SUPPLEMENTAL METHODS

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, Neel Aluru, Sibel I. Karchner, John K. Colbourne, Mark E. Hahn, Joseph R. Shaw, Marjorie F. Oleksiak, Douglas L. Crawford, Wesley C. Warren, Andrew Whitehead (2017).

The Landscape of Extreme Genomic Variation in the Highly Adaptable Atlantic Killifish. Genome Biol Evol 2017; 9 (3): 659-676. doi: 10.1093/gbe/evx023

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. We assessed model quality to select gene locus representatives that captured evidence in the most complete way, and to compare gene sets. 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.

Gene expression evidence included mRNA sequences collected from five sources: 1) expressed sequence tags from various *F. heteroclitus* populations, tissues, and life stages catalogued in NCBI (http://www.ncbi.nlm.nih.gov/nucest?term=Fundulus%20heteroclitus[Organism]); 2) Roche 454 RNA-seq reads from *F. heteroclitus* whole embryos (10 days post fertilization) derived from experiments described in Oleksiak et al. (Oleksiak, et al. 2011) where parental fish were from New Bedford Harbor, MA (41°34.0' N, 70°54.9' W) and Scorton Creek, Sandwich, MA (41°44.0' N, 70°23.0' W) (NCBI/GEO accession GSE86593); 3) Roche 454 RNA-seq reads from multiple developmental stages (embryo and larvae from days 1-15 of development) from parents collected from New Bedford Harbor and Scorton Creek (Oleksiak, et al. 2011) (NCBI/GEO accession GSE86593); 4) RNA-seq reads from adult *F. heteroclitus* gills captured from Northeast Creek (Mount Desert Island, ME) as reported in (Shaw, et al. 2014) (NCBI/GEO accession GSE47035); 5) Illumina HiSeq 2000 strand-specific reads from re-sequencing of the previous three sources of RNA (NCBI/SRA accessions SRR1692373 and SRR1692374). RNAs from these sources were assembled into long, and often complete, mRNA transcripts primarily using de-novo assembly methods. The collection of assembled transcripts was classified to non-redundant coding loci and alternate transcripts with EvidentialGene (Gilbert 2012). These primary and alternate mRNA assemblies are published in NCBI transcript sequence archive TSA under accession GCES00000000.

For gene predictions, these expressed sequences (reads and assembled transcripts) were mapped to genome assembly with GSNAP (Wu and Nacu 2010) for short reads and GMAP (Wu and Watanabe 2005) for transcripts. Intron evidence was collected from properly spliced reads and transcripts mapped onto the genome assembly; the number of spliced reads per intron location was used as a quality

score. Reads and transcripts provide exon evidence, and transcripts provide full gene model span evidence, used during genome-gene modelling.

Gene homology evidence was collected from proteomes of from nine fish species: spotted gar (*Lepisosteus oculatus*), zebrafish (*Danio rerio*), catfish (*Ictalurus punctatus*), stickleback (*Gasterosteus aculeatus*), puffer (*Tetraodon nigroviridis*), tilapia (*Oreochromis niloticus*), zebra mbuna (*Maylandia zebra*), Japanese medaka (*Oryzias latipes*), and platyfish (*Xiphophorus maculatus*), and the human proteome from UniProt 2013 (UniRef 50 set of 39,357 proteins excluding fragments). NCBI Eukaryote genome annotation of 2013 was used for *Maylandia*, spotted gar pre-release of ENSEMBL 2012, ENSEMBL release 70 of platyfish, EvidentialGene 2013 gene construction for catfish, and ENSEMBL release 67 for the remainder. These 338,000 proteins were used for genome-based gene modeling and evidence assessment. Following gene construction and evaluation to the final *Fundulus* gene set, it was re-evaluated for orthology with a later ENSEMBL release 74 set of fish proteomes, including the full initial spotted gar gene set.

For gene predictions, these fish proteins were aligned to genome assembly and used to guide gene predictions, along with other gene evidence. For expressed gene assembly, these proteins were not used in gene construction. Alignment to genome assembly was done using tBLASTn (NCBI) to the repeat and transposon soft-masked genome, then refined with Exonerate (Slater and Birney 2005) to create protein gene models, with options "exonerate --model protein2genome:bestfit --exhaustive 1 --subopt 0 --forcegtag 1 --softmasktarget 1". These proteins were also used in orthology assessment described below.

We assembled paired-end RNA-seq to genes using the EvidentialGene methods from the euGenes.org, with several *de-novo* RNA assemblers, using multiple options for kmer fragmenting, insert sizes, read coverage, digital normalization, and quality and abundance filtering. RNA assemblers used include Velvet/Oases (v1.2.03/o0.2.06), SOAPDenovo-Trans (v2011.12.22) using multi-kmer shredding options from 23bp to 95bp, and Trinity (v2012.03.17) (with fixed kmer option). Accessory methods used for RNA data processing include GMAP/GSNAP and Bowtie for read and transcript mapping to genome assembly, diginorm of khmer package, and sequence artifact filtering. Additional transcripts were assembled with genome-mapping assistance, using methods of PASA (v2.2011), Cufflinks (v1.0.3 and v0.8), and EvidentialGene. We produced 45 alternate gene assemblies with varying methods, parameters and source data (approx. 200,000 transcripts each). EvidentialGene *tr2aacds* pipeline (http://eugenes.org/EvidentialGene/trassembly.html) was used to process and classify these several million transcript assemblies into best representatives of primary and alternate locus for retention, plus redundant duplicates, fragments and mis-assemblies that are discarded.

We constructed gene models upon the draft genome assembly. Transposon and repeat locations were located with RepeatMasker (Smit, et al. 2013-2015) and Repbase database. Gene models are derived using the evidence-directed AUGUSTUS predictor (Stanke, et al. 2006). Several gene prediction sets are produced to create a superset of models that include the models selected to be best, based on matching all gene evidence using EvidentialGene methods (Gilbert 2012).

AUGUSTUS flexibly uses both Hidden Markov Model (HMM) training models and available gene evidence for each locus. Training the predictor HMM involves steps described by the authors (Stanke, et al. 2006), with validated genes for this species. We produced 12 alternate gene model set with AUGUSTUS, varying evidence and parameters, of 30,000 to 70,000 loci per set.

We selected 2,000 *Fundulus* genes that appeared to be full length from EST/RNA transcript assemblies. We split these into subsets for training and validation of the resulting predictor. We created and used several training sets, plus one that is un-optimized. Evidence sets and configuration weightings were constructed to include: *(1)* complete gene structure information (exon, CDS, intron, gene spans); and *(2)* an extra influence of one major component (proteins, EST exons, full transcript assemblies). The first was necessary to reduce aberrant gene models generated by over-influence of one structure component. For example, evidence of exons from only ESTs lead to missed introns and missed gene ends. The second was required to reduce conflicting signals, and returned better models under the influence of an appropriate gene evidence class. For instance, extra influence of homologous proteins returned models that more closely matched those proteins. Following each prediction run, the results were assessed for overall quality and matched to evidence. This assessment then suggested the options for new configurations and evidence mixtures. AUGUSTUS is also able to model alternate transcripts from evidence. But those are seldom supported by transcript assemblies and tend to include aberrations. Therefore, we did not use this option and instead used only transcripts assembled directly from RNA reads in selecting alternate splice-forms.

Gene sets from RNA assembly and gene prediction were classified by evidence and best locus representatives were selected. EvidentialGene (Gilbert 2012) uses gene evidence described above, annotates each model and exon, calculates quality scores per model for each type of evidence, and then determines a combined locus set from maximal evidence scores. A final set of distinct loci, plus alternate transcripts, was chosen to be good quality genes, all having the best match to expression and/or protein homology evidence. These were annotated and compared to other fish gene sets. Genes were annotated with descriptions and attributes drawn from orthology to UniProt and ENSEMBL fish and human gene data, and classified by evidence scores, including transposable elements. Models left out of our final set included redundant models, most with homology to transposon proteins, and those with no or minor expression and protein evidence. However, transposon protein genes that were annotated as being expressed and/or having orthology to other fish species were retained. An annotated, searchable gene and genome database for this *F. heteroclitus* project is available at http://eugenes.org/EvidentialGene/vertebrates/killifish/. A table of attributes for the final gene set, with associated evidence scores, is available in Appendix S1. A GFF file of our gene set that is compatible with the NCBI-hosted genome is provided in Appendix S5.

In the EvidentialGene locus classifier, locus overlaps of gene models are calculated using the primary criteria of CDS-overlap on same DNA strand (reverse-strand CDS-overlap is rare, but locus UTR overlaps are relatively common). A

weighted sum of the various evidence component scores is calculated, configurable to gene set requirements. Selecting the best locus from among a large set of gene models is accomplished according to two basic criteria: (1) gene evidence must pass a minimum threshold score, and (2) the combined score is maximal for all models overlapping the same CDS-locations. Other criteria and tests are included and used for classification, such as orthology scores. One indicator of a joined model error is a homology score for the joined model that is no greater than for un-joined models, though its coding span is larger. Determining a final gene set is an iterative process that involves evaluation after selection, modification of score weights, and reselection. After the majority of optimal models are found, smaller subsets of problem loci are sampled and examined, with additional evaluations to resolve these. This is a negative-feedback process designed to filter out errors and suboptimal gene models, with successive iterations changing fewer models until the optimal set is found. It also involves expert curation to identify and remove suboptimal models, and locate or promote missed high value models (e.g., unique orthologs).

Quality scores per gene model are calculated using the following types of evidence: *(a)* the level of RNA sequence coverage over the gene model coordinates on the genome assembly; *(b)* the number of EST and RNA sequence reads spanning the intron splice sites that matched to annotated exon ends; *(c)* gene structure agreement, as end-to-end match of exons in the model with the evidence in support of gene structure; *(d)* sequence homology to proteins from species-specific reference databases using BLASTp scores of all significant matches to the reference set of genes. Each gene model for each locus is therefore scored by weighted evidence. Finally, the maximal evidence scored, non-overlapping model set is determined, with respect to inter-locus effects of gene joins and other factors.

Gene names have been assigned on the basis of homology (BLASTp) scores to UniProt UniRef50 proteins, reference fish genes, and the consensus gene family names from OrthoMCL orthology analyses, in accordance with UniProt protein naming guidelines.

Orthology of intermediate and final *F. heteroclitus* gene sets was assessed with reciprocal blastp and OrthoMCL gene family clustering (Li, et al. 2003). For OrthoMCL analysis, significance criteria were applied with recommended options: a similarity P-value <= 1e-05, protein percent identity >= 40%, and MCL inflation of 1.5 to achieve optimal clustering granularity. Reciprocal best similarity pairs between species and reciprocal better similarity pairs within species (i.e., recently arisen paralogs, or in-paralogs, proteins that are more similar to each other within one species than to any protein in the other species) were added to a similarity matrix. The protein similarity matrix was normalized by species and subjected to Markov clustering (Enright, et al. 2002) to generate ortholog groups including recent in-paralogs. This clustering provides gene orthology classes of ortholog and out-paralog, abbreviated “orlog”, in-paralog “inpar” genes with greater within-species alignment but clustering with other species, and non-ortholog classes abbreviated as single-copy “noor” and paralog grouped “upar” (Table S4, and Appendix S1. Note that “noor” and “upar” categories are summarized in Table S4 as “unclassified-single” and “unclassified-dup”, respectively). These non-ortholog

classes generally contain genes with significant homology, are not *de-novo* new species genes, but homology is not sufficient for clustering to stricter ortholog groups. Orthologs include those with best alignment to another species gene, but may include groups with two or more genes per species can also be termed out-paralogs. We use ortholog for both grouping types, due to the computational rather than phylogenetic groupings used as they both measure genes with one-to-one best alignment to other species.

The existence of *Fundulus heteroclitus* genes in other fish genomes (Amazon molly *Poecilia formosa*, guppy *Poecilia reticulata*, and African turquoise killifish *Nothobranchius furzeri,* zebrafish *Danio rerio*) was examined, using blastn alignment of *Fundulus heteroclitus* transcripts to genome assemblies of those species. This determines whether unlocated genes may exist in these related fish. Alignment to genome assemblies is done as "blastn -task blastn -evalue 1e-5 -db otherfish-chrasm -query killifish2.mrna", then output is reduced to table of gene coding exon locations on chrasm with evigene script makeblastscore3.pl. Alignment of gene coding sequences is done as "blastn -task dc-megablast -template_type coding -template_length 18 -evalue 1e-9 -ungapped -db killifish2-cds -query otherfish.cds".

REFERENCES

Enright AJ, Van Dongen S, Ouzounis CA 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30: 1575-1584.

EvidentialGene: Perfect Genes Constructed from Gigabases of RNA [Internet]. 2012. Available from: http://arthropods.eugenes.org/EvidentialGene/

Li L, Stoeckert CJ, Roos DS 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178-2189.

Oleksiak MF, et al. 2011. Transcriptomic assessment of resistance to effects of an aryl hydrocarbon receptor (AHR) agonist in embryos of Atlantic Killifish (*Fundulus heteroclitus*) from a Marine Superfund Site. BMC Genomics 12: 263.

Shaw JR, et al. 2014. Natural Selection Canalizes Expression Variation of Environmentally Induced Plasticity-Enabling Genes. Molecular Biology and Evolution 31: 3002-3015.

Slater GS, Birney E 2005. Automated generation of heuristics for biological sequence comparison. Bmc Bioinformatics 6: 31.

Smit A, Hubley R, Green P 2013-2015. RepeatMasker Open-4.0.

Stanke M, Schoffmann O, Morgenstern B, Waack S 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. Bmc Bioinformatics 7.

Wu TD, Nacu S 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873-881.

Wu TD, Watanabe CK 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859-1875.