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

PI: Andrew Whitehead (Louisiana State University)

Collaborator: Joseph Shaw (Indiana University)

Collaborator: Mark Hahn (Woods Hole Oceanographic Institution)

Collaborator: Margie Oleksiak (University of Miami)

Collaborator: David Rand (Brown University)

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 an unusual and 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**. By doing this, we will address three specific aims. The first aim is to identify the number and physical location of loci associated with the evolved tolerance phenotype. The second aim is to test whether the same loci are implicated in each population where tolerance has convergently evolved. The third aim is to test for association between mapped loci and diverged patterns of gene expression.

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

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 distributed across diverse disciplines, the *F. heteroclitus* genome sequence will **enable and accelerate research in environmental genomics, ecological genomics, physiological genomics, comparative genomics, and evolutionary genomics, thereby transforming the research landscape in these rapidly emerging fields.** As part of this project we will host a genome annotation workshop and establish a collaboration wiki, which will be excellent training devices 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 serve to seed and foster cross-disciplinary collaborations since the diverse PIs that form the *Fundulus* Genomics Consortium will again be brought together in a mutually engaging and intense research setting.

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 research in environmental and evolutionary genomics [1]. A robust genome sequence is necessary to enable the next generation of cutting-edge research. The genome will be a significant resource for a diverse community of researchers, but here we propose to use the genome to identify the genomic basis of dramatic, rapid, convergent evolutionary adaptation to extreme environmental stress. The genome sequence will serve as a template to which various existing lines of experimental evidence will be anchored to identify the number and distribution of genomic regions, and identify candidate loci, associated with dramatic adaptive convergence.

A.2. Research Objectives and Specific Aims

Our overarching goal is to identify the genomic architecture and candidate genes underpinning adaptive convergent phenotypes. This will be accomplished by generating a draft genome sequence for *F. heteroclitus* which will serve as a template to which QTL mapping, population genome scan, and comparative transcriptome profiling data (generated from other completed or ongoing projects by the PIs and collaborators) will be anchored.

Aim 1: Identify the number and physical location of loci associated with the evolved tolerance phenotype. To achieve this, we will map existing QTL and population genome scan markers to the genome sequence.

Aim 2: Test whether the same loci are implicated in each population where tolerance has evolved independently. To achieve this, we will test whether the same genomic regions are implicated from QTL and population genome scan mapping of multiple tolerance-derived populations.

Aim 3: Test for association between mapped loci and divergent patterns of gene expression. Multiple transcriptomics datasets are completed or in progress that identify genes and pathways that are differentially regulated between sensitive and tolerant populations.

A.3. Expected Significance

With a genome sequence for *F. heteroclitus*, 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. Importantly, products from this project will serve as a critical resource for accelerating the next generation of research in this important model species, and will serve to transform the research landscape in the rapidly emerging fields of environmental genomics, ecological genomics, physiological genomics, comparative genomics, and evolutionary genomics.

B. PREVIOUS NSF SUPPORT

A recent (June 2007) award to AW (BES-0652006) is supporting comparative genomics studies in tolerant and sensitive populations of *F. heteroclitus*, including genotyping of QTL mapping families, population genomics genotyping, and comparative transcriptomics studies. Importantly, these data will be used to query the genome proposed here, and to address the specific aims proposed here. QTL mapping families have been generated for two of the tolerant populations (in collaboration with Diane Nacci, US EPA). Both have been phenotyped, and genotyping is in progress. Similarly, comparative transcriptomics studies have been initiated, where embryos from tolerant and sensitive populations (having been common-gardened in clean water for at least 2 generations) were challenged with a range of PCB doses. Experiments for one tolerant population are completed, and experiments for other tolerant populations are planned for spring 2009. For population genomics experiments tissue samples have been collected (in collaboration with David Rand, Brown University) from 60 individuals from each of three tolerant and eight sensitive reference populations, and DNA has been extracted and genotyped for a subset of individuals. In the one year since funding, this grant has already supported four publications [2,3,4,5].

JRS received NSF support from 9/2002 to 9/2008 as a co-PI on DEB-BE/GEN-EN (#0221837) titled, "*Development of Methods Linking Genomic and Ecological Responses in a Freshwater Sentinel Species*". This project coupled lab and field studies to explore the evolution of metal tolerance in the micro-crustacean sentinel, *Daphnia pulex*. It created genomic resources for *D. pulex*, including sequencing its transcriptome, constructing a database to disseminate the sequence information (wFleabase), and developing micorarrays for functional genomics investigations of this species complex. These tools, which are now community resources, were applied to lab cultures and natural populations of *Daphnia* to understanding the molecular basis of physiological acclimation and genetic adaptation to metal stress. Results offered important insights into the genetic and biochemical mechanisms that underpin evolved metal tolerance in laboratory-derived and field-derived populations. Analyses of polymorphisms in markers tightly linked to candidate metal responsive genes revealed combinations of genes and alleles that strongly correlated with the adaptive phenotype and identified susceptible populations. 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 (wFleabase; http://wfleabase.org/), produced a set of 70bp oligonucletides that represents over 10,000 genes, developed genome tiling path arrays that contain 4.2 million overlapping 70bp fragments representing the entire sequenced genome chemically synthesized on glass slides, and constructed a 12-plex transcriptional profiling microarray that simultaneously compares12 experiments on a single slide -each interrogating all genes represented by 135,000 probes, and contributed the formation of new model system (http://www.nih.gov/science/models/). This grant has resulted in over 15 publications of which at least six are of specific interest to this proposal [6,7,8,9,10,11]. In addition, it played a significant role in attracting a special issue of *Science* scheduled for release 2009 that will launch the *Daphnia* model system. This issue will feature four manuscripts, including one that describes the evolution of metal tolerance sponsored by this grant, and will be supported by the concordant publication of over 40 companion papers in BioMed Central journals.

C. PRELIMINARY DATA

Preliminary data that will assist with genome assembly and annotation, and that will be 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 [12] and annotation algorithms designed specifically for *F. heteroclitus* [13], are available to assist with genome assembly, gene finding, and annotation. Importantly, collaboration with the Center for Genomics and Bioinformatics at Indiana University exploits their extensive experience in developing the sophisticated infrastructure necessary to enable cutting edge genomics research in new model species (see section F.5).

New sequencing efforts by partners at the Washington University Genome Sequencing Center are serving as proof-of-principle for using all next-generation sequencing for *de novo* assembly of complex eukaryotic genomes. These efforts have sought to estimate the sequencing effort required for draft assembly since the shorter next-gen sequence reads are well-known to assemble less efficiently than traditional Sanger sequence reads given the same depth of coverage. Successful assembly of new chicken and *C. elegans* genomes indicate that ~17X coverage, using a mix of long 454 pyrosequencing reads and short Illumina reads combined with paired-end sequencing of large fragments, is an efficient strategy for *F. heteroclitus* given the estimated genome size. To our knowledge, this is the only group to have assembled a large complex eukaryote genome *de novo* using all 454 pyrosequencing reads, and is therefore uniquely qualified to estimate the sequencing effort necessary for successful assembly.

This proposal seeks support primarily for genome sequencing, assembly, and annotation. Importantly, the data necessary to query the genome to pursue our specific aims are already in hand, or are currently being produced through support from other sources by the PIs and collaborators (see attached letters of collaboration from Rand, Hahn, and Oleksiak). These data include QTL mapping data, population genome scan data, and comparative transcriptomics data, and are described in more detail in section F.6.

D. BACKGROUND and SIGNIFICANCE

Population responses following dramatic environmental changes could span the continuum from extirpation to evolutionary adaptation, and 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 capacity to tolerate the stress, and in yet other species the necessary adaptive genetic variation may quickly emerge or 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 [14,15,16,17,18]. 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 *Fundulus heteroclitus*. Eukaryote genome sequences are now feasible for PI-directed research projects given the longer reads and ultra-high throughput of the newest next-generation sequencing platforms. We will use this genome sequence to anchor transcriptomic and mapping data in order to identify the genomic architecture and candidate genes underlying the dramatic, rapid, and convergent evolution of extreme stress tolerance.

D.1. Fundulus as a model for evolutionary and ecological genomics

The power of comparative genomics, which can be used to identify the influence of positive and negative selection governing the evolution of genomic regions, is expanded significantly with the inclusion of more taxa. Fish are the most diverse group of vertebrates; they exploit and thrive in extraordinarily disparate habitats globally including deep ocean, hot springs, hyper and hypoosmotic habitats, hyper and hypothermal habitats, and display diverse habits associated with life history, mating system, and behavior. In order to understand the characters and mechanisms that underlie and drive organismal diversity, fish have been and will continue to be powerful comparative models [19,20]. By sequencing additional fish genomes, our investment in comparative genomics will be significantly leveraged and expanded. Other genome-enabled fish species represent wonderful biomedical models (zebrafish) or compelling models for studying genome evolution (pufferfish) or morphological evolution (sticklebacks, cichlids), but the promise of *Fundulus* is to stand as a powerful physiological model accessible to diverse research programs in environmental, ecological, and evolutionary science.

The uniqueness of *Fundulus* as model species for ecological and evolutionary genomics research derives from the combination of features that no other current genomics model shares. These features include evolved tolerance to pollution, extreme euryhalinity, eurythermia, and tolerance to hypoxia. They occupy diverse ecological niches across a very wide geographical distribution, along strong environmental clines and within human-impacted landscapes in North America and Europe. Their ecology and evolutionary history is well understood, as is their physiology, developmental biology, and molecular biology [1]. Populations are distributed along strong environmental clines in temperature, salinity, and pollution stress, thereby serving as models for studies in acclimation and adaptation [1,21,22,23,24,25,26]. Individuals are resilient to large variations in environmental conditions, thereby serving as excellent models for research on physiological plasticity [1]. Some populations hybridize with other species, thereby serving as models for studying speciation and the causes of clonal reproduction in vertebrates [27]. Populations live in highly human-altered environments, thereby serving as models for environmental biology and ecotoxicology. Some populations have been recently established in Europe [28], thereby serving as models for research in invasion biology. Large populations, ease of capture, moderate size, and ease of breeding and care enable both laboratory and field-based experimentation. No other teleost model exhibits such a wide range of beneficial characteristics.

The *Fundulus* research community has developed a broad suite of tools for genomics research including extensive expressed sequence tag (EST) libraries and cDNA and oligonucleotide microarrays supported by robust bioinformatics [12,29], a bacterial artificial chromosome (BAC) library (M. Gómez-Chiarri, and C. Amemiya, unpublished), population genetic markers including microsatellite [3,30,31], AFLP [32,33], and RAD markers (A. Whitehead, unpublished data), morpholino knockdown techniques

[34], transgenic lines [35], and a genetic map and complete molecular phylogeny of the genus (A. Whitehead, manuscript in preparation) are in progress. Accordingly, researchers in evolution, ecology, physiology, toxicology, and environmental biology, are poised to quickly exploit exciting new avenues of cutting-edge research enabled by a draft genome sequence (see attached letters of support).

In a recent review of genomic approaches for studying adaptive variation, Ellegren and Sheldon [15] present difficult choices that many evolutionary biologists may soon have to make in the post-genomics era where "the rich get richer". Either field biologists will have to switch focus to species which are genome-enabled, or laboratory-based biologists working with genome model species will have to seek ecological realism in order to enable fundamental insights into the evolutionary process. By sequencing the genome, the *Fundulus* genomics toolkit will be dramatically expanded, thereby accelerating research potential for this organism that bridges the gap between useful laboratory model and compelling ecological and evolutionary model.

D.2. The genetic basis of adaptive variation

Identifying the genetic 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 rate 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 [36] and [16]), it is clear that both types of variation can be important. For example, protein variation accounts for pigmentation variation in mice [37], birds [38], and lizards [39], cold tolerance in antarctic fishes [40], resistance to newt toxin in snakes [41], and hypoxia tolerance in mammals and birds [42,43]. In contrast, regulatory variation plays a role, for example, in eye degeneration in cavefish [44], developmental variation in beak morphology in finches [45,46], and pigmentation variation in fruit flies [47]. 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/or structural variation may underlie the environmental pollution tolerance phenotype 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, which exist from completed and ongoing projects, will be used to interrogate the genome sequence proposed here 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.

D.3. 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 alternatively may involve fundamentally different genes or biochemical pathways. Examples of both types of process (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 [48,49], loss of pelvic structures in multiple stickleback species and also in manatees [50,51], and adaptive coat color variation in beach mice [37] and other animals including lizards, birds, cats, and mammoth (see review in [14]). Intriguingly, in the case of beach mice, the *Mc1r* gene is implicated for light coat coloration in Gulf-coast (USA) derived populations, but not for Atlantic derived populations in which other genes must be responsible [37]. Indeed, there are several other cases in which adaptive convergence is underpinned by different genetic pathways. For example, pigment loss and eye loss have convergently evolved through different genetic mechanisms in Mexican cavefish populations [52], as has pigment gain in rock mouse populations [53]. Clearly, similar or different genetic mechanisms can underlie convergent phenotypes, whether or not convergent lineages are closely or distantly related [14].

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 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 [54]. This is the likely mechanism whereby convergent morphological types have evolved in stickleback fish [48]. 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 sensitive 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 anchoring 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.

D.4. Pollution tolerance in Fundulus

Evolved pollution tolerance in *Fundulus heteroclitus* is an unusual and compelling model system for evolutionary genomics for three primary reasons.

1. The adaptive phenotype is very dramatic (Fig.1); individuals from tolerant populations are more than three orders of magnitude more tolerant to environmental pollution stress (most notably to dioxin-like compounds) compared to individuals from nearby reference populations and other species, whereas *F. heteroclitus* in general is relatively sensitive to this kind of stress among fish species [55,56].
2. This adaptive phenotype has evolved over an extraordinarily short time period (Fig.2). The polluted sites where tolerant populations reside have been contaminated with tolerance-associated chemicals (primarily PCBs and PAHs) for only dozens of generations [57].
3. This extraordinary tolerance in *F. heteroclitus* has evolved independently at least four times [1,30,33].

This provides a very powerful comparative system for determining the genomic basis of rapid and dramatic phenotypic adaptation, and also for testing important questions related to the repeatability of evolutionary change [14] and the relative roles of selection on standing genetic variation versus selection on *de novo* mutation underlying evolutionary innovation [58].

F. heteroclitus has long been known to be highly adaptable to changes in environmental conditions [59,60,61]. Evolved tolerance or resistance to pollutants was demonstrated first for methylmercury [62,63] and later for several structurally related aromatic hydrocarbons, including chlorinated dioxins (Newark, NJ [64,65]), PCBs (New Bedford Harbor, MA [66,67]), and PAHs (from creosote; Elizabeth River, VA [56,68,69]).

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 [56,57]). 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 [56]. (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 [70]. PCB-resistant New Bedford fish are also resistant to PAHs [66,67]. PAH-resistant Elizabeth River fish are also resistant to PCBs [71]. This cross-tolerance suggests that the mechanisms of resistance may converge on a common biochemical pathway. (iii) The tolerance appears to involve both heritable and non-heritable mechanisms [56,66,69,71,72]. (iv) The tolerant populations do not show an overall loss of genetic diversity, as assessed by allozyme and molecular methods [32,73,74,75]. However, the tolerant populations are genetically distinct from nearby sensitive populations and do show evidence for selection at specific loci [74,76,77,78]. (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 [55]. 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 the pollutant tolerant phenotype [56,79]. 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.

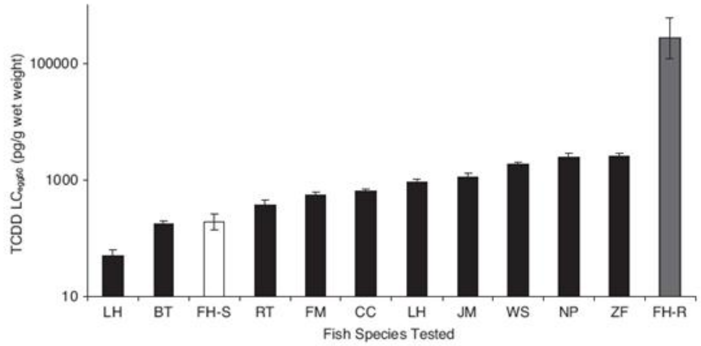


Figure 1. Variation in early life stage tolerance to TCDD of various fish species. *F. heteroclitus* in general are a relatively sensitive species (FH-S), whereas tolerant populations of *F. heteroclitus* (FH-R) are orders of magnitude more tolerant than all species tested [lake trout (LT), brook trout (BT), rainbow trout (RT), fathead minnow (FM), channel catfish (CC), lake herring (LH), Japanese medaka (JM), white sucker (WS), northern pike (NP), and zebrafish (ZF)]. Adapted from Van Veld and Nacci (2008).

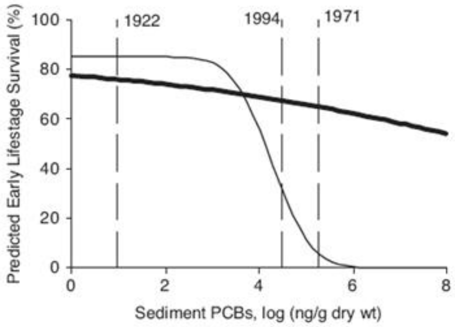


Figure 2. Survival models for sensitive (light line) and tolerant (dark line; New Bedford Harbor, MA) populations of *F. heteroclitus*. Estimated sediment concentrations based on sediment cores are shown for 1922, 1971 when PCBs were banned, and in recent times (adapted from Van Veld and Nacci (2008).

D.5. State of readiness for genome sequencing

Based on flow cytometry and bulk fluorometric analysis, the *Fundulus heteroclitus* haploid genome size is between 1.29 and 1.5 billion bp [27,80,81] which is intermediate in size between Japanese medaka and zebrafish [82]. We will sequence one individual from a stock of fish that have been inbred for ten generations to minimize heterozygosity. Individuals from the most recent generation have allelic diversity (based on 15 microsatellites) equivalent to full sibs. Other sequencing projects have successfully produced high-quality assemblies from fully outbred sometimes highly heterozygous individuals [83,84,85,86,87] from sequencing efforts equivalent to what we propose here.

Extensive EST sequencing with robust annotations exists for *F. heteroclitus* [12]. Our 74,000+ EST sequences will aid with assembly and gene-finding, and for adding UTR information to genes identified through protein homology (see sections F.3, F.5). Extensive additional EST sequencing representing 72 different conditions (life-stages, sexes, tissues, salinities, metal-treatments) has recently been completed by the Shaw lab, in collaboration with Bruce Stanton (Dartmouth Medical School). This project utilized newly developed cDNA normalization procedures for 454 sequencing, which produced 1.5 million more EST sequences to be used for assembly, assembly validation, and gene finding. A novel and highly robust gene annotation pipeline has been developed for *F. heteroclitus* EST collections [13], and will serve to correctly annotate genes identified by protein and EST homology (see section F.5).

The bioinformatics infrastructure that will be used for assembly validation and community-based manual curation of annotations has already been successfully implemented for water flea and aphid genome projects (see https://dgc.cgb.indiana.edu). Adaptation of this resource for application in this *F. heteroclitus* project is highly feasible (see attached letter of support from John Colbourne).

Ongoing projects in the labs of A. Whitehead, P. Schulte (University of British Columbia; see letter of support), and M. Bagley (US EPA, Cincinatti) are generating genetic maps from *F. heteroclitus* QTL mapping experiments. The Whitehead project is using Illumina-generated 36-base restriction site associated sequence tags as polymorphic markers [88] where thousands of markers for dozens of individuals are scored in a single Illumina run [89]. The Schulte and Bagley projects are using hundreds of microsatellite markers. Collectively, these markers and maps will be used for ultracontig assembly (see section F.3). Finally, a complete molecular phylogeny has been generated from multiple populations from all extant members of the *Fundulus* genus (A. Whitehead, manuscript in preparation), and this phylogeny will serve as a template for future comparative experiments.

E. BROADER IMPACTS

A high quality genome sequence for *Fundulus heteroclitus* would **transform the research landscape for environmental, ecological, and evolutionary genomics**. 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 *Fundulus* Genomics Consortium (FGC) (see attached letters of support). For example, the genome sequence will be of immediate and future use in the following domains:

1. Physiological genomics: *Fundulus* research has contributed to understanding fundamental mechanisms associated with physiological resilience, especially to key environmental variables such as salinity, hypoxia, and temperature (see review in [1]). Members of the FGC and others would use the genome sequence to derive lists of genes and associated promoters to test, for example, the roles of gene family expansion and novel regulatory and signaling architectures that may enable physiological resilience, and to design gene knockdown morpholinos for studies dissecting biochemical and physiological mechanisms (D. Crawford, University of Miami; D. Evans, University of Florida; B. Fuller, University of Illinois; F. Galvez, Louisiana State University; W. Marshall, St. Francis Xavier University; J. Podrabsky, Portland State University; R. Preston, Illinois State University; B. Rees, University of New Orleans; P. Schulte, University of British Columbia; J. Shaw, Indiana University; T. Singer, University of Waterloo; D. Towle, Mount Desert Island Biological Laboratory; A. Whitehead, Louisiana State University).

2. Environmental genomics: *Fundulus heteroclitus* populations are resident in many human altered waterscapes, and are important models in environmental stress biology. Numerous groups would use the genome sequence, for example, to anchor environmental stress tolerance associated QTL and genome scan markers, to study the promoters of stress-induced genes and the associated regulatory architecture of diverse environmental stress responses, and to design knockdown morpholinos to study the effects of stressors on development, toxicity, cancer, and physiological function (L. Bain, Clemson University; B. Baldwin, Clemson University; S. Bard, Dalhousie University; G. Callard, Boston University; S. Cohen, San Francisco State University; R. Di Giulio, Duke University; M. Gomez-Chiarri, University of Rhode Island; M. Hahn, Woods Hole Oceanographic Institution; J. Hamilton, Marine Biological Laboratory; D. Nacci, EPA; J. Shaw, Indiana University; J. Stegeman, Woods Hole Oceanographic Institution; D. MacLatchey, Wilfred Laurier University; W. Marshall, St. Francis Xavier University; J. Shaw, Indiana University; P. Van Veld, Virginia Institute of Marine Science; W. Vogelbein, Virginia Institute of Marine Science; K. Willett, University of Mississippi; R. Winn, University of Georgia).

3. Evolutionary genomics: *Fundulus heteroclitus* occupies habitats distributed along strong environmental gradients, and has been an important model for the study of adaptive evolution in promoter regions and protein coding regions, glycolytic enzyme evolution, evolution of pollution tolerance, and new studies in osmotic adaptation (see review in [1]). Members of the FGC and others would use the genome sequence, for example, to study the evolution of protein families (ie: nuclear receptor families, claudins and other transport and signaling proteins, and cytochrome P450 gene families), the regulatory architecture of genes with similar patterns of expression following various environmental challenges, to design gene knockdown morpholinos for studies verifying adaptive relevance of gene variants, and to anchor QTL and genome scan data for identifying genes underlying adaptive variation (D. Crawford, University of Miami; D. Duvernell, Southern Illinois University Edwardsville; F. Galvez, Louisiana State University; M. Gomez-Chiarri, University of Rhode Island; B. Fuller, University of Illinois; M. Hahn, Woods Hole Oceanographic Institution; B. Kreiser, University of Southern Mississippi; D. Nacci, EPA; M. Oleksiak, University of Miami; J. Postlethwaite, University of Oregon; D. Rand, Brown University; J. Schaefer, University of Southern Mississippi; J. Stegeman, Woods Hole Oceanographic Institution; A. Whitehead, Louisiana State University).

At the final stage of genome annotation, the Indiana University Center for Genomics and Bioinformatics will host a collaboration wiki, web-based conferences, and an annotation training workshop for members of the 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). This workshop will serve important additional functions. First, it will serve as an excellent training device for students and post-docs (from FGC member labs) in cutting-edge techniques at the forefront of bioinformatics and comparative genomics (we will provide travel awards for students and post-docs to maximize participation and training potential). Second, it will bring together PIs from the diverse *Fundulus* research community, and foster the kinds of cross-disciplinary collaborations that are most likely to emerge from communication-intensive, information-rich, community-based interactions.

In addition to producing an important resource for the genomics community and as a training device for students and post-docs, this project will contribute to our understanding of the genetic 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 [14,15,16,17,18]. 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.

F. METHODS and MATERIALS

F.1. DNA sequencing

Next-generation massively parallel sequencing technologies provide a dramatic cost savings per megabase compared to traditional Sanger sequencing. The disadvantage has been that sequence reads are typically shorter, so greater depth of sequencing coverage is necessary to generate equivalent quality assemblies. Fortunately, next-generation sequencing reads have been getting longer. Currently, the throughput advantage derived from massively parallel sequencing far exceeds disadvantages posed by shorter reads, such that high quality genome assemblies are feasible at a fraction of the cost compared to technologies available only one or two years ago. As a proof-of-principle, the genome sequencing group at the Washington University Genome Sequencing Center (WUGSC; under the direction of Dr. Wes Warren, Assistant Director, personal communication) re-sequenced the chicken and *C. elegans* genomes (originally sequenced using Sanger technology) using all next-generation technologies in order to estimate the depth of sequence coverage necessary for high-quality *de novo* assemblies of complex eukaryotic genomes. The WUGSC group is finding that the depth of 454 sequencing coverage needs to be at least double that of Sanger sequencing to achieve similar levels of assembly contiguity (Wes Warren, personal communication).

Based on the chicken re-sequencing effort, experts at WUGSC estimate that ~17X coverage, using a mix of reads from the 454 Life Sciences Titanium pyrosequencer and the Illumina Genome Analyzer, will provide sufficient sequence for a draft assembly of the *F. heteroclitus* genome. For example, 12x coverage of various 454 read types produced a *de novo* assembly with N50 scaffolds (supercontigs) of 214kb. Coverage and contiguity assessments of this chicken draft assembly suggest sufficient quality for downstream gene annotation (Warren et al., manuscript in prep.). The inclusion of Illumina coverage in the sequencing plan for the *F. heteroclitus* genome is expected to improve assembly contiguity and to correct insertion and deletion errors observed in pure 454 based assemblies. The proposed Illumina coverage is a moderate cost for the expected gain in assembly quality.

Our strategy, with the chicken genome serving as proof-of-principle, is to use the relatively long reads (350-400 bases) from the 454 titanium technology to provide adequate genome coverage (10x) and paired end read types to address contiguity (7x). The paired end reads will include a mix of long 454 reads from long 20kb insert libraries (1X) and 75-base Illumina reads from 3 kb insert libraries (6X). Both of these paired-end read types will serve to improve assembly contiguity. Indeed, the 20kb inserts used for paired-end sequencing will cover the genome at least 33X. This total depth of sequencing coverage (17X) is predicted to yield N50 contig sizes of at least 10 kb (N50 is a metric for judging assembly contiguity, and refers to the minimum contig size in which half of all bases sequenced reside). As a general guide, sequencing efforts are scaled such that N50 values match or exceed typical gene length. From the Japanese medaka genome sequencing project, contig N50 values of 9.8 kb were generated from 10.6-fold sequencing coverage using Sanger sequencing [90]. Since *Fundulus heteroclitus* is closely related to medaka we expect gene lengths between these species to be similar. The mean length of unspliced genes downloaded from the medaka database at Ensembl is ~10 kb.

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.

F.2. Quality control

Although error rates for the pyrosequencing technology were relatively high when first reported [91], successive generations of the technology have increased accuracy rates to 99.5% [92]. Additional simple quality control filters, such as eliminating a small proportion of reads with ambiguous bases, can help account for errors associated with multi-templated beads and homopolymer extension, and increase the accuracy rate to 99.75% [92]. Prior to assembly the PCAP software has a series of base trimming scripts which will decrease the contribution of low quality bases to the consensus sequence derived from all reads [93].

F.3. Genome assembly

Genome assembly will involve four principal steps that progress from forming contigs from raw sequence reads, to connecting contigs into supercontigs using paired-end sequence of large fragments, to joining supercontigs into ultracontigs using mapping data and ultimately multi-species alignments. In more detail, the four steps are as follows:

1. The PCAP algorithm will be used for initial assembly of all individual sequence reads into contigs by an overlap detection process.
2. Paired-end sequence reads (454 sequence from 20kb inserts, and Illumina sequence from 3kb inserts) will enable supercontig assembly.
3. Markers currently being used for the generation of a *F. heteroclitus* genetic map include ~36-base restriction site associated DNA (RAD) sequences generated by solexa sequencing [89] in an ongoing project in the Whitehead laboratory, and collaborators have also generated ~300 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 ultracontig assembly.
4. For ultracontigs and supercontigs that cannot be joined by any of the above procedures, we will screen multi-species genome alignments (including puffer fish, zebrafish, stickleback, and Japanese medaka) for conserved gene orders and orientations to enable final genome assembly. At this time of writing (December 2008) the cichlid genome is in assembly, so will also be available to our project for multi-species alignments. This alignment strategy was useful for assembling the platypus genome where other distantly-related mammal genomes were aligned to identify conserved order and orientation [87]. We expect this strategy to work nicely since *F. heteroclitus* is relatively closely related to other fish species. For example, percent sequence divergence between *F. heteroclitus* and Japanese medaka is similar to that between human and rat; sequence divergence for genes ENC1, RAG1, and Ldh-b are 13.7%, 13.3%, and 15.5%, respectively, between human and rat, whereas the same genes are 14.8%, 15.9%, and 11.4% divergent between Japanese medaka and *F. heteroclitus*.

F.4. Assessing structural accuracy of assembly

Despite improvements in assembly algorithms, assembling genomes from millions of small partially overlapping sequence reads in automatic fashion is susceptible to producing errors. We will assess the accuracy of the assembled *F. heteroclitus* genome using the methods developed by the Indiana University Center for Genomics and Bioinformatics (CGB) that have been successfully applied to evaluate a *Daphnia pulex* genome [94]. For this analysis, the assembly will be assessed using a machine learning approach that compares several *in silico* measures including: read coverage, clone coverage, compression and extension statistics for unsatisfied mate-pairs, the number of good clones minus the number of bad clones, the ratio of good and bad clones, the average of Z-scores, and the maximum of absolute values of average positive Z-scores and average negative Z-scores. For the *D. pulex* genome project, we have used this approach to successfully verify assembly by building a training model from EST sequences and by comparing assemblies made by different assemblers (Colbourne et al., in prep.).

F.5. Gene finding, annotation & bioinformatics

First-pass gene prediction will use Eannot, which is a modified Ensembl pipeline [95], for evidence-supported gene model building and model merging. Uniprot protein sequences from *Fundulus 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 support 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 [96], Genewise family models [97] and SNAP [98]. Colleagues at the NCBI RefSeq Project Group will provide RefSeq transcript alignments [99] 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 [100] 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 [101], eukaryotic orthologous groups [102], and KEGG metabolic pathways [103].

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 [104,105]. To aid in this task, project partners at 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*), jewel wasp and pea aphid 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 three-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 [104,105]. The workshop will be held during the first summer 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.

The *Fundulus* genome annotation workshop will follow the design of those for the water flea (http://conferences.cgb.indiana.edu/daphnia2007/index.html) that was also hosted at Indiana University and the aphid (https://dgc.cgb.indiana.edu/display/aphid/Workshop+I). As was the case for these meetings, ours will be combined with a FGC meeting to ensure broad participation from the community. The meeting will include presentations from community members that will address research applications of the genome sequence and those by invited participants from other genome projects who will highlight the annotation process and speak from experience on ways to improve our annotation efforts. Participants will also work hands-on with annotation training modules. 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.

This dispersed annotation approach will initially focus on genes of interests 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 labs 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 [104,106].

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; [107]) components and open source software shared with other genome databases. GMOD leverages expertise from existing genome projects with the goal to fully develop and extend a genome database tool set to the level of quality needed to create and maintain new genome databases. Use of common components ensures rapid construction and interoperability. The genome access tools of GMOD (GBrowse [107], BioMart [108] and BLAST [109]) will be available for searching the *F. heteroclitus* genome. The GMOD Chado relational database schema (www.gmod.org/chado/) will be used for managing genome information. 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. The *Fundulus* database will be operable on several Unix computers, including Apple

Macintosh OSX and Intel Linux, and is portable enough to run on laptop computers. 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 (www.ensembl.org) and the University of California Santa Cruz (http://genome.ucsc.edu/) 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 [110]. The Apollo interface is similar to GBrowse in that users navigate by coordinates, gene name, or BLAST search to view an expandable number of tracks, including predicted gene sets, homologs, ESTs, marker genes and microsattelite locations. It differs from the genome browser in that it offers users full editorial capabilities of the gene models. With Apollo, we will probe, create, delete, split, merge, classify, and comment on annotations, according to the community adapted guidelines. 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 [13]. 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 [13].

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. TeraGrid provides collaborative, cost-effective scientific computing infrastructure is particularly suitable for genome assembly, annotation, gene finding and phylogenetic analyses. 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, microsatellite, and EST locations.

F.6. Testing Hypotheses:

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 [15], some maintain that regulatory elements are most likely to underpin new phenotypes [36], whereas others contend that protein variation accounts for much of adaptive variation [16]. Since there is no *a priori* reason to exclude the influence of either structural or regulatory variation in extremely stress tolerant phenotypes in *Fundulus heteroclitus*, we have surveyed both types of variation in comparative experiments. We will interrogate the genome sequence with genes and markers primarily derived from three sources in order to address our specific aims. These sources include:

1. Markers from QTL mapping families currently created for two different tolerant populations.
2. Markers from population genome scan studies including multiple tolerant and sensitive populations.
3. Genes from comparative transcriptome profiling that are associated with differences in stress response between tolerant and sensitive populations.

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 the genome in hand, physical associations can be made between markers and between markers and genes with interesting patterns of gene expression, facilitating closing of this gap.

F.6.1. QTL mapping data

In a recent review, Stinchcombe and Hoekstra [18] 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 [33] and additional data are currently being generated through support from other sources, and we do not seek support for such data collection through this proposal. In collaboration with Dr. Diane Nacci (US EPA Atlantic Ecology Division; see letter of support) we have generated F2 QTL mapping families from full-sib 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 is ongoing. In addition, we have QTL mapping families derived from another tolerant population which have been phenotyped, and are awaiting genotyping.

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 36-base Illumina-generated sequence tags that are associated with restriction sites [88,89]. A polymorphism between two individuals is scored as either presence/absence of one of thousands of sequence tags present in sequencing pools of different individuals (indicating a restriction site polymorphism) or through detection of SNPs in a specific sequence tag that is present in both individuals. 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 [88], thereby validating data from laborious AFLP and microsatellite mapping [48]. Importantly, these markers were sensitive enough to identify additional loci associated with morphology that were missed by AFLP and microsatellite mapping [88]. Over 13,000 SNPs were identified in the stickleback studies [89], 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). This approach will enable characterizing the number of unique locations in the genome identified by QTL markers, testing whether similar genomic regions are implicated in independently derived tolerant populations, testing for association between QTL loci and genes implicated from comparative transcriptomics studies, and will 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 [111] 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 CGB or Ensembl genome browsers to visualize the physical locations of markers relative to each other, relative to markers from population genome scans (section F.6.2), relative to genes with divergent expression patterns (section F.6.3), and relative to candidate genes.

F.6.2. Population mapping data

Our group will use two “population genomics” data sets to test for outlier loci between tolerant and sensitive populations implicating loci under selection, and these markers represent unique sequences that can be anchored to unique locations in the genome. For identifying the genomic basis of adaptive traits, population genomic data are naturally complementary to QTL mapping [15,18]. 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 [112,113,114]. If genome scan markers coincide with QTL-associated markers, this provides compelling evidence that genes within these regions are significantly involved in adaptive evolution.

Markers from two population data sets (an AFLP marker dataset and a RAD marker dataset) will be anchored to the genome. In a data set of ~300 AFLP markers from collaborator Margie Oleksiak (University of Miami), 24 outlier loci were detected in genome scans of three tolerant populations and six flanking sensitive reference populations [33]. Most outliers were specific to different tolerant populations, but four were shared among at least two of the tolerant populations. These data indicate that either different genomic regions are under selection in different tolerant populations, or that AFLP markers were sorting differently in separate ancestral populations but are physically liked to the same genomic region under selection in separate populations. The strongest way to test this is to anchor relevant AFLP markers to the genome sequence and identify marker associations to each other and to QTL loci and candidate genes.

A second population genomics study is using the same RAD markers as those used for QTL analysis. These experiments are being conducted in collaboration between labs of David Rand (Brown University) and PI Whitehead (see attached letter of collaboration). We have collected DNA samples from 60 field-caught individuals from each of three tolerant populations (two of which are subjects for QTL studies) and eight sensitive reference populations. Each of 60 individuals from each population are being genotyped at thousands of Illumina-sequenced RAD tag loci. Since QTL analyses are limited to a relatively small number of meioses from F2 full-sib 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 4 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 among-taxon divergence) between adaptively diverged populations are identified [33,112,113,115,116,117,118]. 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 [119], 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 anchored 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.

F.6.3. Transcriptome profiling data

Completed and ongoing projects in the laboratories of PI Whitehead, and collaborators Mark Hahn (Woods Hole Oceanographic Institution) and Margie Oleksiak (University of Miami), and other FGC members including Baldwin, Bain, and Di Giulio, are generating functional genomic datasets in studies of tolerant and sensitive populations. The common purpose of these studies is to identify regulatory patterns that indicate genes and biochemical pathways functionally involved in producing the derived tolerant phenotype.

The Whitehead lab, in collaboration with 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. RNA has been extracted from embryos, and hybridization to 7,500 gene oligonucleotide microarrays has started. These comparative data will be available in spring 2009. Similar additional experiments are planned for spring 2009 which include other tolerant populations and paired sensitive reference populations. The goals of these studies are to identify what tolerant individuals are doing differently, at the regulatory and biochemical pathway levels, to enable their extremely tolerant physiology, and to test whether convergent tolerant populations have converged on similar regulatory responses.

The collaborating lab of Margie Oleksiak (University of Miami; see attached letter of collaboration) 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 2009). This experiment was designed to differentiate both evolved and physiologically induced changes in gene expression in tolerant versus sensitive populations. The Oleksiak group has also completed a dose-response experiment in collaboration with Rich DiGiulio (Duke University) to explore synergistic effects of representative PAH-type CYP1A inducers and inhibitors on morphology, physiology and gene expression in developing fish embryos from both sensitive and tolerant populations. RNAs from individual, dosed embryos at four specific stages have been hybridized to the 7,000 gene cDNA arrays and data currently are being analyzed. A dose-response experiment in collaboration with Mark Hahn is ongoing and soon will be completed (below).

The collaborating lab of Mark Hahn (Woods Hole Oceanographic Institution; see attached letter of collaboration) 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) [67,77,120,121,122]. In an ongoing collaboration with Oleksiak (University of Miami), Hahn's group has exposed F1 embryos from the two sites to PCBs 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. The embryo RNA samples are currently (January 2009) being analyzed using a 7,000 gene cDNA array to identify gene-, site- and stage-specific responses to PCB exposure. 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, by examining the sensitivity of NBH and SC embryos to environmental hypoxia. Embryos from the two sites are being raised under conditions in which the extent and duration of hypoxia are varied. The ability of the embryos to mount an adaptive transcriptional response to hypoxia will be measured (Summer 2009) using the 7,000 gene cDNA array, 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).

F.7. Potential pitfalls and alternate strategies

Two of the main concerns with new genome projects are with having enough sequence information for *de novo* assembly, and enough cDNA sequence to support gene finding and for adding UTR information. One might initially consider the 17X coverage proposed here to be excessive. However, greater depth of coverage is necessary for shorter next-gen reads compared to Sanger reads for equivalent assembly contiguity, and our partners at WUGSC have done the proof-of principle experiments (with chicken) to indicate that 17X coverage including paired end reads is an efficient strategy for *F. heteroclitus*. However, given these concerns, we will add an assembly checkpoint after 12X coverage is achieved, and retain the remaining sequencing capability to be used in one of three discretionary ways, depending on circumstances. If assembly quality is insufficient, given low contig and supercontig N50 values, then these funds would be directed toward additional shotgun and paired-end sequencing. If assembly is good but difficulties are encountered in gene finding, then these funds would be directed toward additional EST sequencing. In addition, it is important to note that the cost of generating large-scale sequence data is continually falling, and by the earliest date that this project would start assembly (Fall/Winter 2009) the cost for producing 17X genome coverage will likely have decreased, thereby enabling more sequencing if necessary while remaining within the budgetary scope of this proposal.

If the genome does not require additional sequence for assembly or gene finding, then this extra sequencing capacity will be used to generate a second genome sequence of one individual from a highly tolerant population. This will be done at low coverage using short sequence reads from one run of the Illumina sequencer which will generate ~10Gb of paired-end sequence (~6X genome coverage from ~75bp reads). This is feasible because the high quality primary genome sequence will enable assembly of a second low-coverage genome that is good enough to allow comparative screening of proteins and promoters for signatures of selection. This is strategically sound because if the primary genome does not require 17X coverage, then the second low-coverage genome will serve as an additional powerful tool for our analyses, accelerating progress toward the research goals proposed. Even if the sequences for the primary genome do not ultimately assemble into the expected number of chromosomes (haploid chromosome number = 24), the ultracontigs produced will be more than sufficient to address the research questions posed here and for most of the research questions being pursued by the broader FGC community.

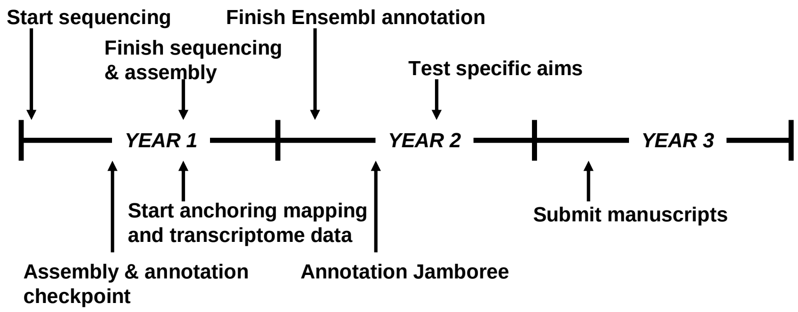
Beyond sequencing efforts, assembly strategies vary. Fortunately, our Washington University partners have produced a high quality *de novo* assembly for the chicken genome using all next-generation sequence using the PCAP algorithm. Other assemblers, such as Arachne [123] and Newbler (454 Life Sciences), may have difficulties handling this volume and type of sequence data, though these programs are likely to soon be updated and become useful for our purposes. One particular advantage for fish genomes, and especially for *F. heteroclitus*, is that several other genomes are available for alignment and joining ultracontigs. Japanese medaka and stickleback are particularly closely related to *F. heteroclitus* (similar to the genetic distance between human and rat) and are expected to significantly assist with long-range assembly. Indeed, the platypus genome assembly benefited from this approach by aligning with human, dog, opossum, and chicken: relatively distantly related species [87]. In addition to using multi-species alignments, we will also use the algorithms designed by the Genomics and

Bioinformatics group at the University of Indiana for assembly validation. These tools were successfully used to validate other genome assemblies including water flea and aphid (https://dgc.cgb.indiana.edu).

Though sequencing and assembly are quickly becoming routine, gene finding and functional annotation still represent challenges. We have chosen to apply *in silico* gene prediction models that incorporate evidence-based biological data (e.g., ESTs, microarrays) and implement an open annotation strategy that emphasizes genes of interest to the *F. heteroclitus* community. Given the considerable protein information available for teleost fishes and the extensive EST sequencing resources in hand for *F. heteroclitus*, we do not anticipate significant difficulties here. Indeed, comparable EST sequencing efforts were sufficient to discover many novel genes in the water flea genome (John Colbourne, personal communication). However, if using Uniprot sequences and EST sequences for seeding gene finding leads to discovery of too few genes, then we will include more transcript sequencing from strategically selected tissue and life stage libraries and use those sequences as seeds to find new genes. We may focus on biasing sequencing toward the 5’ end of transcripts by using the Illumina platform to sequence cDNAs derived from 5’ end-capped mRNAs [124], similar to the strategy used for the Japanese medaka genome [90]. One run of the Illumina machine would generate 10Gb of additional cDNA data, and this would be accomplished by using the discretionary sequencing capacity introduced above.

Accurate and informative gene annotation – that is, deciding what to name each gene in a way that best reflects gene function and in a way that allows informative comparisons across taxa – is an ongoing research endeavor in bioinformatics and genome science. Following the model of prior genome projects it is our intention to refine gene functional annotations after first rounds of analyses by using a phylogenetic perspective (i.e., PhIGs; [125]) and incorporating additional functional data as they become available (i.e., from tiling arrays, proteomics). The CGB is committed to regularly update frozen gene models for a period of 5 additional years, based on such evidence, and to provide a clear versioning of gene IDs for consistency in the growing literature referencing this data.

G. TIMELINE


Start sequencing
Finish Ensembl annotation
Finish sequencing & assembly
Test specific aims
YEAR 1
YEAR 2
YEAR 3
Start anchoring mapping and transcriptome data
Submit manuscripts
Assembly & annotation checkpoint
Annotation Jamboree

H. MANAGEMENT PLAN

Wes Warren (Washington University Genome Sequencing Center sub-contract PI) will oversee genome sequencing and assembly. Collaborator J. Shaw will oversee gene finding, annotation, and bioinformatics. AW and JS will organize the annotation workshop. Collaborators David Rand (Brown University), Mark Hahn (Woods Hole Oceanographic Institute), and Margie Oleksiak (University of Miami) will contribute data, and with PI Whitehead, will anchor their respective mapping and transcriptome data sets to the genome and test hypotheses (see attached letters of collaboration from Rand, Hahn, and Oleksiak). AW will coordinate the merging of data sets and writing of papers among collaborators.

I. 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 (2008) Simultaneous Extraction of High-Quality RNA and DNA from Small Tissue Samples. J Hered: esn083.

5. Whitehead A (in press) Comparative mitochondrial genomics within and among species of killifish. BMC Evolutionary Biology.

6. 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: -.

7. Colbourne JK, Singan VR, Gilbert DG (2005) wFleaBase: The *Daphnia* genome database. Bmc Bioinformatics 6: -.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. Young S (2008) Improved EST Annotation Using Multiple External Databases. Raleigh: North Carolina State University.

14. 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.

15. Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. Nature 452: 169-175.

16. Hoekstra HE, Coyne JA (2007) The locus of evolution: Evo devo and the genetics of adaptation. Evolution 61: 995-1016.

17. Hoffmann AA, Willi Y (2008) Detecting genetic responses to environmental change. Nat Rev Genet 9: 421-432.

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

19. Cossins AR, Crawford DL (2005) Opinion - Fish as models for environmental genomics. Nature Reviews Genetics 6: 324-333.

20. Streelman JT, Peichel CL, Parichy DM (2007) Developmental genetics of adaptation in fishes: The case of novelty. Annual Review of Ecology Evolution and Systematics 38: 655-681.

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

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

23. Pierce VA, Crawford DL (1997) Phylogenetic analysis of glycolytic enzyme expression. Science 276: 256-259.

24. Powers DA, Smith M, Gonzalez-Villasenor I, DiMichelle L, Crawford DL, et al. (1993) A multidisciplinary approach to the selectionist/neutralist controversy using the model teleost, *Fundulus heteroclitus*. In: Futuyma D, Antonovics J, editors. Oxford Surveys in Evolutionary Biology. New York, NY: Oxford University Press. pp. 43-108.

25. 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.

26. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. Molecular Ecology 15: 1197-1211.

27. 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.

28. Gisbert E, Lopez MA (2007) First record of a population of the exotic mummichog *Fundulus heteroclitus* (L., 1766) in the Mediterranean Sea basin (Ebro River delta). Journal of Fish Biology 71: 1220-1224.

29. 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.

30. 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.

31. Adams SM, Oleksiak MF, Duvernell DD (2005) Microsatellite primers for the Atlantic coastal killifish, *Fundulus heteroclitus*, with applicability to related *Fundulus* species. Molecular Ecology Notes 5: 275-277.

32. 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.

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

34. 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.

35. Winn RN, Norris MB, Brayer KJ, Torres C, Muller SL (2000) Detection of mutations in transgenic fish carrying a bacteriophage lambda cll transgene target. Proceedings of the National Academy of Sciences of the United States of America 97: 12655-12660.

36. 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.

37. 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.

38. 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.

39. Rosenblum EB, Hoekstra HE, Nachman MW (2004) Adaptive reptile color variation and the evolution of the Mc1r gene. Evolution 58: 1794-1808.

40. 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.

41. 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.

42. 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.

43. 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.

44. Yamamoto Y, Stock DW, Jeffery WR (2004) Hedgehog signalling controls eye degeneration in blind cavefish. Nature 431: 844-847.

45. 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.

46. 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.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. Hoekstra HE, Nachman MW (2003) Different genes underlie adaptive melanism in different populations of rock pocket mice. Molecular Ecology 12: 1185-1194.

54. Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. Trends in Ecology & Evolution 23: 38-44.

55. 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.

56. 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.

57. 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.

58. Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. Nature Reviews Genetics 8: 675-688.

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

60. 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.

61. Schulte PM (2001) Environmental adaptations as windows on molecular evolution. Comp Biochem Physiol B Biochem Mol Biol 128: 597-611.

62. 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.

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

64. 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.

65. 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.

66. 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.

67. 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.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. 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.

78. 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.

79. Meyer JN, Di Giulio RT (2002) Heritable adaptation and fitness costs in killifish (*Fundulus heteroclitus*) inhabiting a polluted estuary. Ecological Applications 13: 490-503.

80. Hardie DC, Hebert PDN (2004) Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences 61: 1636-1646.

81. Hinegardner R (1968) Evolution of cellular DNA content in teleost fishes. American Naturalist 102: 517-523.

82. Gregory TR (2001) Animal Genome Size Database.

83. 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.

84. 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.

85. 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.

86. 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.

87. 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.

88. 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.

89. 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.

90. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714-719.

91. Margulies M, Egholm M, Altman W, Attiya S, Bader J, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376 - 380.

92. Huse S, Huber J, Morrison H, Sogin M, Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8: R143.

93. Huang XQ, Wang JM, Aluru S, Yang SP, Hillier L (2003) PCAP: A whole-genome assembly program. Genome Research 13: 2164-2170.

94. 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.

95. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, et al. (2004) The Ensembl automatic gene annotation system. Genome Research 14: 942-950.

96. Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Research 10: 516-522.

97. Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. Genome Research 10: 547-548.

98. Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5: -.

99. 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.

100. 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.

101. 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.

102. 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: -.

103. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Research 32: D277-D280.

104. 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.

105. Stein L (2001) Genome annotation: From sequence to biology. Nature Reviews Genetics 2: 493-503.

106. 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.

107. 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.

108. 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.

109. 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.

110. 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.

111. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403-410.

112. Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: Genome-wide sampling of insect populations. Annual Review of Entomology 46: 441-469.

113. 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.

114. Schlotterer C (2003) Hitchhiking mapping - functional genomics from the population genetics perspective. Trends in Genetics 19: 32-38.

115. Beaumont MA (2005) Adaptation and speciation: what can F-st tell us? Trends in Ecology & Evolution 20: 435-440.

116. 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.

117. 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.

118. Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular Ecology 14: 671-688.

119. 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.

120. 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.

121. 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.

122. 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.

123. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, et al. (2002) ARACHNE: A Whole-Genome Shotgun Assembler. Genome Research 12: 177-189.

124. Hashimoto S-i, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, et al. (2004) 5'-end SAGE for the analysis of transcriptional start sites. Nat Biotech 22: 1146-1149.

125. Dehal PS, Boore JL (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. BMC Bioinformatics 7: -.