with the appearance of developmental abnormalities. Similar experiments were completed in the summer 2009 by the US EPA that include other tolerant populations and paired sensitive reference populations, to test whether tolerant populations have converged on similar compensatory responses to PCBs.

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). Preliminary experiments with a targeted metabolic array have found that up to 17% of metabolic genes have evolved adaptive changes in gene expression. RNA samples are ready to be hybridized to a ~7,000 gene cDNA array (spring 2010). This experiment was designed to differentiate both evolved and physiologically induced changes in gene expression in tolerant *versus* sensitive populations. The Oleksiak group also has completed dose-response experiments in collaboration with Rich DiGiulio (Duke University; see attached letter of collaboration) to explore effects of polluted sediment extracts 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. Gene expression results using a 7,000 gene array identify genes with significant differences in expression due to population by treatment interactions, treatment alone, and population alone. We found a strong correlation (82%) between the severity of morphological deformities and cardiac physiology among sensitive embryos as well as a striking relationship between the deformity index and differential expression of genes. Several genes whose expression is significantly different due to morphology play important roles in heart development, physiology, and major metabolic processes during embryogenesis.

The Hahn group has conducted a series of studies comparing targeted gene expression and receptor polymorphisms in killifish from the New Bedford Harbor (NBH; MA) Superfund site and a reference site (Scorton Creek (SC), MA) [81,92,140-142]. In an ongoing collaboration with Oleksiak, Hahn's group has exposed F1 embryos from the two sites to a dioxin-like PCB and sampled at three different stages of development. Targeted analysis of gene expression (CYP1A, AHRR) by qRT-PCR has demonstrated that the NBH fish remain highly resistant, despite recent clean-up efforts at the site. Analysis of the embryo RNA samples using a 7,000 gene cDNA array revealed dramatic population-specific differences in responsiveness to the PCB at all stages. In a second set of experiments, Hahn's group is examining the costs of PCB resistance in terms of its effect on sensitivity to other stressors, such as environmental hypoxia. F1 embryos and larvae from SC and NBH were exposed to normoxia or two levels of hypoxia, in the presence or absence of PCB. Targeted analysis of PCB-inducible (CYP1A) and hypoxia-inducible (IGFBP-1) genes demonstrated population-specific differences in responsiveness that mirrored the enhanced hypoxia sensitivity of the PCB-resistant fish. Transcriptional profiling of these samples using the 7,000 gene cDNA array will be performed early in 2010, in collaboration with Oleksiak.

These data on evolved differences among populations and induced differences in gene expression due to chronic exposure to pollutants are greatly enhanced with the *Fundulus* genome. The first goal is to test for association between mapped loci and genes with evolved differences in gene expression. Additionally, we will determine if the architecture of genes with evolved differences in gene expression are different from genes that lack this difference. For example, are these genes more or less likely to lack TATAA sites, contain GC-boxes, or have a higher density of known binding sites? Finally, we will determine if physiologically induced patterns of gene expression share critical transcription factor binding sites (e.g., AhR or HIF binding sites). For instance, the expression of approximately 33% of

genes is highly correlated in *F. heteroclitus* and show significant positive and negative correlations for both proteins and mRNAs [31,39]. Some of these genes have patterns of expression indicative of natural selection [19,28] and some are strongly associated with variation in metabolism [24]. We suggest that genes with positively correlated expression will be enriched for the same transcription factor binding sites (TFBS) and these sites will have patterns of sequence variation indicative of stabilizing selection. Putative TFBS will be identified using FootPrinter [143] and MULAN [144,145] and sites confirmed by comparison to TRANSFAC [146-149]. 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. Comparative genomics data

The eight genomes from experimental populations generated in Aim 1 will be used to buttress evidence for genes associated with adaptation, to aid in candidate gene prioritization, and accelerate the overall pace of discovery and hypothesis testing. Using the reference genome, the alignment of reads from each respective experimental sample is the obvious first test to screen for polymorphisms (for example, that alter amino acid sequence) that are shared only among tolerant genomes, thereby rejecting the neutral expectation that geographically paired sensitive and reference populations should share most variation [71]. Such loci, especially if implicated by other lines of evidence (G.2.1 – G.2.3), would be strong candidates.

The WUGC cancer genome analysis pipeline is established for variant detection for both targeted capture and whole genome sequencing [150]. All predicted somatic variants will be annotated with gene and function classes. These data will be a source of structural variation (SV) data as well. We have developed BreakDancer for genome-wide detection of SVs including deletions, insertions, and inversions [151] which has been successfully used for predicting all five types of SVs in cancer samples. Copy number alterations (CNAs), another source of gene variation, can also be detected using our internal hidden Markov model following GC- and mappability-corrections. As with the mutation data, SVs and CNAs will be annotated with known protein-coding and RNA genes, conserved regions, and regulatory elements. Our data release policies for the proposed work will require us to submit all reads from the eight experimental genomes into the short read trace archive consistent with NHGRI requirements. We look forward to continuing to work with our colleagues at the other centers to plan, produce and understand these data, and to share our results with the research community.

Because adaptation was rapid (dozens of generations), strict reliance on d_N/d_S ratio tests among genomes for identifying genes under strong selection may be of limited power [152]. Rather, the power of these additional genomes is to significantly enhance the scope for inference from other lines of evidence. For example, mapping data may implicate a genome region spanning many genes. If comparative genome data indicate that some of those genes lack non-synonymous changes, whereas other genes do (especially if they are shared among tolerant genomes), then this would greatly aid in prioritizing genes for further validation through population screening (G.3).

G.2.5. 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. 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 genome sequence data, and what we know about functionally relevant genes from the literature. For example, if the linkage block identified using mapping markers includes genes with divergent patterns of gene expression between tolerant and sensitive populations, this would strengthen candidacy. Importantly, comparative genome sequence data from experimental populations (Aim 1) would be critically useful for ruling out or supporting candidacy; that is, if candidate genes within linkage blocks do not include non-synonymous substitutions (or transcription factor biding site substitutions in promoters) between tolerant and sensitive populations whereas others genes do, then this evidence would help prioritize candidacy. Where candidate regions appear shared only among tolerant populations, this would be the strongest evidence for candidacy, and contribute to testing our hypothesis that the same ancestral mutations are responsible

for convergent evolution among tolerant populations. This hypothesis will be further tested by candidate gene screening in experimental population samples in Aim 3 (section G.3).

G.3. Specific Aim 3 – Validation of candidate genes

We expect to generate a prioritized list of candidate genomic regions (from Aim 2). Our proposed method of population screening is easily scalable from including dozens to hundreds of candidates. We propose to screen patterns of variation in the sequences and promoters of these candidate genes among 20 individuals from each of the eight experimental tolerant and sensitive populations identified in section G.1. DNA has already been isolated from at least 20 individuals from each of the eight populations (tissues provided by Diane Nacci, US EPA). We will use population genetic tests to verify or rule out whether candidates show adaptive patterns of polymorphism. Multiple tests, each with strengths and weaknesses [43,152], will be brought to bear, including Tajima's D [153], McDonald-Kreitman test [154], detection of F_{ST} outliers [132,134,137], and possibly selective sweep [138] or linkage disequilibrium [155] tests. Additionally, we will return to DNA samples from QTL mapping families to verify that candidate polymorphisms segregate between tolerant and sensitive family members.

Our screening approach will use DNA capture microarrays to enrich genomic DNA fragment libraries for target genomic regions, coupled with Illumina sequencing of enriched libraries [156,157], to sequence 20 individuals from each of eight experimental tolerant and sensitive populations, and a subset of individuals from QTL mapping families (20 of the most tolerant, and 20 of the most sensitive individuals). Briefly, overlapping microarray probes are designed to tile target candidate genome regions. The Nimblegen capture arrays include from 385K to 2.1M probes (60-90 bases each). These arrays enable capture of up to 34Mb of contiguous or dispersed target DNA and therefore afford high scalability depending on the number and size of candidate regions we identify. Nimblegen capture arrays are already in use at the Indiana University CGB. Personal communication (AW) with Nimblegen design teams have indicated their willingness to collaborate in designing custom capture arrays for *Fundulus*. An alternative is to use solution-based capture technologies for fragment library enrichment (for example, from Agilent). Since these technologies are evolving so quickly, final enrichment platform decision will be based on the latest data regarding efficiency, enrichment yields, and enrichment consistency. Genomic DNA from each individual is fragmented (by sonication) then individual barcoded Illumina adapters are ligated. Individuals with different barcodes can be combined into a single library to enable multiplexing and significant increase in efficiency. Our lab (AW) has already tested a set of barcoded adapters enabling multiplexing of 64 individuals (unpublished data). Multiplexed fragment libraries are then enriched for target regions using the sequence capture arrays, and enriched libraries sequenced using the Illumina platform. Even if targets are many (100) and large (5,000 bases each) and we seek high (50X) coverage per fragment per barcode, we theoretically anticipate being able to multiplex 100 individuals, but more conservatively 50 individuals per Illumina channel. Therefore, to screen 200 individuals (20 individuals X 8 populations, + 40 individuals from QTL families) would require five libraries to each be enriched and sequenced, which would consume five capture arrays and less than one full Illumina flow cell (seven channels per flow cell).

These population data are expected to strongly implicate only a few genes of major effect, considering the rapidity and repeatability of derived tolerance. To test our second major hypothesis, that adaptive mutations are present at low frequency in ancestral populations and repeatedly swept to fixation in each pollution-exposed population, we propose to screen sensitive populations for adaptive alleles. We will re-design our DNA capture probes to cover each of our adaptive genes. Solution-based capture techniques may be more appropriate here due to the relatively small number of probes needed. To detect a minimum population frequency of 0.5% (the adaptive *Eda* allele was detected at 3.8% frequency in ancestral stickleback populations [60]), we propose to pool genomic DNA from 100 individuals from each of the four sensitive populations (1 pool of 100 individuals per population). Each population pool will be barcoded, enriched for target sequences using sequence capture, and the enriched barcoded pool multiplex sequenced in one Illumina channel. Sequence counts for the adaptive allele compared to other alleles should offer an estimate of frequency in the population. For example, if sequence fragments containing the adaptive allele are counted in the sequencing pool at $1/100^{th}$ the frequency of other alleles, this would indicate that the adaptive allele is present in the population at 1% frequency.

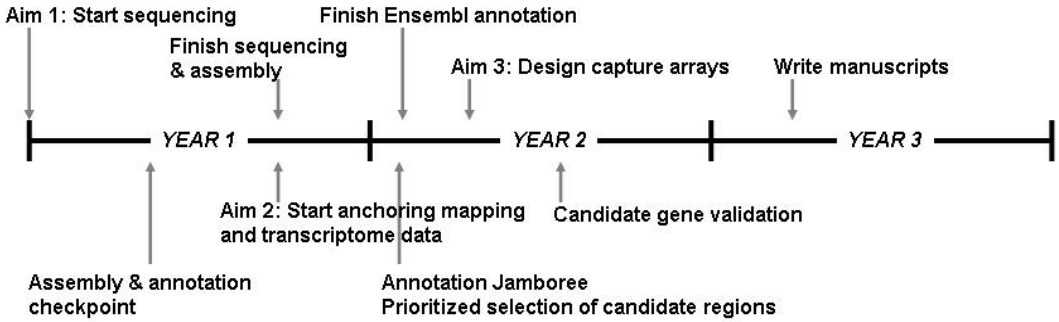
Together, patterns of polymorphism of candidate genes among eight populations and among individuals from mapping families will provide a rich data set with which to rigorously test our hypotheses.

If our hypotheses are true, that the same adaptive variants present at low frequency in ancestral populations repeatedly swept to fixation in independently derived tolerant populations, we expect to find the same candidate adaptive alleles fixed in each tolerant population, associated with tolerant and sensitive phenotypes in mapping families, and appearing at low but detectable frequency in sensitive populations (as is the case for convergent adaptive variation in sticklebacks [60]). Alternatively, adaptive variants may include different mutations, in the same or different genes, in each of the tolerant populations, indicating repeated independent origins of adaptive variation. Here, adaptive mutations may be present at low frequency in each paired sensitive population indicating selection on standing variation, or undetectable indicating selection on *de novo* mutations (or present but at a frequency below our limit of detection). A third possibility (a hybrid of the above two) is that some combination of tolerant populations (for example, those north of the phylogeographic break at New Jersey) may share adaptive mutations, whereas other loci may explain adaptive variation in remaining tolerant populations, as is the case for convergent adaptive pigmentation variation in cavefish [158] and coat color variation in mice [48,65].

H. FUTURE STUDIES

Though genome and population data will provide very strong evidence for genes associated with rapid, convergent adaptation, the gold-standard for demonstrating adaptive relevance is functional studies. Such studies could include gene knockdown experiments and functional enzyme and promoter experiments. Our group is experienced in promoter function and enzyme function assays [14,18,81,140,159-172] and morpholino knockdown technology has been successfully adapted for use in *Fundulus* [105]. These tools position our group well to capitalize on the data generated in this study for future functional experiments.

I. TIMELINE



J. MANAGEMENT PLAN

AW will oversee selection of genomes 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 and JS will organize the annotation workshop, and everyone will participate. AW, WW, JS, MO, DC, and MH will form the annotation steering committee. AW, MH, and MO will contribute experimental data, anchor their respective mapping and transcriptomics data sets to the genome, and assemble a prioritized list of candidate genes (Aim 2). AW and JS will screen population samples for patterns of candidate gene polymorphism (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.

K. REFERENCES

1. Burnett KG, Bain LJ, Baldwin WS, Callard GV, Cohen S, et al. (2007) *Fundulus* as the premier teleost model in environmental biology: Opportunities for new insights using genomics. Comparative Biochemistry and Physiology, Part D 2: 257-286.
2. Cheviron ZA, Whitehead A, Brumfield RT (2008) Transcriptomic variation and plasticity in rufous-collared sparrows (*Zonotrichia capensis*) along an altitudinal gradient. Molecular Ecology 17: 4556-4569.
3. Duvernell DD, Lindmeier JB, Faust KE, Whitehead A (2008) Relative influences of historical and contemporary forces shaping the distribution of genetic variation in the Atlantic killifish, *Fundulus heteroclitus*. Molecular Ecology 17: 1344-1360.
4. Triant DA, Whitehead A (2009) Simultaneous extraction of high-quality RNA and DNA from small tissue samples. J Hered 100: 246-250.
5. Whitehead A (2009) Comparative mitochondrial genomics within and among species of killifish. BMC Evolutionary Biology 9: -.
6. Whitehead A (in press) The evolutionary radiation of diverse osmotolerant physiologies in killifish (*Fundulus* sp.). Evolution.
7. Whitehead A, Galvez F, Zhang S, Williams LM, Oleksiak MF (in review) Functional genomics of physiological plasticity and local adaptation in killifish.
8. Colbourne JK, Eads BD, Shaw J, Bohuski E, Bauer DJ, et al. (2007) Sampling *Daphnia's* expressed genes: preservation, expansion and invention of crustacean genes with reference to insect genomes. BMC Genomics 8: -.
9. Colbourne JK, Singan VR, Gilbert DG (2005) wFleaBase: The *Daphnia* genome database. Bmc Bioinformatics 6: -.
10. Cook JC, Denslow ND, Iguchi T, Linney EA, Miracle A, et al. (2006) “Omic” approaches in the context of environmental toxicology. In: Benson WH, DiGulio R, editors. Genomic Approaches for Cross-Species Extrapolation in Toxicology. Boca Raton, FL: CRC Press.
11. Denslow ND, Colbourne JK, Dix D, Freedman J, Helbing C, et al. (2006) Selection of surrogate animal species for comparative toxicogenomics. In: Benson WH, DiGulio R, editors. Genomic Approaches for Cross-Species Extrapolation in Toxicology. Boca Raton, FL: CRC Press.
12. Shaw J, Colbourne J, Davey J, Glaholt S, Hampton T, et al. (2007) Gene response profiles for Daphnia pulex exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC Genomics 8: 477.
13. Shaw JR, Pfrender ME, Eads BD, Klaper R, Callaghan A, et al. (2008) *Daphnia* as an emerging model for toxicological genomics. In: Hogstrand C, Kille P, editors. Advances in Experimental Biology: Comparative Toxicogenomics. London: Elsevier Science.
14. Podrabsky JE, Javillonar C, Hand SC, Crawford DL (2000) Intraspecific variation in aerobic metabolism and glycolytic enzyme expression in heart ventricles. American Journal of Physiology-Regulatory Integrative and Comparative Physiology 279: R2344-R2348.
15. Crawford DL (2001) Functional genomics does not have to be limited to a few select organisms. Genome Biology 2: interactions1001.1001–interactions1001.1002.
16. Oleksiak MF, Kolell KJ, Crawford DL (2001) Utility of natural populations for microarray analyses: Isolation of genes necessary for functional genomic studies. Marine Biotechnology 3: S203-S211.
17. Crawford DL (2002) Evolution of physiological adapatation. In: Storey KB, Storey JM, editors. Cell and Molecular Responses to Stress. NY: Elsevier Publishing.
18. Kolell KJ, Crawford DL (2002) Evolution of Sp transcription factors. Molecular Biology and Evolution 19: 216-222.
19. Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nature Genetics 32: 261-266.
20. Oleksiak MF, Crawford DL (2002) 5' genomic structure of human Sp3. Molecular Biology & Evolution 19: 2026-2029.
21. Clark MS, Crawford Douglas L, Cossins A (2003) Worldwide genomic resources for non-model fish species. Comparative and Functional Genomics 4: 502-508.
22. Paschall JE, Oleksiak MF, VanWye JD, Roach JL, Whitehead JA, et al. (2004) FunnyBase: a systems level functional annotation of *Fundulus* ESTs for the analysis of gene expression. BMC Genomics 5: 96.

23. Cossins AR, Crawford DL (2005) Fish as models for environmental genomics. Nature Reviews Genetics 6: 324-333.
24. Oleksiak MF, Roach JL, Crawford DL (2005) Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. Nature Genetics 37: 67-72.
25. Whitehead A, Crawford DL (2005) Variation in tissue-specific gene expression among natural populations. Genome Biology 6: -.
26. Oleksiak M, F., Crawford Douglas L (2006) Functional Genomics in Fishes, Insights into Physiological Complexity. In: Evan D, Claiborne J, editors. The Physiology of Fishes. Boca Raton: CRC Press. pp. 523-550.
27. Richardson DE, VanWye JD, Miyake AM, Cowen RK, Crawford DL (2006) High-throughput species identification: from DNA isolation to bioinformatics. Molecular Ecology 7: 199-207.
28. Whitehead A, Crawford DL (2006) Neutral and adaptive variation in gene expression. Proceedings of the National Academy of Sciences of the United States of America 103: 5425-5430.
29. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. Molecular Ecology 15: 1197-1211.
30. Crawford DL (2007) Human reference sequence makes sense of names. Nature 447: 142.
31. Crawford DL, Oleksiak MF (2007) The biological importance of measuring individual variation. Journal of Experimental Biology 210: 1613-1621.
32. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology 17: 1636-1647.
33. Williams DA, Brown SD, Crawford DL (2008) Contemporary and historical influences on the genetic structure of the estuarine-dependent Gulf killifish Fundulus grandis. Marine Ecology-Progress Series 373: 111-121.
34. Everett MV, Crawford DL (in press) Adaptation versus allometry: Population and body mass effects on hypoxic metabolism in *Fundulus grandis*. Physiological and Biochemical Zoology.
35. Scott CP, VanWye J, McDonald MD, Crawford DL (2009) Technical analysis of cDNA microarrays. PLoS ONE 4: e4486.
36. Scott CP, Williams DA, Crawford Douglas L (2009) The effect of genetic and environmental variation on gene expression Molecular Ecology 18: 2832-2843.
37. Everett MV, Crawford DL (in preparation) Time course for hypoxic induction of mRNA expression
38. Oleksiak M, F., Crawford DL (in review) Inter and intra-specific variation in cardiac metabolism using three different substrates.
39. Rees BB, Crawford Douglas L (in preparation) Individual variation in protein expression: a protemic analysis of gene expression
40. Young S, Crawford D, L. (in review) Improved Annotations with FuzzyMatch and Incremental p-values.
41. Young S (2008) Improved EST Annotation Using Multiple External Databases. Raleigh: North Carolina State University.
42. Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends in Ecology & Evolution 23: 26-32.
43. Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. Nature 452: 169-175.
44. Hoekstra HE, Coyne JA (2007) The locus of evolution: Evo devo and the genetics of adaptation. Evolution 61: 995-1016.
45. Hoffmann AA, Willi Y (2008) Detecting genetic responses to environmental change. Nat Rev Genet 9: 421-432.
46. Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100: 158-170.
47. Carroll SB (2005) Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom. New York: Norton. xi, 350 p. p.
48. Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. Science 313: 101-104.
49. Mundy NI, Badcock NS, Hart T, Scribner K, Janssen K, et al. (2004) Conserved genetic basis of a quantitative plumage trait involved in mate choice. Science 303: 1870-1873.

50. Rosenblum EB, Hoekstra HE, Nachman MW (2004) Adaptive reptile color variation and the evolution of the Mc1r gene. Evolution 58: 1794-1808.
51. Chen LB, DeVries AL, Cheng CHC (1997) 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 94: 3817-3822.
52. Geffeney SL, Fujimoto E, Brodie ED, Brodie ED, Ruben PC (2005) Evolutionary diversification of TTX-resistant sodium channels in a predator-prey interaction. Nature 434: 759-763.
53. Jessen TH, Weber RE, Fermi G, Tame J, Braunitzer G (1991) 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 88: 6519-6522.
54. Storz JF, Sabatino SJ, Hoffmann FG, Gering EJ, Moriyama H, et al. (2007) The Molecular Basis of High-Altitude Adaptation in Deer Mice. PLoS Genetics 3: e45.
55. Yamamoto Y, Stock DW, Jeffery WR (2004) Hedgehog signalling controls eye degeneration in blind cavefish. Nature 431: 844-847.
56. Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, et al. (2006) The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. Nature 442: 563-567.
57. Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ (2004) Bmp4 and morphological variation of beaks in Darwin's finches. Science 305: 1462-1465.
58. Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, et al. (2008) The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell 132: 783-793.
59. Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB (2009) Stepwise modification of a modular enhancer underlies adaptation in a *Drosophila* population. Science 326: 1663-1667.
60. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, et al. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307: 1928-1933.
61. Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, et al. (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. Plos Biology 2: 635-641.
62. Cresko WA, Amores A, Wilson C, Murphy J, Currey M, et al. (2004) 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 101: 6050-6055.
63. Shapiro MD, Bell MA, Kingsley DM (2006) Parallel genetic origins of pelvic reduction in vertebrates. Proceedings of the National Academy of Sciences of the United States of America 103: 13753-13758.
64. Wilkens H, Strecker U (2003) Convergent evolution of the cavefish *Astyanax* (Characidae, Teleostei): genetic evidence from reduced eye-size and pigmentation. Biological Journal of the Linnean Society 80: 545-554.
65. Hoekstra HE, Nachman MW (2003) Different genes underlie adaptive melanism in different populations of rock pocket mice. Molecular Ecology 12: 1185-1194.
66. Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. Trends in Ecology & Evolution 23: 38-44.
67. Nacci DE, Champlin D, Coiro L, McKinney R, Jayaraman S (2002) Predicting the occurrence of genetic adaptation to dioxinlike compounds in populations of the estuarine fish *Fundulus heteroclitus*. Environmental Toxicology and Chemistry 21: 1525-1532.
68. Van Veld PA, Nacci DE (2008) Toxicity resistance. In: Di Giulio RT, Hinton DE, editors. The Toxicology of Fishes. Boca Raton, FL: Taylor and Francis.
69. Nacci DE, Gleason T, Gutjahr-Gobell R, Huber M, Munns WRJ (2002) Effects of environmental stressors on wildlife populations. In: Newman MC, editor. Coastal and Estuarine Risk Assessment: Risk on the Edge. Washington, DC: CRC Press/Lewis Publishers.
70. Adams SM, Lindmeier JB, Duvernell DD (2006) Microsatellite analysis of the phylogeography, Pleistocene history and secondary contact hypotheses for the killifish, *Fundulus heteroclitus*. Molecular Ecology 15: 1109-1123.
71. Williams LM, Oleksiak MF (2008) Signatures of selection in natural populations adapted to chronic pollution. BMC Evolutionary Biology 8: -.
72. Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. Nature Reviews Genetics 8: 675-688.
73. Mitton JB, Koehn RK (1975) Genetic organization and adaptive response of allozymes to ecological variables in *Fundulus heteroclitus*. Genetics 79: 97-111.

74. Powers DA, Schulte PM (1998) Evolutionary adaptations of gene structure and expression in natural populations in relation to a changing environment: a multidisciplinary approach to address the million-year saga of a small fish. J Exp Zool 282: 71-94.
75. Schulte PM (2001) Environmental adaptations as windows on molecular evolution. Comp Biochem Physiol B Biochem Mol Biol 128: 597-611.
76. Weis JS, Heber M, Weis P, Vaidya S (1981) Methylmercury tolerance of killifish (*Fundulus heteroclitus*) embryos from a polluted vs non-polluted environment. Mar Biol 65: 283-287.
77. Weis JS, Weis P (1989) Tolerance and stress in a polluted environment. Bioscience 39: 89-95.
78. Prince R, Cooper KR (1995) Comparisons of the effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin on chemically impacted and nonimpacted subpopulations of *Fundulus heteroclitus*: I. TCDD toxicity. Environmental Toxicology and Chemistry 14: 579-587.
79. Prince R, Cooper KR (1995) Comparisons of the effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin on chemically impacted and nonimpacted subpopulations of *Fundulus heteroclitus*: II. Metabolic considerations. Environmental Toxicology and Chemistry 14: 589-595.
80. Nacci D, Coiro L, Champlin D, Jayaraman S, McKinney R, et al. (1999) Adaptations of wild populations of the estuarine fish *Fundulus heteroclitus* to persistent environmental contaminants. Marine Biology 134: 9-17.
81. Bello SM, Franks DG, Stegeman JJ, Hahn ME (2001) 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 60: 77-91.
82. Van Veld PA, Westbrook DJ (1995) Evidence for depression of cytochrome P4501A in a population of chemically resistant mummichog (*Fundulus heteroclitus*). Environmental Sciences 3: 221-234.
83. Meyer JN, Nacci DE, Di Giulio RT (2002) Cytochrome P4501A (CYP1A) in killifish (*Fundulus heteroclitus*): Heritability of altered expression and relationship to survival in contaminated sediments. Toxicological Sciences 68: 69-81.
84. Elskus AA, Monosson E, McElroy AE, Stegeman JJ, Woltering DS (1999) Altered CYP1A expression in *Fundulus heteroclitus* adults and larvae: a sign of pollutant resistance? Aquatic Toxicology 45: 99-113.
85. Meyer J, Di Giulio R (2002) Patterns of heritability of decreased EROD activity and resistance to PCB 126-induced teratogenesis in laboratory-reared offspring of killifish (*Fundulus heteroclitus*) from a creosote-contaminated site in the Elizabeth River, VA, USA. Marine Environmental Research 54: 621-626.
86. Ownby DR, Newman MC, Mulvey M, Vogelbein WK, Unger MA, et al. (2002) Fish (*Fundulus heteroclitus*) populations with different exposure histories differ in tolerance of creosote-contaminated sediments. Environmental Toxicology and Chemistry 21: 1897-1902.
87. Roark SA, Nacci D, Coiro L, Champlin D, Guttman SI (2005) Population genetic structure of a nonmigratory estuarine fish (*Fundulus heteroclitus*) across a strong gradient of polychlorinated biphenyl contamination. Environ Toxicol Chem 24: 717-725.
88. McMillan AM, Bagley MJ, Jackson SA, Nacci DE (2006) Genetic diversity and structure of an estuarine fish (*Fundulus heteroclitus*) indigenous to sites associated with a highly contaminated urban harbor. Ecotoxicology 15: 539-548.
89. Mulvey M, Newman MC, Vogelbein W, Unger MA (2002) Genetic structure of *Fundulus heteroclitus* from PAH-contaminated and neighboring sites in the Elizabeth and York Rivers. Aquatic toxicology 61: 195-209.
90. Mulvey M, Newman MC, Vogelbein WK, Unger MA, Ownby DR (2003) Genetic structure and mtDNA diversity of *Fundulus heteroclitus* populations from polycyclic aromatic hydrocarbon-contaminated sites. Environ Toxicol Chem 22: 671-677.
91. Cohen S (2002) Strong positive selection and habitat-specific amino acid substitution patterns in MHC from an estuarine fish under intense pollution stress. Mol Biol Evol 19: 1870-1880.
92. Hahn ME, Karchner SI, Franks DG, Merson RR (2004) Aryl hydrocarbon receptor polymorphisms and dioxin resistance in Atlantic killifish (*Fundulus heteroclitus*). Pharmacogenetics 14: 131-143.
93. Hahn ME, Karchner SI, Franks DG, Evans BR, Nacci D, et al. (2005) Mechanism of PCB- and Dioxin-Resistance in Fish in the Hudson River Estuary: Role of Receptor Polymorphisms.
94. Meyer JN, Di Giulio RT (2002) Heritable adaptation and fitness costs in killifish (*Fundulus heteroclitus*) inhabiting a polluted estuary. Ecological Applications 13: 490-503.

95. Nacci D, Champlin D, Jayaraman S (in press) Adaptation of the estuarine fish *Fundulus heteroclitus* to toxic pollutants. Estuaries and coasts.
96. Dawley RM (1992) Clonal hybrids of the common laboratory fish *Fundulus heteroclitus*. Proceedings of the National Academy of Sciences of the United States of America 89: 2485-2488.
97. Hardie DC, Hebert PDN (2004) Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences 61: 1636-1646.
98. Hinegardner R (1968) Evolution of cellular DNA content in teleost fishes. American Naturalist 102: 517-523.
99. Gregory TR (2001) Animal Genome Size Database - http://www.genomesize.com/fish.htm.
100. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222-234.
101. Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.
102. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064-U1063.
103. Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, et al. (2006) Research article - The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314: 941-952.
104. Warren WC, Hillier LW, Graves JAM, Birney E, Ponting CP, et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution. Nature 453: 175-U171.
105. Matson CW, Clark BW, Jenny MJ, Fleming CR, Hahn ME, et al. (2008) Development of the morpholino gene knockdown technique in *Fundulus heteroclitus*: A tool for studying molecular mechanisms in an established environmental model. Aquatic Toxicology 87: 289-295.
106. Nacci D, Coiro L, Wasserman D, Di Giulio R (2005) A non-destructive technique to measure Cytochrome P4501A enzyme activity in living embryos of the estuarine fish *Fundulus heteroclitus*. In: Ostrander GK, editor. Techniques in Aquatic Toxicology, Volume 2. Boca Raton, FL: Taylor and Francis. pp. 209-225.
107. Li R, Fan W, Tian G, Zhu H, He L, et al. (2009) The sequence and *de novo* assembly of the giant panda genome. Nature advance online publication.
108. Li R, Zhu H, Ruan J, Qian W, Fang X, et al. (2009) *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Research: -.
109. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, et al. (2009) ABySS: A parallel assembler for short read sequence data. Genome Research 19: 1117-1123.
110. Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Meth 6: S6-S12.
111. Choi JH, Kim S, Tang H, Andrews J, Gilbert DG, et al. (2008) A machine-learning approach to combined evidence validation of genome assemblies. Bioinformatics 24: 744-750.
112. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3: e3376.
113. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, et al. (2004) The Ensembl automatic gene annotation system. Genome Research 14: 942-950.
114. Salamov AA, Solovyev VV (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Research 10: 516-522.
115. Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. Genome Research 10: 547-548.
116. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: -.
117. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33: D501-D504.
118. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research 31: 5654-5666.
119. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32: D258-D261.
120. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5: -.

121. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Research 32: D277-D280.
122. Elsik CG, Worley KC, Zhang L, Milshina NV, Jiang H, et al. (2006) Community annotation: Procedures, protocols, and supporting tools. Genome Research 16: 1329-1333.
123. Stein L (2001) Genome annotation: From sequence to biology. Nature Reviews Genetics 2: 493-503.
124. Cameron RA, Samanta M, Yuan A, He D, Davidson E (2009) SpBase: the sea urchin genome database and web site. Nucl Acids Res 37: D750-754.
125. Stein LD, Mungall C, Shu SQ, Caudy M, Mangone M, et al. (2002) The Generic Genome Browser: A building block for a model organism system database. Genome Research 12: 1599-1610.
126. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, et al. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439-3440.
127. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.
128. Lewis S, Searle S, Harris N, Gibson M, Iyer V, et al. (2002) Apollo: a sequence annotation editor. Genome Biology 3: research0082.0081 - 0082.0014.
129. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research 17: 240-248.
130. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403-410.
131. Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: Genome-wide sampling of insect populations. Annual Review of Entomology 46: 441-469.
132. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: From genotyping to genome typing. Nature Reviews Genetics 4: 981-994.
133. Schlotterer C (2003) Hitchhiking mapping - functional genomics from the population genetics perspective. Trends in Genetics 19: 32-38.
134. Beaumont MA (2005) Adaptation and speciation: what can F-st tell us? Trends in Ecology & Evolution 20: 435-440.
135. Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). Molecular Biology and Evolution 23: 773-783.
136. Campbell D, Bernatchez L (2004) 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 21: 945-956.
137. Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular Ecology 14: 671-688.
138. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genetics 3: 901-915.
139. Mathew LK, Sengupta SS, LaDu J, Andreasen EA, Tanguay RL (2008) Crosstalk between AHR and Wnt signaling through R-Spondin1 impairs tissue regeneration in zebrafish. FASEB J 22: 3087-3096.
140. Karchner SI, Franks DG, Powell WH, Hahn ME (2002) Regulatory interactions among three members of the vertebrate aryl hydrocarbon receptor family: AHR repressor, AHR1, and AHR2. Journal of Biological Chemistry 277: 6949-6959.
141. Meyer JN, Wassenberg DM, Karchner SI, Hahn ME, DiGiulio RT (2003) 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 22: 2337-2343.
142. Powell WH, Bright R, Bello SM, Hahn ME (2000) Developmental and tissue-specific expression of AHR1, AHR2, and ARNT2 in dioxin-sensitive and -resistant populations of the marine fish, *Fundulus heteroclitus*. Toxicological Sciences 57: 229-239.
143. Blanchette M, Tompa M (2003) FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res 31: 3840-3842.

144. Ovcharenko I, Loots GG, Giardine BM, Hou M, Ma J, et al. (2005) Mulan: multiple-sequence local alignment and visualization for studying function and evolution. Genome Research 15: 184-194.
145. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, et al. (2005) Evolution and functional classification of vertebrate gene deserts. Genome Research 15: 137-145.
146. Fu Y, Weng Z (2005) Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences. Genome Inform 16: 68-72.
147. Fogel GB, Weekes DG, Varga G, Dow ER, Craven AM, et al. (2005) A statistical analysis of the TRANSFAC database. Biosystems 81: 137-154.
148. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378.
149. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, et al. (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29: 281-283.
150. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18: 1851-1858.
151. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6: 677-U676.
152. Nielsen R (2005) Molecular signatures of natural selection. Annual Review of Genetics 39: 197-218.
153. Tajima F (1989) Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123: 585-595.
154. Mcdonald JH, Kreitman M (1991) Adaptive protein evolution at the adh locus in *Drosophila*. Nature 351: 652-654.
155. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.
156. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27: 182-189.
157. Okou DT, Locke AE, Steinberg KM, Hagen K, Athri P, et al. (2009) Combining Microarray-based Genomic Selection (MGS) with the Illumina Genome Analyzer Platform to Sequence Diploid Target Regions. Annals of Human Genetics 73: 502-513.
158. Gross JB, Borowsky R, Tabin CJ (2009) A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. PLoS Genetics 5: -.
159. Crawford DL, Pierce VA, Segal JA (1999) Evolutionary physiology of closely related taxa: Analyses of enzyme expression. American Zoologist 39: 389-400.
160. Crawford DL, Segal JA, Barnett JL (1999) Evolutionary analysis of TATA-less proximal promoter function. Molecular Biology and Evolution 16: 194-207.
161. Evans BR, Karchner SI, Allan LL, Pollenz RS, Tanguay RL, et al. (2008) Repression of aryl hydrocarbon receptor (AHR) signaling by AHR repressor: role of DNA binding and competition for AHR nuclear translocator. Mol Pharmacol 73: 387-398.
162. Karchner SI, Jenny MJ, Tarrant AM, Evans BR, Kang HJ, et al. (2009) 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 29: 3465-3477.
163. Oleksiak MF, Wu S, Parker C, Karchner SI, Stegeman JJ, et al. (2000) Identification, functional characterization, and regulation of a new cytochrome P450 subfamily, the CYP2Ns. Journal of Biological Chemistry 275: 2312-2321.
164. Oleksiak MF, Wu S, Parker C, Qu W, Cox R, et al. (2003) 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 411: 223-234.
165. Pierce VA, Crawford DL (1994) Rapid enzyme assays investigating the variation in the glycolytic pathway in field-caught populations of *Fundulus heteroclitus*. Biochemical Genetics 32: 315-330.
166. Pierce VA, Crawford DL (1996) Variation in the glycolytic pathway: The role of evolutionary and physiological processes. Physiological Zoology 69: 489-508.
167. Pierce VA, Crawford DL (1997) Phylogenetic analysis of glycolytic enzyme expression. Science 276: 256-259.

168. Pierce VA, Crawford DL (1997) Phylogenetic analysis of thermal acclimation of the glycolytic enzymes in the genus *Fundulus*. Physiological Zoology 70: 597-609.
169. Powell WH, Morrison HG, Weil EJ, Karchner SI, Sogin ML, et al. (2004) Cloning and analysis of the CYP1A promoter from the Atlantic killifish (*Fundulus heteroclitus*). Marine Environmental Research 58: 119-124.
170. Segal JA, Barnett JL, Crawford DL (1999) Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. Journal of Molecular Evolution 49: 736-749.
171. Stegeman JJ, Woodin BR, Singh H, Oleksiak MF, Celander M (1997) 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 116: 61-75.
172. Yang X, Liu D, Murray TJ, Mitchell GC, Hesterman EV, et al. (2005) The aryl hydrocarbon receptor constitutively represses c-myc transcription in human mammary tumor cells. Oncogene 24: 7869-7881.