The Landscape of Extreme Genomic Variation in the Highly Adaptable Atlantic Killifish

Noah M. Reid  
*University of California*

Craig E. Jackson  
*Indiana University*

Don Gilbert  
*Indiana University*

Patrick Minx  
*Washington University*

Michael J. Montague  
*Washington University School of Medicine*

*See next page for additional authors*

Follow this and additional works at: [https://digitalcommons.dartmouth.edu/facoa](https://digitalcommons.dartmouth.edu/facoa)

Part of the [Ecology and Evolutionary Biology Commons](https://digitalcommons.dartmouth.edu/eeb)

**Recommended Citation**

Reid, Noah M.; Jackson, Craig E.; Gilbert, Don; Minx, Patrick; Montague, Michael J.; Hampton, Thomas H.; Helfrich, Lily W.; King, Benjamin L.; Nacci, Diane E.; and Aluru, Neel, "The Landscape of Extreme Genomic Variation in the Highly Adaptable Atlantic Killifish" (2017). Open Dartmouth: Faculty Open Access Articles. 3841.  
[https://digitalcommons.dartmouth.edu/facoa/3841](https://digitalcommons.dartmouth.edu/facoa/3841)
Authors
Noah M. Reid, Craig E. Jackson, Don Gilbert, Patrick Minx, Michael J. Montague, Thomas H. Hampton, Lily W. Helfrich, Benjamin L. King, Diane E. Nacci, and Neel Aluru

This article is available at Dartmouth Digital Commons: https://digitalcommons.dartmouth.edu/facoa/3841
The Landscape of Extreme Genomic Variation in the Highly Adaptable Atlantic Killfish

Noah M. Reid,1,11 Craig E. Jackson,2 Don Gilbert,3 Patrick Minx,4 Michael J. Montague,4,12 Thomas H. Hampton,5 Lily W. Helfrich,6,13 Benjamin L. King,7,14 Diane E. Nacci,8 Neel Aluru,6 Sibel I. Karchner,6 John K. Colbourne,9 Mark E. Hahn,6 Joseph R. Shaw,2 Marjorie F. Oleksiak,10 Douglas L. Crawford,10 Wesley C. Warren,4 and Andrew Whitehead1,*

1Department of Environmental Toxicology, University of California, Davis, CA
2School of Public and Environmental Affairs, Indiana University, Bloomington, IN
3Biology Department, Indiana University, Bloomington, IN
4McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO
5Department of Microbiology and Immunology, Dartmouth College Geisel School of Medicine, Hanover, NH
6Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA
7Mount Desert Island Biological Laboratory, Salisbury Cove, ME
8US Environmental Protection Agency, Office of Research and Development, Narragansett, RI
9School of Biosciences, University of Birmingham, United Kingdom
10Department of Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, FL
11Present address: Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT
12Present address: Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
13Present address: Department of Molecular & Cell Biology, Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA
14Present address: Department of Molecular and Biomedical Sciences, University of Maine, Orono, ME

*Corresponding author: E-mail: awhitehead@ucdavis.edu.

Accepted: February 6, 2017

Data deposition: This project has been deposited at NCBI under the BioProject accession PRJNA269290 (http://www.ncbi.nlm.nih.gov/bioproject/269290) which includes genome assembly and sequences and gene transcript assembly and sequences. Accompanying the NCBI-hosted genome is NCBI’s genome annotation. We provide a GFF file (supplementary appendix S5, Supplementary Material online) to make our custom annotation (supplementary appendix S1, Supplementary Material online) compatible with the NCBI genome. Micro-RNA expression profiles have been archived at the Gene Expression Omnibus under accession number GSE70953 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70953, last accessed 14 February 2017).

Abstract

Understanding and predicting the fate of populations in changing environments require knowledge about the mechanisms that support phenotypic plasticity and the adaptive value and evolutionary fate of genetic variation within populations. Atlantic killfish (Fundulus heteroclitus) exhibit extensive phenotypic plasticity that supports large population sizes in highly fluctuating estuarine environments. Populations have also evolved diverse local adaptations. To yield insights into the genomic variation that supports their adaptability, we sequenced a reference genome and 48 additional whole genomes from a wild population. Evolution of genes associated with cell cycle regulation and apoptosis is accelerated along the killifish lineage, which is likely tied to adaptations for life in highly variable estuarine environments. Genome-wide standing genetic variation, including nucleotide diversity and copy number variation, is extremely high. The highest diversity genes are those associated with immune function and olfaction, whereas genes under greatest evolutionary constraint are those associated with neurological, developmental, and cytoskeletal functions. Reduced genetic variation is detected for tight junction proteins, which in killifish regulate paracellular permeability that supports their extreme physiological flexibility. Low-diversity genes engage in more regulatory interactions than high-diversity genes, consistent with the...
Influence of pleiotropic constraint on molecular evolution. High genetic variation is crucial for continued persistence of species given the pace of contemporary environmental change. Killifish populations harbor among the highest levels of nucleotide diversity yet reported for a vertebrate species, and thus may serve as a useful model system for studying evolutionary potential in variable and changing environments.

Key words: population genomics, genome sequence, comparative genomics, adaptation, genetic diversity.

Introduction

Phenotypic variation enables populations to persist in variable and changing environments. When individual phenotypes change in response to environmental change, this plasticity can serve to maintain or improve fitness. Alternatively, when phenotypes vary among individuals and this variation is heritable, natural selection may shape underlying genotype frequencies across generations resulting in the evolution of locally adapted phenotypes. Furthermore, when phenotypic responses that vary among individuals are heritable, phenotypic plasticity itself may be subject to evolutionary change and contribute to maintaining or elevating fitness in changing environments (Via and Lande 1985; Lande 2009). Indeed, phenotypic plasticity may interact with natural selection to either enable or disable local adaptation depending on the nature of genetic variation within populations, the costs and limits of plasticity, and gene flow (Crispo 2008). Long-standing goals for ecophysiology and evolutionary biology are to understand the genetic and genomic mechanisms that underpin phenotypic plasticity (Scheiner 1993; Via et al. 1995; Schlichting and Smith 2002) and that enable local adaptation (Savolainen et al. 2013). We develop the Atlantic killifish (*Fundulus heteroclitus*) as a genome-enabled model system for addressing such research questions; individuals exhibit high physiological plasticity, populations demonstrate local adaptation to both old and very recent environmental change, and plasticity itself has evolved within and between *Fundulus* species.

Estuarine habitats present diverse challenges for native fauna because periodic and stochastic variability in environmental parameters such as salinity, dissolved oxygen, nutrients, and temperature is common and often dramatic (Wolanski 2007), and because they are often transformed by human activities (Lotze et al. 2006). Because of this, estuarine species must exhibit much physiological resilience. *Fundulus heteroclitus* are estuarine specialists, and they are among the most abundant vertebrate animals inhabiting salt marshes along the North American Atlantic coast (Valiela et al. 1977). They are physiologically resilient to variable temperature, salinity, and dissolved oxygen, contributing to their success and numerical dominance in estuarine habitats. Accordingly, *F. heteroclitus* has emerged as an important model in physiological ecology (Burnett et al. 2007), especially for the study of ion transport (Wood and Marshall 1994) where their extreme physiological plasticity is rare among fishes (Schultz and McCormick 2013). Indeed, their resilience likely contributed to the choice of *F. heteroclitus* as the first fish to be used for experiments in outer space (von Baumgarten et al. 1975). Coupled with decades of physiology research, transcriptomics studies are revealing functional genomic mechanisms that enable physiological plasticity (Whitehead, Roach, et al. 2011; Shaw et al. 2014; Dayan et al. 2015), such that comparisons of whole-genome sequences within and between species are now poised to yield crucial insights about the genes and pathways that are important for the physiological plasticity that defines this species.

In addition to phenotypic plasticity, the evolutionary potential of species is intertwined with the abundance, distribution, and adaptive value of the genetic variation that they harbor. Yet understanding the evolutionary forces, and the features of populations and species, that govern the origin and fate of genetic variation within populations remains an enduring challenge. Species with 1) populations distributed along environmental gradients, 2) large population sizes, 3) low migration rates, and 4) closely related species that exhibit parallelism or variation in niche and population size are advantageous models to better understand the genetic basis of environmental adaptation. Atlantic killifish (*F. heteroclitus*) have all of these attributes. As mentioned, environmental parameters are highly variable within individual coastal estuaries, yet these parameters also vary among different sites. Living along the east coast of North America, *F. heteroclitus* are subject to gradients in temperature, salinity, dissolved oxygen, and environmental pollution. As *F. heteroclitus* populations are large with low migration (Able et al. 2012), selection can be efficient such that populations may evolve adaptations to local environments. Indeed, *F. heteroclitus* populations have derived variable tolerance to temperature and salinity extremes (e.g., Scott et al. 2004; Fangue et al. 2006; Whitehead, Roach, et al. 2011). They have also rapidly and repeatedly evolved adaptations to abruptly and radically altered habitats that are contaminated with industrial pollutants (Van Veld and Nacci 2008). Accordingly, *F. heteroclitus* is a classic model for understanding evolution along environmental gradients (Powers et al. 1993) and is a novel vertebrate model for studying rapid adaptation to human-induced environmental change and evolutionary rescue (Bell 2013; Reid et al. 2016). Advances in high-throughput DNA sequencing have enabled the characterization of genetic variation across the whole genome for many individuals within a population, opening the door to the...
development of new model species for studying environmental adaptation (Leffler et al. 2012).

We have developed a well-annotated F. heteroclitus reference genome and characterized whole-genome variation among 48 individuals within a population located near the northern end of their range. Our goal was to produce a high-quality reference genome, with careful annotation of functional features including protein-coding genes, mobile elements, and microRNAs (miRNAs). By comparing genome sequence variation with other fish species, we identify genes and processes that may contribute to the physiological plasticity and adaptability for which F. heteroclitus is known. By characterizing genome-wide sequence variation among many individuals within a population, we model demographic history, reveal surprisingly high genetic variation, and offer insight into the evolutionary forces governing the fate of this variation.

Materials and Methods

Genome Sequencing and Assembly

Four sets of sequencing libraries were prepared from DNA fragments of different sizes including 300-bp, 3-, 8-, and 40-kb (fosmid) fragments. One individual female fish provided source DNA for the 300-bp, 3-, and 8-kb libraries (captured from Chewonki Campground, ME: 43°57’15.01”N, 69°43’13.64”W in August, 2010), whereas a second individual (captured from Northeast Creek, ME: 44°25’29.04”N, 68°19’36.91”W in July, 2010) provided source DNA for the 40-kb fosmid libraries. Fish from Maine were chosen because they have the lowest heterozygosity among F. heteroclitus populations (Duvernell et al. 2008). All sequences were 100-bp paired-end reads collected on an Illumina HiSeq2000. The combined sequence reads were assembled using the ALLPATHS-LG assembler (Gnerre et al. 2011) (release 44849), followed by local scaffold gap filling with Illumina data (300 bp) (Tsaí et al. 2010) using PyGap perl scripts (available upon request) and removal of contaminating contigs. To remove contaminating contigs we aligned them to a custom database (at the McDonnell Genome Institute [MGI]) consisting of microbial sequences, vector sequences, and sequences from organisms sequenced at the MGI to identify project cross-contamination, or human contamination. Additional removal and trimming of sequences was determined by the automated NCBI pipeline. All sequences identified by NCBI were manually removed. All assembled sequences less than 200 bp were removed from the assembly as required by NCBI.

Gene Model Prediction and Annotation

We developed and implemented software methods to incorporate data from expressed sequences and protein orthology to human and fish into evidence-based gene models. These gene reconstruction methods are a hybrid of genome gene prediction (EvidentialGene-G; Gilbert 2012) and mRNA assembly (Evigene-R; Gilbert 2013), which produce gene models from independent lines of evidence. EvidentialGene methods have produced accurate gene sets for a number of animal and plant genome projects, and are publicly available (EvidentialGene 2012). We assessed model quality by scoring all models per locus for gene evidence from all sources, then selecting locus representatives that maximize evidence supporting that locus, weighted by evidence quality (e.g., protein homology and RNA-seq intron recovery scores weight above protein or transcript sizes). This assessment was followed by error analyses using orthology and expression evidence, expert evaluations of genes, refinement of model parameters, and some manual curation, contributing to sequential refinement of the gene set. All gene models were supported by expression and/or homology evidence; ab initio predictions without expression or homology evidence were not included. Detailed methods for gene model prediction and annotation are in the supplementary methods, Supplementary Material online. A table of attributes for the final gene set, with associated evidence scores, is available in the supplementary appendix S1, Supplementary Material online. A GFF file of our gene annotations that is compatible with the NCBI-hosted genome is provided in the supplementary appendix S5, Supplementary Material online.

We also annotated the miRNAs; small (19–25 nt) noncoding RNAs that are involved in posttranscriptional regulation of gene expression that play important roles in physiology and development (Mishima 2012). In this study, miRNA precursor sequences were identified using Basic Local Alignment Search Tool (BLAST) searches of known miRNAs from teleost fish species from miRBase (version 21) (Kozomara and Griffiths-Jones 2014) including Danio rerio (N=349 precursor sequences), Takifugu rubripes (N=131), Tetraodon nigroviridis (N=132), and Oryzias latipes (N=168). Secondary structures of the precursor sequences were checked using MiPred (Jiang et al. 2007). A small-RNA-seq library from developing F. heteroclitus embryos (GEO data set GSE70953) sequenced using methods described in Jenny et al. (2012) also was scanned for the presence of mature miRNA sequences. Clustered miRNAs were determined to be those that were less than 36 kb apart. This is the distance between miR-133 and miR-1: Two miRNAs that bound a cluster that is conserved across deuterostomes (Campo-Paysaa et al. 2011).

Fundulus heteroclitus gene models were linked to UniProt accession numbers (supplementary appendix S1, Supplementary Material online) to provide a handle for gene ontology (GO) enrichment analyses through commonly used tools, such as DAVID Bioinformatics Resources (Huang et al. 2009) and Ingenuity Pathway Analysis (Qiagen). Orthology (from OrthoMCL analysis) to human and zebrafish was used in that order of priority as UniProt entries for those species tend to have good assignment of GO terms, UniProt accessions for these species are retrievable through commonly used databases such as DAVID, and Ingenuity Pathway Analysis is based mainly on human annotations.
Repetitive Element Analysis

Repetitive elements in the *F. heteroclitus* genome were annotated using both a de novo approach to identify *F. heteroclitus* lineage-specific transposable elements (TEs) and a genome-wide search of known teleost fish TEs. RepeatModeler (Smit and Hubley 2008–2015), which utilizes the programs RECON (Bao and Eddy 2002), RepeatScout (Price et al. 2005), and Tandem Repeat Finder (Benson 1999), was used to conduct the de novo TE search (~220 Mb of the genome assembly sampled to build de novo TE models). RepeatMasker (Smit et al. 2013–2015) was used to conduct a genome-wide exhaustive search combining the known teleost fish and ancestral TEs from RepBase Update (Jurka et al. 2005) with the background set included the genes showing signatures of selection along the *F. heteroclitus* lineage and the background set included the 7,042-gene input set. GO enrichment was also examined using GOrilla (Eden et al. 2009). As DAVID and GOrilla results were highly congruous, only results from DAVID analyses are reported here. To further validate genes determined to be under positive selection, amino acid substitutions that were predicted to impact protein function were assessed with PROVEAN (Choi et al. 2012) using default run parameters and a significance threshold score of −1.3. Genes showing signatures of selection were also included in gene interaction network analysis using the Ingenuity Pathways Analysis database of interactions (settings: direct relationships, all data sources, all species, confidence experimentally observed).

Phylogenetic Analysis

To identify signatures of selection in the lineage leading to *F. heteroclitus*, genome-wide analyses of nonsynonymous/synonymous (dN/dS) substitution ratios were calculated using alignments of orthologous protein-encoding sequences from six fish species: zebrafish (*D. rerio*), stickleback (*G. aculeatus*), tilapia (*Ore. niloticus*), zebra mbuna (*M. zebra*), platyfish (*K. maculatus*), and Atlantic killifish (*F. heteroclitus*). Coding sequences from all taxa were downloaded from Ensembl or NCBI BLAST for repetitive elements (RMBlast) was used to include well-annotated TEs (Schartl et al. 2013; Brawand et al. 2014). The modified version of NCBI BLAST for repetitive elements (RMBlast) was used to conduct the genome-wide search. The *F. heteroclitus* TE landscape was compared with seven other teleost genomes that include well-annotated TEs (*D. rerio*, *X. maculatus*, *O. latipes*, *M. zebra*, *Orecromis niloticus*, *Gasterosteus aculeatus*, *T. nigroviridis*) (Jaillon et al. 2004; Kasahara et al. 2007; Jones et al. 2012; Howe et al. 2013; Schartl et al. 2013; Brawand et al. 2014).

Genome-Wide Population Genetic Diversity

The pattern of genomic variation within a species provides a window into its history of natural selection and demography. To assess these factors in *F. heteroclitus*, we resequenced 48 individuals collected from Great Salt Pond on Block Island, RI (41°10’54.51"N, 71°34’45.50"W). Genomic DNA from individual fish was barcoded, pooled, and sequenced to produce 4.5 billion reads (100-bp paired end) on an Illumina 2500. Reads were mapped to the reference genome using bowtie2 (Langmead and Salzberg 2012). All subsequent analyses were restricted to read pairs that mapped properly, had phred-scaled map qualities >30 and bases with quality >20. Coverage depth per base was assessed across the reference genome using SAMtools (Li et al. 2009), excluding approximately 5 Mb with aberrantly high coverage (>2× the mode). Effective coverage of our assembled genome averaged 5.24× per individual (supplementary table S1, Supplementary Material online). Short variants (single nucleotide polymorphisms and indels) were called using FreeBayes (Garrison and Marth 2012). An allele frequency spectrum was estimated using the 20 largest contigs in the genome (encompassing 50 Mb) using an expectation maximization algorithm implemented in ANGSD (Korneliussen et al. 2014). ANGSD was also used to estimate genetic diversity statistics (*r*, Watterson’s theta, and Tajima’s *D*) using an empirical Bayesian approach that accounts for uncertainty in variant calling and genotyping at individual loci (Korneliussen et al. 2013). Individual-level inbreeding coefficients were calculated using ngsF (Vieira et al. 2013). Estimates of genetic diversity were processed over several types of genomic intervals using bedtools (Quinlan and Hall 2010): 1) Sliding windows...
(50-kb window/10-kb step and 5-kb window/1-kb step), 2) coding sequence by gene region (CDS) with codon positions calculated independently, 3) noncoding sequence by gene region (non-CDS) including introns and untranslated regions (UTRs), 4) proximal promoter regions (defined as 1 kb upstream of the start codon) by gene region (Pro), 5) miRNAs, and 6) intergenic regions. The analysis was restricted to intervals longer than 150 bases. Enrichment of functional classes of genes in the 5% and 95% tails of the distributions of \( \pi \) and Tajima’s \( D \) was tested using DAVID (Huang et al. 2009). Historical demographic models were fit to the estimated allele frequency spectrum using \( \text{fastsimcoal2} \) (Gutenkunst et al. 2009; Excoffier and Foll 2011; Excoffier et al. 2013). Models fit included single-population models with up to four epochs of constant population size, a two-population model with migration and an isolation-with-migration model. The program ms (Hudson 2002) was used to simulate distributions of our diversity statistics in 50-kb windows.

**Genome-Wide Population Variation in Copy Number**

For each of the 48 Block Island individuals, copy number variants (CNVs) were called based on the combination of signals from a read-pair approach (lumpy), split-read approach (lumpy), and read depth approach (CNVnator) (Abyzov et al. 2011; Layer et al. 2014). Deletions and duplications were defined as the decrease or increase of copy-number relative to the reference genome. Final CNV calls required support from all three approaches and were therefore conservative. Discordant alignments (insert size all three approaches and were therefore conservative. final CNV calls required support from both the read-pair module and split-read module. Copy number variants identified in multiple individuals were merged using a 50% reciprocal overlap approach as the estimated breakpoints of CNVs can vary between individuals, yielding population CNV regions. Each population CNV region was copy-number genotyped in all individuals using CNVnator’s genotyping module after partitioning read depth bins for each individual using a 100-bp sliding window. The CNVnator genotyping module computes the estimated copy number of the given locus compared with the reference genome, such that each reference locus was considered to be in a diploid state with two single-copy alleles. The estimated copy numbers from CNVnator were rounded to the closest integer as a proxy for copy-number. We retained all putative biallelic CNV loci, and final CNV calls were generated from putative lumpy CNV calls that were also supported by CNVnator read depth genotyping in at least one individual. CNV allele frequencies in the population were determined from the CNVnator copy-number genotyping calls for all 48 individuals. To estimate nucleotide diversity (\( \pi \)) within CNV regions, we computed \( \pi \) in 500-bp windows across the genome using the window-pi function in VCFtools (Danecek et al. 2011). We separated these 500-bp windows into three categories: Those overlapping or contained within deletion regions, those overlapping or contained within duplication regions, or those outside of CNV regions.

**Results and Discussion**

**Genome Sequencing and Assembly**

Based on flow cytometry and bulk fluorometric analysis, the estimated \( F. \) heteroclitus haploid genome size is between 1.3 and 1.5 Gb (Hinegardner 1968; Dawley 1992; Hardie and Hebert 2004), which is intermediate in size between Japanese medaka and zebrafish (Gregory 2001). Our total sequence coverage was 85× from 300-bp, 3-, 8-, and 40-kb libraries using a genome size estimate of 1.3 Gb. The assembly spans 932 Mb and is made up of 10,180 scaffolds with an N50 contig and scaffold length of 16.7 kb and ~1.2 Mb, respectively. Contiguity measures of the \( F. \) heteroclitus genome are similar to those of other teleost fish genomes that have been assembled using similar strategies (e.g., Illumina sequence only and using the same assembly algorithm; supplementary table S2, Supplementary Material online). The Fundulus heteroclitus-3.0.2 assembly is available under GenBank accession number JXMV00000000.1. Extensive assembly statistics are available at http://www.ncbi.nlm.nih.gov/assembly/GCF_000826765.1 (last accessed 14 February 2017). Assembled gene representation as measured by CEGMA (Parra et al. 2007) showed that 235 of 248 (95%) ultraconserved eukaryotic genes are present in the genome as complete sequences, where the remaining 13 are present as partial sequences (supplementary table S3, Supplementary Material online), demonstrating sufficient genomic coverage for the studies described herein.

**Gene Model Prediction and Annotation**

The \( F. \) heteroclitus genome has data supporting 34,924 gene loci: 27,512 (79%) from mRNA assembly, 7,407 (21%) from genome gene prediction (table 1; supplementary appendix S1, Supplementary Material online), and 5 are curated models of orthologs. For these gene loci, mRNA sequences provide the most convincing evidence. For the 27,512 genes with strong mRNA evidence, 73% (20,067) are complete proteins and 76% are orthologs or paralogs (table 1). For the 7,407 gene models that do not have corresponding complete transcript assemblies, 96% have measurable expression. Also 4,097 (55%) of these have orthologs or paralogs in other species.
(3,054 are orthologs, and 1,043 are paralogs). Five genes were manually curated from ortholog proteins mapping to this genome that lacked corresponding gene models or full transcripts (four are odorant receptor genes common to fish and human). Thus, 31,614 (91%) genes have mRNA sequences, or are orthologs or paralogs. This leaves 3,310 genes with less evidence; yet nearly all of these gene models (94%) are complete proteins and have low counts of RNA-seq reads that map to the gene model from RNA-seq studies.

In comparison to humans (fig. 1), F. heteroclitus has 17,072 orthologs, on average 1,899 (range 201–3,991) more than the other nine fish species that we used for comparison. In comparison to other fish and humans, F. heteroclitus has 24,801 homologs (21,122 orthologs plus 3,679 paralogs) (table 1). Of the 24,801 genes supported by homology evidence, 85% (20,702) are also supported by mRNA sequence assemblies. Fundulus heteroclitus also has 10,123 genes with weak evidence for orthology with other species (specified in the supplementary appendix S1, Supplementary Material online, as “noor” or “udup,” and as “unclassified” in the supplementary table S4, Supplementary Material online), but 6,810 (67%) of these are supported by evidence from mRNA assemblies. The remaining 3,313 of genes with weak orthology are supported by evidence from gene models or expert annotation.

The quality of the F. heteroclitus genome annotation can be assessed by determining the number of orthologous genes (Waterhouse et al. 2013) it contains and how many are missing. The F. heteroclitus genome has the most orthologous genes (21,122) when compared among the other fish and human genomes (supplementary table S4, Supplementary Material online). Fundulus heteroclitus is only missing ten of the orthologous genes found in all other fish species (supplementary table S4, Supplementary Material online: Missing Orthologs) whereas other species are missing from 54 to 1,514. These data suggest that the annotated genes in other fish species are most often also annotated in F. heteroclitus.

We compared the frequencies of shared genes among species (supplementary table S5, Supplementary Material online) by first calculating the maximum number of genes shared between each species and two other species (see first data column [Max Shared with 2 Spp orthologs] in the supplementary table S5, Supplementary Material online). We find that the F. heteroclitus genome has 15,415 genes that have orthologs in at least two other species, which is the most genes for all comparisons. For each species, we then calculated the proportion of two-species orthologs that are present between all species pairs. This comparison creates an asymmetric matrix where comparisons are across rows. Thus F. heteroclitus has 100% of platyfish and medaka orthologs (supplementary table S5, Supplementary Material online). That is, all of platyfish and medaka genes are found in F. heteroclitus. Yet, platyfish (X. maculatus) has only 94% of F. heteroclitus orthologs and medaka (O. latipes) has 82% of F. heteroclitus orthologs. Supplementary table S5, Supplementary Material online, demonstrates that among genes shared by two species, F. heteroclitus has all of these. A number of notable human genes that are conspicuously missing in the zebrafish genome (Howe et al. 2013) are present in the killifish genome, including BRCA1 and interleukin-6 (IL6). The orthoDB database (Waterhouse et al. 2013) has been updated to include F. heteroclitus genes (genes draft version 1, at http://cegg.unige.ch/orthodb7, last accessed 14 February 2017).

To summarize, F. heteroclitus has approximately 35K gene loci, and 90% have mRNAs or homologs in other species. Importantly, we have identified gene models for 100% of the genes found in seven of nine other species and for 98% in the other two species. Thus, the F. heteroclitus genome is well annotated; it includes a more complete gene set than for other currently available fish genomes.

---

**Table 1**  
Gene Set Numbers (version kfish2rae5g, Dec 2013)

<table>
<thead>
<tr>
<th>34,924 Genes</th>
<th>Total</th>
<th>Orthology</th>
<th>Mapping</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Orthologs to Other Species</td>
<td>Inparalogs of Orthologs</td>
<td>Weak Homology or Are Species-Unique (unclassified)</td>
<td>Properly Mapped to Genome (&lt;80% coverage)</td>
</tr>
<tr>
<td>Total</td>
<td>34,924</td>
<td>21,122</td>
<td>3,679</td>
<td>10,123</td>
</tr>
<tr>
<td>Evidence from mRNA assemblies</td>
<td>27,512</td>
<td>18,066</td>
<td>2,636</td>
<td>6,810</td>
</tr>
<tr>
<td>Evidence: are genome-modeled</td>
<td>7,407</td>
<td>3,054</td>
<td>1,043</td>
<td>3,310</td>
</tr>
<tr>
<td>Expert added</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

(100% = 18,066, 60% = 11%, 29%, 21% = 3,679, 41% = 10%, 25%, 21% = 3,310, 41% = 1,043, 45%, 21% = 1,034, 0% = 3,054, 0% = 1,043, 0% = 3,310, 0% = 1,034).

---

Advan ce Access publication February 13, 2017
High Repetitive Element Content

The *F. heteroclitus* genome assembly harbors an overall repeat content of 33%, which is among the largest for known teleost fish genomes and second only to *D. rerio* (fig. 1). Teleost genomes show wide variation in genome size and repeat content, where genome sizes range 4-fold from 310 Mb (*T. nigroviridis*) to 1.4 Gb (*D. rerio*), with genome repeat content varying from <1% (*T. nigroviridis*) to 55.2% (*D. rerio*) (fig. 1). Using seven published teleost genomes, we found a significant positive correlation between teleost genome assembly size and repeat content (Pearson’s correlation, \( r = 0.955, n = 8, P < 0.001 \); supplementary fig. S1, Supplementary Material online), consistent with expansion and contraction of TEs contributing to expansion and contraction of genome sizes within fish. All genome assemblies are currently incomplete partly because repeat elements complicate assembly, such that current highly fragmented genome assemblies probably underestimate the total repeat content.

As in other teleosts, TEs in *F. heteroclitus* were highly diverse, including many families present in fish but absent in mammals (Volf et al. 2003). The *F. heteroclitus* TE landscape includes at least 22 DNA transposon families, 13 long interspersed nuclear element (LINE) families, and 7 long terminal repeat element families. In total, 1.46 million TEs were identified in the *F. heteroclitus* genome (supplementary table S6, Supplementary Material online). In contrast to mammalian genomes where LINE retrotransposons are the most abundant (Han and Boeke 2005; Wicker et al. 2007), DNA transposons were the most abundant type of TE in the *F. heteroclitus* genome, with >587,000 copies representing 12.7% of the genome (130 Mb). The most abundant families of DNA transposons included the Tc1–Mariner superfamily (>387,000 copies, 8.24% of genome) and the hAT superfamily (>135,000 copies, 2.84% of genome). Retrotransposons represent 9.1% of the *F. heteroclitus* genome (93.5 Mb), with the most abundant families consisting of the LINE L2 element (>127,000 copies, 3.25% of genome) and the LINE Rex-Babar element (>62,000 copies, 1.65% of genome). We found that 2.2% of the *F. heteroclitus* transcriptome was derived from TE sequences, indicating that some *F. heteroclitus* TEs are likely still active.

Over 500,000 novel *F. heteroclitus* lineage-specific TEs, which are diverged from currently known TEs, were identified in the *F. heteroclitus* genome, representing 11.1% of the
genome (113.5 Mb). These we refer to as novel because they could not be automatically placed into a known annotated TE family. The *O. latipes* genome also contained a similar percentage of novel repeats (Kasahara et al. 2007), suggesting that several uncharacterized TE families are present in teleost fish genomes. With the large amount of DNA transposons relative to mammals and many novel repeat element families, teleost genomes provide a new opportunity for the comparative study of repetitive element evolution and diversification.

**Positive Selection with Cell Cycle Gene Enrichment**

Orthology across six fish species produced 7,042 genes that were used for phylogenetic tests for positive selection. PAML analysis (codeml package) revealed 649 of 7,042 genes (9%) showing signatures of positive selection along the lineage leading to *F. heteroclitus* (supplementary table S7, Supplementary Material online). This gene set is significantly enriched for genes associated with “cell cycle regulation” (*N* = 46, *P* < 0.0001) (particularly “M-phase of mitotic cell cycle”); *N* = 24, *P* < 0.001) and “regulation of cell death” (apoptosis; *N* = 50, *P* < 0.001) functions (supplementary table S8, Supplementary Material online). A subset of these genes has amino acid substitutions that are predicted to impact protein function (PROVEAN analysis), and these are also enriched for the same functions. These genes form a highly connected interaction network (supplementary fig. S2, Supplementary Material online) implying parallel evolution of a functional module of genes. This network retains significant enrichment for “cell cycle” and “cell death and survival” functions, in particular for apoptosis of epithelial cells (*P* < 0.001).

*Fundulus heteroclitus* are euryhaline and can quickly remodel the physiology and morphology of their gill epithelium to maintain osmotic homeostasis in the face of very broad changes in environmental salinity (Wood and Marshall 1994). This ability enables a level of extreme physiological plasticity that is rare among fishes (Schultz and McCormick 2013). Cell cycle arrest at the G2/M-phase of mitosis is activated during osmotic challenge in diverse species, as is apoptosis during gill epithelial remodeling in euryhaline fish (Kammerer and Kultz 2009). Genes and gene networks associated with cell cycle regulation and apoptosis are transcriptionally regulated during acclimation to osmotic challenge in *F. heteroclitus* (Whitehead et al. 2012, 2013). These observations are consistent with the hypothesis that the evolved network of genes that we report here contributes to the exceptional ability of killifish to remodel gill epithelia in response to variable salinity, and thereby supports their persistence and dominance in osmotically dynamic environments as well as across the continuum of osmotic niches within coastal ecosystems. These data set the stage for network-based analyses to characterize and explore the genomic infrastructure that supports the extraordinary physiological plasticity of *F. heteroclitus*.

**High Genome-Wide Population Genetic Diversity**

Among 48 fish within a single population, the genome-wide average nucleotide diversity (*π*) is 0.0119, whereas genome-wide Tajima’s *D* is −0.1130 (bootstrapped 95% confidence interval: −0.1185, −0.1078). The allele frequency spectrum is not consistent with a population in equilibrium, reflected by the negative Tajima’s *D*. To test some possible causes of this pattern, we conducted several analyses. A small number of newly immigrated individuals from a differentiated population could skew the allele frequency spectrum in this way, a pattern that might manifest as excess homozygosity within individuals. We first estimated individual-level inbreeding coefficients, which ranged from 0 to 0.13, with an average of 0.03. We interpret this as evidence that we have not sampled any recent migrants, or across any substantial population breaks. Additionally, a plot of genetic distances between individuals suggests no outliers. Demographic history may also impact this pattern, so we tested demographic models to explain this shift in the site frequency spectrum, including competing models that differed in the number of discrete historical epochs and that exchanged migrants with an unsampled population. Each epoch could model stable population size, exponential expansion or contraction, or migration with an unsampled population. Models were compared by fit to the data, as well as by stability and plausibility of parameter estimates. The best model was composed of three epochs showing historical population growth from the ancestral population through the first two epochs (accounting for negative Tajima’s *D*), followed by a substantial reduction in the most recent epoch. If we assume a per base mutation rate on the order of 10−8 per generation (Burrige et al. 2008), our model suggests a reduction of 50–80% in effective population size since the last glacial maximum (LGM). This model implies a reduction of genetic diversity following the LGM, which is consistent with a northern refugium or a founder effect from northward expansion following the LGM (Duvernell et al. 2008). Though this model provided a good fit to the allele frequency spectrum, it poorly reproduced the distribution of summary statistics across 50-kb windows (supplementary fig. S3, Supplementary Material online), particularly the long negative tail of Tajima’s *D* and low *π*. This implies either a highly complicated demographic history, or that nonneutral processes, such as background selection and selective sweeps, have impacted a large fraction of genetic variation and this signature is superimposed upon a signal of demographic history.

Genome-wide nucleotide diversity measures are higher for *F. heteroclitus* (*π* = 0.0119) than for other species for which comparable whole-genome data are available. For example, genome-wide average *π* in *F. heteroclitus* is 4.8-fold higher than that among six marine stickleback fishes (*π* = 0.0025) (Feulner et al. 2013). Similarly, variation in *F. heteroclitus* is higher than variation among black cottonwood trees, rice, wild and cultivated maize, *Drosophila melanogaster*.
yeast isolates, crows, and Darwin’s finches (Liti et al. 2009; Sackton et al. 2009; Kolaczkowski et al. 2011; Huang et al. 2012; Hufford et al. 2012; Pool et al. 2012; Evans et al. 2014; Poelstra et al. 2014; Lamichhaney et al. 2015). Placed within the context of a recent broad comparison of genetic diversity across the tree of life, where estimates of π were calculated from a few loci to genome-wide depending on the species, F. heteroclitus ranks as the most genetically diverse among fishes, and ranks in the 98th percentile among vertebrates (Leffler et al. 2012); F. heteroclitus diversity falls firmly within the range of arthropods and at the extreme upper end of the diversity range of vertebrates (median π for arthropod and chordate species was 0.0125 and 0.0026, respectively) (Leffler et al. 2012). Furthermore, the Block Island (RI) population we assess here is from the northern end of the species range, where genetic diversity is lower than in the south (Duvernell et al. 2008). Indeed, nucleotide diversity in coastal Virginia populations of F. heteroclitus is higher (π=0.016) (Reid et al. 2016).

High nucleotide diversity within a F. heteroclitus population is consistent with the pattern of positive correlation between population census size and genetic diversity (Leffler et al. 2012). Within their coastal marsh habitats F. heteroclitus are the numerically dominant fish (Yozzo and Smith 1998) where densities are typically ~14 fish per square meter (Sweeney et al. 1998) and neighborhood sizes can range into the many tens of thousands (Meredith and Lotrich 1979; Teo and Able 2003). Very large populations can harbor much standing genetic variation that is a crucial substrate for adaptation to novel environments, thereby relaxing reliance on new mutations and alleviating the variation-reducing influence of selection on new mutations (Hermisson and Pennings 2005). Indeed, F. heteroclitus populations have recently and quickly evolved extreme tolerance to the normally lethal effects of highly polluted environments (Nacci et al. 2010), suggesting the adaptive importance of standing genetic variation within this species (Reid et al. 2016). However, repeated selection may reduce variation at linked sites through genetic hitchhiking (Maynard Smith and Haigh 1974) or background selection (Charlesworth et al. 1993) and may thereby reduce neutral diversity to levels lower than one might expect based on population size. For both genetic hitchhiking and background selection, recombination is important for determining the size of the footprint of reduced genetic variation. Genetic hitchhiking and background selection may be weak within F. heteroclitus, leading to higher genetic variation, because recombination rates are relatively high. This has yet to be tested in a robust comparative framework (Romiguier et al. 2014).

Across the genome, coding regions have less nucleotide diversity compared with all classes of noncoding regions (mean coding π=0.0060, mean noncoding π=0.0122, Mann–Whitney U test P<2.2 × 10^-16) consistent with coding regions being subject to greater selective constraint than noncoding regions (fig. 2, table 2). In contrast, among classes of noncoding DNA (introns and UTRs within gene models, proximal promoter, intergenic) distributions of nucleotide diversity vary by only ~1%. Similarly, Tajima’s D is substantially more negative for coding regions (D=−0.5937) than for all classes of noncoding sequence (mean D=−0.1235, Mann–Whitney U test P<2.2 × 10^-16). This indicates that coding regions have more rare variation than noncoding regions. We interpret this discrepancy to be a result of weak purifying selection failing to efficiently purge rare weakly deleterious alleles (Fay et al. 2001).

Though the pervasive reduction in Tajima’s D in coding sequence compared with noncoding sequence is best explained by the action of purifying selection on weakly deleterious mutations, we also find that genetic variation within
coding sequence is correlated with both variation in linked genic noncoding sequences ($R^2=0.43$) and variation in linked promoter sequences ($R^2=0.29$) (fig. 3A and B). This correlation is consistent with the expectation that the evolutionary fates of DNA sequence variation in physically linked regions of the genome are coupled and suggestive of a role for linked selection in shaping patterns of genetic variation in this population (Maynard Smith and Haigh 1974; Begun and Aquadro 1992). Indeed, we find evidence for linked selection in *F. heteroclitus*, insofar as Tajima’s $D$ negatively correlates with gene density (supplementary fig. S4, Supplementary Material online). In examining the pattern of correlation in Tajima’s $D$, the overall mass of Tajima’s $D$ is above the 1:1 line for coding versus noncoding sequence (fig. 3D and E), which is consistent with weak purifying selection of greater influence in coding versus noncoding DNA. However genes within the lower left quadrant of the plots (fig. 3D and E) tend to converge on a 1:1 correlation, where Tajima’s $D$ is very low for both coding and linked noncoding sequences, which is consistent with a single dominant process (such as selective sweeps) influencing variation in regions of the genome that are linked with coding sequence. We infer pervasive roles for both weak purifying selection and adaptive selective sweeps for shaping coding sequence variation within this population, and that these processes likely account for variation that is not efficiently explained by a purely demographic model (supplementary fig. S3, Supplementary Material online).

**Low-Diversity Genes Are Associated with Neurological, Developmental, and Cytoskeletal Functions, and Tight Junctions**

We tested whether genes with shared functional attributes clustered within the tails of the diversity distributions. Genes (coding sequence) with the lowest diversity among individuals (those genes that occupy the 5th percentile $\pi$ measurements) tend to be associated with biological processes that include neurological systems and development (fig. 2). Within the context of neurological systems, this gene set was significantly enriched for GO terms including “neurological system process” ($N=31$, $P<0.0001$), “neuron differentiation” ($N=46$, $P=0.002$), “synapse” ($N=39$, $P=0.001$) including cholinergic receptor, GABA receptor, glutamate receptor, and insulin receptor genes, “sensory organ development” ($N=31$, $P<0.0001$), and “ion transport” ($N=63$, $P=0.003$) including sodium, chloride, calcium, and potassium channel genes—both voltage and ligand gated—and solute carrier genes. Within the context of developmental processes, this gene set was significantly enriched for GO terms including “HOX” genes ($N=29$, $P<0.001$), “fore-brain development” ($N=20$, $P=0.013$), and “embryonic morphogenesis” ($N=39$, $P<0.001$). Other significantly enriched GO terms included “actin cytoskeleton” ($N=34$, $P<0.0001$) including actin, actinin, myosin, tropomyosin, and troponin subunits, “protein transport” ($N=72$, $P=0.002$), and “protein kinase activity” ($N=52$, $P=0.048$). Since $\pi$ and Tajima’s $D$ are correlated, and $\pi$ of coding sequences is correlated with $\pi$ of linked noncoding genic and promoter sequences (fig. 3), we find that GO terms enriched in the tails of each of these distributions are highly overlapping, so we report here only data from the tails of the coding sequence distributions. We also compared nucleotide diversity between genes that were expressed only in adult gill tissues versus those expressed only in embryos (Illumina RNA-seq data, threshold for detection of expression=10 reads per kilobase per million; supplementary appendix S1, Supplementary Material online) and found that diversity was lower for genes with embryo-specific expression (supplementary fig. S5, Supplementary Material online). This indicates that genes expressed specifically during early developmental stages are more likely to be involved in conserved biological processes compared with those expressed in specialized adult tissues.

Low-diversity genes within this *F. heteroclitus* population tend to be associated with neurological, developmental, and cytoskeletal functions, which is consistent with findings in other species. For example, in the *Drosophila melanogaster* genetic reference panel, those GO terms that were enriched among the lowest diversity genes included “synapse, ion transport, appendage development”, and “actin cytoskeleton” (Mackay et al. 2012). Similarly, among primates, the most slowly evolving genes include those preferentially expressed in the brain, and those enriched for functions including neurogenesis, synaptic transmission, and intracellular signaling (Mikkelsen et al. 2005). This suggests that these processes, and the genes that govern them, are subject to shared selective constraint in diverse taxa. Surprisingly, neurological system genes are not among the most slowly evolving genes in Japanese medaka (Kasahara et al. 2007), which are closely related to *F. heteroclitus*.

**Table 2** Genetic Diversity among 48 Individuals within a Population

<table>
<thead>
<tr>
<th>Nucleotide Diversity ($\pi$)</th>
<th>Waterson’s Theta per Site</th>
<th>Tajima’s $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome-wide</td>
<td>0.0119</td>
<td>0.0124</td>
</tr>
<tr>
<td>Genic—coding</td>
<td>0.0060</td>
<td>0.0072</td>
</tr>
<tr>
<td>Genic—noncoding</td>
<td>0.0122</td>
<td>0.0126</td>
</tr>
<tr>
<td>Intergenic</td>
<td>0.0122</td>
<td>0.0125</td>
</tr>
<tr>
<td>Proximal promoter*</td>
<td>0.0123</td>
<td>0.0129</td>
</tr>
<tr>
<td>Codon position 1</td>
<td>0.0040</td>
<td>0.0056</td>
</tr>
<tr>
<td>Codon position 2</td>
<td>0.0034</td>
<td>0.0049</td>
</tr>
<tr>
<td>Codon position 3</td>
<td>0.0101</td>
<td>0.0109</td>
</tr>
<tr>
<td>5′-UTR</td>
<td>0.0083</td>
<td>0.0095</td>
</tr>
<tr>
<td>3′-UTR</td>
<td>0.0096</td>
<td>0.0105</td>
</tr>
<tr>
<td>miRNAs</td>
<td>0.0012</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

*Proximal promoter is a subset of intergenic.*
Although \( \pi \) and Tajima’s \( D \) are often correlated, one notable exception is “tight junction” proteins, which are not enriched in the \( \pi \) distribution but are enriched (\( N=13, P=0.008 \)) in the 5th percentile of the Tajima’s \( D \) distribution for coding sequences. This enrichment is consistent with selective constraint on regulation of paracellular permeability in this euryhaline estuarine specialist. Atlantic killifish are unusual among fish insofar as they appear to lack the mechanisms that would enable active chloride uptake to maintain osmotic homeostasis in low-salinity environments; instead, they adjust paracellular permeability of gill epithelia to limit ion loss in low-salinity conditions (Patrick et al. 1997; Patrick and Wood 1999; Scott et al. 2004). Claudin proteins are localized to gill epithelial tight junctions and regulate paracellular permeability and ion...
selectivity (Van Itallie and Anderson 2006). Transcripts of tight junction proteins including claudins are consistently differentially regulated during both hypo- and hyperosmotic acclimation in killifish (Whitehead, Galvez, et al. 2011; Whitehead, Roach, et al. 2011; Whitehead et al. 2012, 2013; Brennan et al. 2015) and show signatures of diversifying selection between populations native to different osmotic environments (Kozak et al. 2014). Claudins, and other tight junction components, are likely important proteins that contribute to the extreme physiological plasticity and estuarine success that distinguishes Fundulus species.

miRNAs Are among the Least Variable of Genomic Elements

We identified 197 precursor miRNA sequences in the _F. heteroclitus_ genome (supplementary appendices S2 and S3, Supplementary Material online), 126 of which were detected as mature miRNAs expressed in developing embryos. Average nucleotide diversity of the 197 miRNA loci is the lowest among genomic regions compared (table 2): average \( \pi = 0.0012 \) is approximately 3-fold less than at first and second codon positions and nearly 10-fold less than at third codon positions. Similarly, average Tajima’s \( D \) for miRNAs was very low (−1.99), at the extreme negative end of the genome-wide distribution (supplementary fig. S6, Supplementary Material online). To test whether discovery bias might have skewed these results, we compared the distribution of diversity statistics for miRNA in humans (Quach et al. 2009) to the human miRNA subset that is homologous to _F. heteroclitus_ miRNAs. We find that within humans, the genetic diversity of _F. heteroclitus_ miRNA orthologs is representative of the genetic diversity of all miRNAs in humans. Thus, we predict that the diversity of the 197 _F. heteroclitus_ miRNAs that we initially discovered is representative of _F. heteroclitus_ miRNA diversity. We found that 48.7% (96 of 197) of miRNAs appear in 39 clusters (supplementary appendix S4, Supplementary Material online). This is similar to what is observed in zebrafish (Chen et al. 2005); indeed, many of these clusters are conserved among vertebrates and, in some cases, across bilateria (Campo-Paysaa et al. 2011). These patterns of variation are consistent with the influence of strong purifying selection of mostly deleterious mutations resulting in a skewed distribution of segregating variation that is mostly very rare. In humans, miRNA nucleotide diversity is similar to that of coding sequence (Rawlings-Goss et al. 2014), whereas average Tajima’s \( D \) is close to zero for European and East Asian populations but shifted negative in African populations. A similar pattern is reported for nematodes, where nucleotide diversity tends to be low for miRNAs, but Tajima’s \( D \) centers on zero (Jovelin and Cutter 2014). These important regulatory elements appear to be under similar selective constraint in _F. heteroclitus_ compared with other species, but the marked skew toward excess rare variation may be unusual compared with other species. As single miRNAs can have hundreds of binding partners (Selbach et al. 2008), their low nucleotide diversity is consistent with pleiotropy constraining their evolution.

Balancing Selection Governs Variation in Immune System and Olfaction Genes

At the opposite end of the among-individual diversity distribution, we find that the highest diversity genes (coding sequence in genes that occupy the 95th percentile \( \pi \) measurements) tend to be associated with immune system processes (fig. 2). This gene set was significantly enriched for GO or INTERPRO terms including “immunoglobulin and immunoglobulin-like” (\( N = 32, P < 0.0001 \), and \( N = 32, P < 0.0001 \), respectively), “C-type lectin” (\( N = 22, P < 0.0001 \), “MHC protein complex” (\( N = 10, P < 0.0001 \), “cell adhesion” (\( N = 46, P < 0.0001 \), and “positive regulation of immune system process” (\( N = 16, P < 0.0001 \)). Again, as coding sequence variation correlates with promoter sequence variation (fig. 3), these immune system genes are also among the most variable of genes in their promoter sequence. Consistent with this, we find that the gene set with the highest variation in gene expression levels among individuals within populations (measured as coefficient of variation from microarray experiments in _F. heteroclitus_; Whitehead et al. 2012) is significantly enriched for immune system genes (CV 95th percentile, significant GO enrichment). Another enriched GO term within this set included “olfactory receptor” (\( N = 5, P = 0.007 \)).

The high-diversity genes with both high \( \pi \) and positive Tajima’s \( D \) suggest the influence of balancing selection. This is consistent with classic studies in mammals that reveal the influence of diversity-enhancing selection in immunoglobulin genes (Tanaka and Nei 1989) and major histocompatibility complex genes (Hughes and Nei 1988, 1989). Similarly, comparative genomics reveals immune system genes and olfaction genes to be among the fastest evolving between chimpanzees and humans (Mikkelsen et al. 2005). In fish, within-species variation in immunoglobulin genes is particularly high compared with other vertebrates (Das et al. 2008), and the genes with highest sequence diversity among six stickleback fish included immune response genes (Feulner et al. 2013).

Low- and High-Diversity Genes Occupy Different Regions within Interaction Networks

As purifying selection contributes to the fate of “low-diversity” genes and as genes that are positioned at hubs within interaction networks tend to be subject to selective constraint (Fraser et al. 2002), we predicted that the low-diversity genes would more likely be “hub” genes and the high-diversity genes would more likely be connected toward the periphery (tips) of interaction networks. We selected 1,000 UniProt annotated genes with the highest nucleotide diversity (median
\( \pi = 1.29 \times 10^{-2} \) and compared them with 1,000 UniProt annotated genes with the lowest nucleotide diversity (median \( \pi = 7.23 \times 10^{-4} \)). Cellular location and biological function annotations for the two sets showed substantial systematic differences. Genes with the lowest nucleotide diversity were significantly more likely to be located in the cytoplasm or nucleus (Benjamini Hochberg adjusted \( P \) value < 0.01, odds ratio > 1.5), whereas high-diversity genes were more likely to be found at the plasma membrane or in the extracellular space (Benjamini Hochberg adjusted \( P \) value < 0.001, odds ratio < 0.6) (supplementary fig. S7, Supplementary Material online). Furthermore, low-diversity genes were more likely to be transcription regulators, transporters or kinases (Benjamini Hochberg adjusted \( P \) value < 0.02, odds ratio > 1.8), whereas high-diversity genes were more likely to be transmembrane receptors or G-protein coupled receptors (Benjamini Hochberg adjusted \( P \) value < 1.0 \times 10^{-5}, odds ratio < 0.14) (supplementary fig. S7, Supplementary Material online).

To further test our hypotheses about “connectivity” differences between low- and high-diversity genes (not including miRNAs), we used Ingenuity’s (Qiagen) knowledge base to identify downstream targets of each set. As biological networks are scale free, one can summarize their general degree of connectedness as the regression line slope fitted to the log frequency of observing a particular node degree and the log of the node degree observed; that is, scale free networks can be described by a power law (Wagner 2001). We tested for significant differences between the power laws of high- and low-diversity genes using an analysis of covariance of the log node frequency as a function of log node degree (as in Shaw et al. [2014]). The regression was performed in R as an analysis of covariance of log-transformed node degree as a function of log-transformed node frequency to test the null hypothesis that both high- and low-diversity genes do not significantly differ in the power law that governs this relationship. Low-diversity gene sets were significantly more connected than high-diversity gene sets (\( P = 0.008 \), analysis of covariance of their respective power law slopes) (fig. 4). These results are consistent with the proposal that pleiotropy constrains evolution (Fisher 1930), and consistent with a negative relationship between connectivity and evolutionary rate (Fraser et al. 2002).

**Much Genome-Wide Copy Number Variation Is Segregating among Individuals**

Across all 48 *F. heteroclitus* genomes within the Block Island population we detected 24,764 deletions (median length of 656 bp) and 4,069 duplications (median length of 3,576 bp) (table 3). CNV calling methods are sensitive to deletions and simple tandem duplications but have limited ability to detect more complex duplications, which may explain some of the skew in the number of deletions and duplications identified. These copy number variable regions affect 57.3 Mb of the *F. heteroclitus* genome, representing 6.37% of the genome assembly (excluding gaps), which is comparable to the CNV density within three-spined sticklebacks (Chain et al. 2014).

We examined the distribution of CNV lengths in the population and found differing distributions between deletions and duplications (fig. 5A). Deletion events follow a negative exponential distribution, which is similar to the CNV length distribution found in the human genome (Chaisson et al. 2015) where the majority of human CNV events are <1,000 bp in length. These are often difficult to call using Illumina-generated sequence. Due to the insert size used for our Illumina sequencing libraries and our CNV calling methods, our ability to detect CNVs <500 bp was limited. Even

---

**Table 3**

<table>
<thead>
<tr>
<th>Copy Number Variation across 48 Individuals within a Population</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Events</strong></td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>Deletions</td>
</tr>
<tr>
<td>Duplications</td>
</tr>
<tr>
<td>All CNV</td>
</tr>
</tbody>
</table>
with this limitation, 70.8% of *F. heteroclitus* deletion events were <1,000 bp in length. Among humans, 85% of CNVs called with single molecule sequencing were unable to be detected using Illumina sequencing, with most of these undetected CNVs being short length (<1,000 bp) (Chaisson et al. 2015). Therefore, there is most likely an excess of short CNVs in this killifish population that we were unable to detect using Illumina sequencing, in addition to the CNVs reported here. Duplication lengths followed a bimodal distribution, with a peak at <1,000 bp and another peak at ~4,000 bp (fig. 5B).

To assess whether TEs played a role in these duplication events, we scanned duplications with lengths 2.5–5 kb for repeat element sequences. The duplication sequences showed no significant enrichment of repeat elements compared with random sampling from the genome.

We explored the distribution of CNVs among individuals using both the allele frequency spectrum and CNV presence/absence per individual. CNV allelic frequencies were analyzed using depth of coverage in each individual at biallelic sites. CNVs generally occur at low frequency in the population; 52.5% of CNVs have allele frequencies <0.125 (<12/96 alleles) and 45.8% of CNVs are present in <10 individuals (supplementary fig. S8, Supplementary Material online). CNVs occurring in only one individual account for 7% of all CNVs. Though the majority of CNVs occur at low frequency, this *F. heteroclitus* population harbors higher CNV levels compared with sticklebacks; for sticklebacks about half of all CNVs are found in a single individual among 66 individuals sampled from across 11 populations (Chain et al. 2014).

To assess the impact of CNVs on nucleotide diversity within the population, we compared the distribution of nucleotide diversity within regions that harbor deletions, within duplicated regions, and within non-CNV regions in the genome (supplementary fig. S9, Supplementary Material online). We found lower nucleotide diversity within regions that harbor deletions compared with non-CNV regions, due to the loss of heterozygosity caused by deletion events. We found no difference in the distribution

**Fig. 5.**—Distribution of the lengths of deletions (white) and duplications (black) within the Block Island population, binned into 2.5-kb bins with the last bin containing all CNVs greater than 20 kb in length (A). Distribution of the lengths of duplications less than 10 kb in length, showing the abundance of duplications with length of ~4 kb (B).
of nucleotide diversity between duplicated regions and non-CNV regions.

To evaluate the potential functional impacts of CNVs, we calculated the proportion of deletions and duplications that are completely intergenic or intronic, that overlap an exon (but not an entire gene), or that completely encompass a gene (Fig. 6). We found that 94% of deletions are located completely within intergenic or intronic sequence whereas only 66% of duplications are completely intergenic or intronic; 28% of duplications overlap an exon and 6% of duplications completely encompass a gene. Genes affected by the deletion or duplication of exonic sequence (but not the entire gene) make up 9.1% of all genes (2,414 genes). In total, 820 complete genes are copy number variable (3.1% of all genes) in non-CNV regions.

**Fig. 6.** Proportion of deletions (white) and duplications (black) overlapping four mutually exclusive genomic categories (from left to right): entirely within intergenic regions, entirely within intronic regions, partially overlapping a gene including an exonic region, or completely overlapping at least one gene.

acknowledgments

供认

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**Acknowledgments**

Dawoon Jung (Dartmouth College, present address at Seoul National University, Korea) collected the fish used for fosmid sequencing. Dr Pieter de Jong (Children’s Hospital Oakland Research Institute) prepared fosmid libraries. Dr Haixu Tang (Indiana University) contributed to the transposable elements analysis. This work was primarily supported by a grant from the National Science Foundation (collaborative research grants DEB-1265282, DEB-1120512, DEB-1120013, DEB-1120263, DEB-1120333, DEB-1120398 to J.K.C., D.L.C., M.E.H., S.I.K., M.F.O., J.R.S., W.W., and A.W.). Further support was provided by the National Institute of Environmental Health Sciences (1R01ES021934-01 to A.W., P42ES07373 to T.H.H., P42ES007381 to M.E.H., and R01ES019324 to J.R.S.), the National Institute of General Medical Sciences (P20GM103423 and P20GM104318 to B.L.K.), and the National Science Foundation (DBI-0640462 and XSEDE-MCB100147 to D.G.). Extensive comments

**Conclusions**

The evidence-based approach used to identify *F. heteroclitus* gene models, with all loci supported by expression and/or homology (usually both), produced a robust and thorough final gene set. To enable gene function analyses, gene models were linked to human and zebrafish UniProt accesses based on orthology. Genome-wide variation among 48 individuals within a northern *F. heteroclitus* population (which has less variation than southern populations) shows that *F. heteroclitus* ranks among the most genetically diverse of vertebrates (98th percentile) and firmly within the range of arthropod diversity. This high level of standing genetic variation likely contributes to *F. heteroclitus*’ ability to adapt to environmental change. High and low variation is nonrandomly distributed across gene functional categories, and across positions within interaction networks, such that evolutionary constraint imposed by pleiotropy linked with core biological processes can be inferred. Patterns of population genetic variation in epithelial tight junction proteins, and patterns of accelerated evolution of cell cycle regulation and apoptosis genes along the *F. heteroclitus* lineage, offer insight into the mechanisms that may contribute to their physiological plasticity. We detected a genome-wide skew in the relative frequencies of common and rare variation that suggests a post-Pleistocene colonization, which is consistent with the geological history of northwestern Atlantic coastal habitats. This new genome resource, and catalog of variation among many individuals within a population, coupled with our knowledge of their ecology and physiology, provides a foundation for advancing *F. heteroclitus* as a key model system for understanding gene-by-environment interactions in an evolutionary context.
contributed by two anonymous reviewers served to significantly improve the manuscript. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the US EPA.

**Literature Cited**


