EPA Public Access

Author manuscript

Environ Toxicol Chem. Author manuscript; available in PMC 2018 December 01.

About author manuscripts

Submit a manuscript

Published in final edited form as:

Environ Toxicol Chem. 2017 December ; 36(12): 3436–3442. doi:10.1002/etc.3929.

FIRST-GENERATION ANNOTATIONS FOR THE FATHEAD MINNOW (*PIMEPHALES PROMELAS*) GENOME

Travis W. Saari^a, Anthony L. Schroeder^b, Gerald T. Ankley^a, and Daniel L. Villeneuve^{a,*}

^aOffice of Research and Development, National Health and Environmental Effects Research Laboratory, Mid-Continent Ecology Division, US Environmental Protection Agency, Duluth, Minnesota

^bMath, Science, and Technology Department, University of Minnesota Crookston, Crookston, Minnesota, USA

Abstract

Ab initio gene prediction and evidence alignment were used to produce the first annotations for the fathead minnow (*Pimephales promelas*) genome. We also describe a genome browser, the Society of Environmental Toxicology and Chemistry's Fathead minnow genome project, that provides simplified access to the annotation data in context with the genomic sequence. The present study extends the utility of the fathead minnow genome and supports the continued development of this species as a model organism for predictive toxicology.

Keywords

Genome; Annotation; Fathead; Pimephales; Browser

INTRODUCTION

The fathead minnow (*Pimephales promelas*) has served as an important model organism in aquatic ecotoxicology research for decades, and is the most commonly used species in aquatic ecotoxicology regulatory testing in North America [1]. The sequencing and assembly of the fathead minnow genome, reported by Burns et al. [2], was an important step in advancing molecular characterization of the fathead minnow and facilitating its ongoing use in 21st century predictive toxicology. However, the assemblies published by Burns et al. [2] were unannotated, with no direct indication of which regions of the genome contain functional components or what those functional components were. Without annotations, much of the utility of the genomic resources available for this species has yet to be realized.

* Address correspondence to villeneuve.dan@epa.gov.

Supplemental Data—The Supplemental Data are available on the Wiley Online Library at DOI: 10.1002/etc.3929.

Disclaimer—This manuscript has been reviewed in accordance with the requirements of the US Environmental Protection Agency (EPA) Office of Research and Development and Office of Pesticide Programs. The views expressed in the present study are those of the authors and do not necessarily reflect the views or policies of the USEPA, nor does the mention of trade names or commercial products constitute endorsement or recommendation for use.

Data availability—The fathead minnow genome browser and all associated annotation data files can be accessed at setac.org/fhm-genome.

In the present study we describe the use of an automated annotation pipeline to generate the first genomic annotations for the fathead minnow. By utilizing ab initio gene prediction and evidence alignment (Figure 1), we have created the first set of putative gene models for this important model species. In addition, we have created a web-accessible genome browser that provides simplified access to these genomic resources. The present study will aid in predictive toxicology work supporting both research and regulation (including facilitation of the use of next-generation sequencing data) in the fathead minnow.

MATERIALS AND METHODS

State of the genome assembly

The sequencing effort described by Burns et al. [2] produced 2 genome assemblies for the fathead minnow, which are now publicly available through the National Center for Biotechnology Information (NCBI) GenBank database [3]. High-throughput Illumina paired-end sequencing was used with deoxyribonucleic acid (DNA) library preparations of varying insert sizes to achieve over 100× coverage of the genome, and these sequence reads were assembled using 2 different methods: String Graph Assembly [4] and SOAPdenovo [5] assembly. The authors noted that the assembly generated using the SOAPdenovo method had higher N50 values than the assembly generated with the String Graph Assembly method, indicative of a more contiguous genome assembly [4,5]. Further evaluation of the assemblies using a Core Eukaryotic Gene Mapping Approach (CEGMA) revealed that a higher number of genes could be mapped to the SOAPdenovo assembly [5]. Taken together, these results suggest that the SOAPdenovo assembly is more contiguous and amenable to gene annotation; therefore, this assembly was chosen for the present study.

Ab initio gene prediction

Ab initio gene prediction is often used as the first step in generating gene annotations, because it can be performed with only the genome sequence as an input. The most effective programs for ab initio gene prediction use Hidden Markov Models to determine the most likely sequence components in a given DNA sequence (e.g., exons, introns, intergenic regions, etc.) [6]. The basis for this determination is made using the probabilities of state transitions (e.g., intergenic to genic regions, splice sites) as well as the probabilities of observing a particular nucleotide composition in one of the available states (untranslated region, exon, intron, etc.). This method allows for complete gene models to be created without any extrinsic data for a particular organism, using only a model that has the defined aforementioned probabilities for a typical DNA sequence.

To perform ab initio gene prediction in the fathead minnow, the program AUGUSTUS [7], Ver 3.03, was used (Figure 1). This program can be fine-tuned to specific genomes through training, that is, providing a set of known gene structures that the program uses to create a model with maximized accuracy for a particular species. In addition, AUGUSTUS includes several models that have already been optimized for various eukaryotic organisms.

Before beginning ab initio gene prediction, a model was custom-trained for the fathead minnow using the limited amount of known gene sequences in the species, and this model

was tested for accuracy and compared with models that were provided with AUGUSTUS. To create the fathead minnow-specific model, all available fathead minnow protein sequences (423 total) were downloaded from NCBI. After mitochondrial proteins and partial sequences were removed, just 49 full-length fathead minnow proteins remained. An additional 55 multiexonic partial protein sequences were added, for a total of 104 fathead minnow protein sequences. These 104 sequences were randomly split into 2 different groups—one for training and one for testing AUGUSTUS. Subsequently, the training set was used as described in the AUGUSTUS training tutorial [8]. This custom-trained model was evaluated using the built-in accuracy testing capabilities of AUGUSTUS, which employ the set of known genes that were reserved for testing in the previous steps. In addition, 2 other predefined models included in the AUGUSTUS package were tested for accuracy in the same fashion: a model trained for a generic eukaryotic genome, and a model trained for the zebrafish genome, for a total of 3 tested models.

When compared, the model that was custom-trained based on 104 fathead minnow protein sequences performed the worst (Table 1), with sensitivity and specificity values at the exon level of 0.28 and 0.395, respectively. This is likely because of the relatively small number of known fathead minnow sequences that were available for training the model. It is recommended to use over 1000 genes for training [8], but this was not possible using known fathead minnow sequences. The zebrafish model had the highest exon level sensitivity and specificity, with values of 0.656 and 0.727, respectively (Table 1). Based on results of the accuracy test, the zebrafish model was used with the genome assembly to perform ab initio gene prediction.

The original fathead minnow genome assemblies, after their inclusion in the GenBank database, were automatically repeat-masked by NCBI using RepeatMasker [9]. This important step prevents repetitive and low-complexity regions from interfering with the gene prediction process. The repeat-masked sequence of the SOAPdenovo assembly was supplied to AUGUSTUS as the template sequence, and the softmasking option was enabled. This option allowed gene predictions to continue into repeat-masked regions of the genome assembly while still preventing them from being initiated in these regions. In configuration of the output, the gff3 option of AUGUSTUS was enabled to produce a result in a standard and portable format, and stop codons were included in the coding DNA sequence in accordance with the Sequence Ontology definition of the term [10].

The output from AUGUSTUS provided translated protein sequences for each gene prediction. To help evaluate the potential identity of these gene predictions, the protein sequences were queried against both the NCBI nonredundant protein database and the zebrafish proteome provided by ENSEMBL [11] using the BLAST+ software package [12], Ver 2.2.28. The top hit from these blast results was added to the annotation file containing the ab initio gene predictions.

Evidence alignment

In contrast to ab initio gene prediction, which only requires a genomic DNA sequence as input, evidence alignment uses various other types of sequence data to add information to the genome assembly (Figure 1). When sequences (e.g., transcripts) are aligned, these

alignments can serve as potential links between the known query sequence and the genomic assembly, and various types of evidence can then be used to build confidence in a particular gene model and annotation. For example, a spliced alignment of transcript data can be used to more precisely define the coordinates of the exons in a putative gene model. In addition, the alignment of homologous transcripts from other well-characterized organisms can be used to assign putative gene identities to a particular genomic region.

The BLAST+ and Exonerate Ver 2.4.0 [13] programs were used in concert to generate spliced alignments (Figure 2). After completion, poor alignments were removed based on the percentage of alignment length (details specific for each type of alignment), and the best remaining alignment for each query sequence was chosen based on Exonerate's alignment scores. Using this approach, it was often necessary to employ a simple program, written in an interpreted language, to move from one step to the next. The programs and configuration files that were used (e.g., for coordinating the BLAST and Exonerate alignment programs, converting the output into gff3 format, filtering alignments) are publicly available at GitHub [14].

A number of fathead minnow expressed sequence tags were available for download from the NCBI expressed sequence tag database [15]. These sequenced fragments of fathead minnow transcripts were generated as part of previous experiments. When aligned to the genome, these transcripts correspond to expressed genomic regions, and these alignments can be used to precisely define the coordinates of a gene model. In all, 258 504 fathead minnow expressed sequence tags were downloaded from NCBI, split into individual sequence files, and run through BLASTN/Exonerate to generate spliced alignments. The blastn task supplied by the BLASTN application was used to query expressed sequence tags against the fathead minnow SOAPdenovo genome assembly, and the 6 highest scoring subject scaffolds were passed on to Exonerate (Figure 2). Exonerate was employed using the est2genome model, the maximum intron size was limited to 20 000 base pairs, and the minimum intron size was set to 10 base pairs. In its output, Exonerate was configured to produce a Generic Feature Format (GFF) output with respect to the target sequence. After alignment, the output of Exonerate was converted into the newer, extensible GFF3 format [16], and additional alignment characteristics from Exonerate's output (e.g., alignment length) that were not included in the original Generic Feature Format lines were added to the processed GFF3 lines. The alignments that had aligned less than 90% of the query sequence length were removed from this GFF3 file, and then the best remaining entry for each query sequence was kept.

Similarly, zebrafish coding sequences were downloaded from the ENSEMBL database [11], split into individual sequence files, and run through BLASTN and Exonerate to generate spliced alignments. The discontinuous megablast task in the BLASTN application was used to detect alignments that may have larger stretches of mismatches, insertions, or deletions that are commonly seen in interspecies nucleotide alignments. Once again, the 6 highest scoring scaffolds were processed using Exonerate. In this instance, Exonerate was run using the coding2genome model, but was otherwise configured as described previously for fathead minnow. The output was converted and processed similarly to the output from the expressed sequence tag alignments, although the minimum alignment length percentage threshold was

lowered to 50%. After limiting the alignments to only the best scoring for each query sequence, the GFF3 annotation entries were extended with the gene symbols and descriptions were derived from the query sequences' identities. This way, in addition to precisely defining the exonic regions of gene models, the aligned coding DNA sequence might also serve as a link between characterized genes in the zebrafish and fathead minnow genomic regions.

In addition, an RNA-seq dataset representative of the fathead minnow transcriptome under standard laboratory rearing conditions (R. Packer, The George Washington University, Washington, DC, USA, unpublished manuscript) was obtained from NCBI Sequence Read Archive database, (accession no. SRR1582202) [17]. This dataset was produced using Illumina sequencing technology with read lengths of 202 bp, which generated over 470 million paired-end reads. These transcript sequences were recognized as a useful source of additional evidence for the gene annotation, so they were sorted and then aligned to the fathead minnow genome using the HISAT2 package [18]. The maximum intron size was constrained to 20 000 base pairs, and default settings were used otherwise. The estimated insert size of the paired end reads was 184 ± 68 bp. Insert size was not used to filter the mappings.

Development of a genome browser

The annotation files produced by the previous steps contain an incredible amount of information, the large machine-readable files are difficult for researchers to navigate and to extract information from. To help remedy this, the web-based genome browser JBrowse [19] was employed (Figure 1). The genome browser provides a visual and interactive method to explore the genome assembly along with its associated annotation data. In addition, different data sources can be viewed side by side. When utilizing annotation data from an automated pipeline, this ability is crucial; there are often complementary data points, derived from multiple sources, that can help to fill in information gaps and create a holistic picture of a putative gene-coding region.

There are several programs that facilitate loading various data types into the application included in JBrowse, and these were used to add the reference sequence and associated annotation files to the genome browser. The fathead minnow SOAPdenovo assembly was used as the reference sequence, and annotation files (GFF3 format) resulting from the ab initio gene prediction, zebrafish coding DNA sequence alignment, and fathead minnow expressed sequence tag alignment are related to this sequence. The gene symbols and accessions from the zebrafish homolog alignments were added as alternative names for that annotation file, and these were used to generate a searchable index for the genome browser. This way, users can navigate to a particular genomic region using a well-known identifier (e.g., gene symbol or accession number), rather than a genomic coordinate, which may not be known a priori. The RNA-sequencing alignment results were converted from a per-alignment format to a depth-based format before loading into the genome browser. Because the RNA-seq reads were derived from a single experimental treatment, distinction between individual reads carries little benefit. However, their aggregated result (e.g., depth of genomic coverage) is informative in determining expressed regions of the genome, so

converting the data in this way retains the most relevant information while drastically reducing file size and server requirements. A combination of SamTools [20], BedTools [21], and the utility bedGraphToBigWig [22] were utilized to convert the Sequence Alignment/Map format [23] into the BigWig format [22]. The resulting BigWig file was loaded into the genome browser application and configured to display coverage density of RNA-seq alignments on the fathead minnow genome assembly.

RESULTS AND DISCUSSION

The ab initio analysis produced a total of 43 345 gene predictions on the fathead minnow SOAPdenovo assembly. These predictions indicate regions of the genome that potentially encode genes, with a single gene prediction assigned to a given genomic region. The current 43 345 gene predictions are likely a significant overestimate of the number of true coding genes based on comparison with other fish genomes, which average approximately 21 000 genes (Supplemental Data, Table S1). The overestimate likely indicates representation of numerous gene fragments resulting from incomplete assembly of the short-read sequences. This is also suggested by the lower median exon count (3) for the fathead minnow gene predictions compared with those of other species (e.g., 9 for *Danio rerio*; 8 for *Oryzias latipes*; 14 for *Takifugu rubripes*; Supplemental Data, Table S2).

The evidence alignments provide potential links between a particular query sequence and the genomic regions they align to, and so multiple pieces of evidence can align to the same place. For each type of evidence, the successfully aligned queries represent these links between the evidence and the genome (Table 2). The majority of the alignments in each dataset were successful, with 82% of zebrafish coding DNA sequence, 82% of fathead minnow expressed sequence tags, and 77% of RNA-seq reads aligning to a single genomic scaffold with scores above threshold cutoffs (Table 2). In addition, the coverage of the genome, as roughly indicated by the number of scaffolds that the various lines of sequence evidence aligned to, was greater for the RNA-seq reads than for the other alignments, as one might expect. The number of alignments attempted is multiple orders of magnitude higher in the RNA-seq dataset than in the other datasets, and so these are more likely to align to a given genomic region through random chance.

The extent to which the different data sources (e.g., gene predictions, expressed sequence tag alignments) overlap with one another can be useful in determining the accuracy and completeness of the gene models (Table 3). Ideally, a fathead minnow genomic region would have a gene prediction, a coding DNA sequence alignment, and several transcript alignments from expressed sequence tag or RNA-seq data that all corroborate one another; in this case, we would have high confidence that a given region encodes a gene. However, because of the limited nature of the datasets and their inherent differences in coverage, it is inevitable that some genomic regions may have sparse or conflicting information, which makes it more difficult to identify a particular gene in that region.

To generate an overview of the extent of overlap between the datasets, each gene prediction, coding DNA sequence alignment, expressed sequence tag alignment, or RNA-seq alignment was taken as a genomic interval based on the coordinates of the fathead minnow genome

assembly where it exists. Evaluated this way, the datasets were tested for intersection with at least 100 base pairs of overlap (Table 3). Notably, the majority of the zebrafish coding DNA sequence sequences overlapped with at least one of each data type, and nearly all of them (92%) overlapped with a gene prediction (Table 3). This indicates that the zebrafish coding DNA sequence alignments were of high confidence, because they were nearly universally corroborated by other information sources. Conversely, the majority of ab initio gene predictions were supported by RNA-seq alignments, with only a smaller fraction being supported by other data sources (Table 3). This is expected because the ab initio gene predictions are only characterized based on their genomic sequence, so these predicted genes may or may not be expressed in the organism under any given set of conditions.

In addition to evaluating the global overlap among the various data sources, we manually evaluated the quality of the annotations for the 104 genes for which either partial or full fathead minnow protein sequences were available in NCBI GenBank or Ensembl. Among the 104 targets evaluated, 92 (88%) of the corresponding ab initio gene predictions aligned correctly with evidence tracks that supported the annotation provided in the original source. Incorrect annotations, indicated by mismatch between the ab initio annotation and other available evidence or the original source were detected for 8 of the 104 sequences ($\approx 8\%$), whereas the remainder either lacked a corresponding ab initio gene prediction or the alignment was inconclusive based on the evidence available.

The genome browser

The genome browser allows users to navigate the fathead minnow SOAPdenovo genome assembly and view various types of data (Figure 3). It is possible to move between the genomic scaffolds in the assembly by typing their accession numbers into the text search. It is also possible to select various types of annotation data to view, and to view annotation files that are locally stored on a user's computer. The browser enables one to easily find detailed information about a particular annotation, to pull underlying sequence from the genome assembly, and to conduct in silico splicing using the exons and introns derived from the annotations.

Utility and applications

The first-generation annotations developed as part of this effort should immediately expedite a wide range of molecular analyses and applications with this species. For example, our research team has already used it to aid design and development of polymerase chain reaction (PCR) primers (e.g., Ankley et al. 2017 [24]). Quantitative (q)PCR is routinely used to quantify messenger ribonucleic acid (mRNA) expression of genes of interest, but primers for the fathead minnow historically have been designed using limited information from related species, application of degenerate primers, or from individual expressed sequence tags. With the genome browser, it is now possible to view putative gene structures and conduct in silico splicing, which can be used to approximate the intron–exon boundaries of genes. The use of qPCR primers that span intron–exon boundaries within an mRNA molecule (generally not feasible when designing is based on expressed sequence tags) is a simple way to reduce the possibility of genomic DNA contamination in qPCRs.

Likewise, development of these fathead minnow genome annotations and genome browser supports the expanded use of RNA sequencing in the fathead minnow as a tool for understanding toxicity. Although it has been common practice for several years to observe specific transcript levels in response to chemical exposure via qPCR or microarrays, recently there has been an increased interest in whole-transcriptome sequencing. Van Delft et al. demonstrated that RNA-sequencing can be used to detect differentially expressed genes in human cells exposed to benzo[*a*]pyrene [25], and others are beginning to lay the foundations for the use of RNA-seq in *Daphnia* [26] and in fish species [27] for the purpose of ecotoxicology research. The ability to map RNA-seq reads to a reference genome gives a more complete and unbiased view into the transcriptome, and allows researchers to identify novel transcripts, even while using short read lengths such as those offered by Illumina technologies [28]. Furthermore, proteomic analysis can benefit from cross-validation with genomic and transcriptomic data [29]. The characterization of all the genes in the organism also allows for other types of whole-genome analyses, including the detection of different isoforms of genes, or comparison of genes between the fathead minnow and other species. One type of whole-genome analysis is already underway—the mapping of quantitative trait loci in the fathead minnow (E. Waits, US EPA, Cincinnati, OH, USA, personal communication), which will allow for genetic regions to be associated with particular phenotypes.

Next steps

As with most genome projects, genome assembly and annotation is an ongoing process; in this context, what we have done so far is incomplete and subject to error. Although Burns et al. [2] were able to achieve an adequate sequencing depth, they point out that the short read lengths provided by Illumina sequencing along with the highly repetitive nature of fish genomes ultimately limit the contiguity of the assemblies they were able to generate. Similar to their CEGMA results, we found several genes that could not be mapped to the current genome assembly, and many that were partially aligned to one or more scaffolds. This finding suggests that if another, more contiguous genome assembly were created, through integration of additional long-read sequencing data and/or improved scaffolding methods, additional iterations of automated annotation should produce a more complete gene set. The current assemblies for the fathead minnow still have plenty of room for improvement—the scaffold and contig N50 values for the relatively well-characterized zebrafish and medaka genomes are orders of magnitude higher than those in the current fathead minnow genome assemblies [30,31]. Consequently, our results should appropriately be viewed as a first release annotation that, while very useful, invariably will continue to improve in quality with additional work.

Despite their recognized limitations, it is nonetheless important and valuable to make these resources available to the scientific community. The pace at which omics data are being generated and the associated methods and bioinformatics tools are evolving nearly guarantee that genome annotations will continue to evolve and improve over time. The results described in the present study provide a starting point for illuminating the genes that exist in the current fathead minnow genome assemblies, and it is assumed that manual curation, along with continued use of high-throughput sequencing, will improve our understanding of

this organism into the future. In light of this, the tools that were used for this annotation process have all been made publicly available so that other researchers can use them to perform similar work, but with new datasets. In addition, the genome browser available at the Society of Environmental Toxicology and Chemistry's website [32] allows new users to view their own annotation files alongside those that have been generated as part of the present study. We encourage others to utilize these resources and to contribute to the fathead minnow genome annotation effort if they are able. Ultimately, improvements to the current genome and its annotation can be large or small, including the use of newer scaffolding technologies to improve the assembly itself, incorporation of additional tissue-specific transcript libraries, or even simple manual corrections of the automated gene predictions. By making these improvements accessible through the browser, the community can continue to benefit from the incremental contributions of many investigators as the genomic resources for the fathead minnow evolve.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Olker for providing comments.

References

1. Ankley GT, Villeneuve DL. The fathead minnow in aquatic toxicology: Past, present, and future. *Aquat Toxicol.* 2006; 78:91–102. [PubMed: 16494955]
2. Burns FR, Cogburn AL, Ankley GT, Villeneuve DL, Waits E, Chang YJ, Llaca V, Deschamps SD, Jackson RE, Hoke RA. Sequencing and de novo draft assemblies of a fathead minnow (*Pimephales promelas*) reference genome. *Environ Toxicol Chem.* 2015; 35:212–217. [PubMed: 26513338]
3. National Center for Biotechnology Information. [cited 2017 February 9] GenBank Genomes—*Pimephales promelas*. 2017. Available from: <https://www.ncbi.nlm.nih.gov/genome/genomes/13167>
4. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 2011; 22:549–556. [PubMed: 22156294]
5. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaogian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience.* 2012; 1:1–18. [PubMed: 23587310]
6. Wang Z, Chen Y, Li Y. A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics.* 2004; 2:216–221. [PubMed: 15901250]
7. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics.* 2008; 24:637–644. [PubMed: 18218656]
8. Stanke, M. [cited 2017 February 9] Training AUGUSTUS. 2011. Available from: <http://molecularrevolution.org/molevolfiles/exercises/augustus/training.html>
9. Smit, AFA., Hubley, R., Green, P. [cited 2017 February 9] RepeatMasker Open-4.0. 2013–2015. Available from: <http://www.repeatmasker.org>
10. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biol.* 2005; 6:R44. [PubMed: 15892872]
11. Yates A, et al. Ensembl 2016. *Nucleic Acids Res.* 2016; 44:D710–D716. [PubMed: 26687719]

12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009; 10:421. [PubMed: 20003500]
13. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2006; 6:31.
14. Saari, T. [cited 2017 March 6] Tsaari88 config files & R script. GitHub. 2017. Available from: <https://github.com/tsaari88/annotation-scripts>
15. National Center for Biotechnology Information. [cited 2016 August 3] Genbank EST—*Pimephales promelas*. 2017. Available from: <https://www.ncbi.nlm.nih.gov/nucest/?term=pimephales+promelas+%5BOrganism%5D>
16. Stein, L. [cited 2017 February 9] Generic Feature Format Version 3. 2013. Available from: <https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md>
17. National Center for Biotechnology Information. [cited 2017 February 9] Fathead Minnow RNA-seq. 2017. Available from: <https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1582202>
18. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods*. 2015; 12:357–360. [PubMed: 25751142]
19. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: A next-generation genome browser. *Genome Res*. 2009; 19:1630–1638. [PubMed: 19570905]
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009; 25:2078–2079. [PubMed: 19505943]
21. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010; 26:841–842. [PubMed: 20110278]
22. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics*. 2010; 26:2204–2207. [PubMed: 20639541]
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009; 25:2078–2079. [PubMed: 19505943]
24. Ankley GT, Feifarek D, Blackwell B, Cavallin JE, Jensen KM, Kahl MD, Poole S, Saari T, Villeneuve DL. Re-evaluating the significance of estrone as an environmental estrogen. *Environ Sci Technol*. 2017; 51:4705–4713. [PubMed: 28328210]
25. Van Delft J, Gaj S, Lienhard M, Albrecht MW, Kirpiy A, Brauers K, Claessen S, Lizarraga D, Lehrach H, Herwig R, Kleinjans J. RNA-Seq provides new insights in the transcriptome responses induced by the carcinogen benzo[a]pyrene. *Toxicol Sci*. 2012; 130:427–439. [PubMed: 22889811]
26. Orsini L, et al. *Daphnia magna* transcriptome by RNA-Seq across 12 environmental stressors. *Sci Data*. 2016 Data Descriptor.
27. Hahn CM, Iwanowicz LR, Cornman RS, Mazik PM, Blazer VS. Transcriptome discovery in non-model wild fish species for the development of quantitative transcript abundance assays. *Comp Biochem Physiol D*. 2016; 20:27–40.
28. Conesa A, Madriga P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczekniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016; 17:13. [PubMed: 26813401]
29. Tyers M, Mann M. From genomics to proteomics. *Nature*. 2003; 422:193–197. [PubMed: 12634792]
30. Howe K, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 2013; 496:498–503. [PubMed: 23594743]
31. Kasahara M, et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature*. 2007; 447:714–719. [PubMed: 17554307]
32. Society of Environmental Toxicology and Chemistry. Fathead minnow genome project. Pensacola, FL, USA: 2017. Available from: setac.org/fhm-genome [cited 2017 March 6]

Figure 1.

Overview. Gene prediction and evidence alignment steps were performed on the fathead minnow SOAPdenovo genome assembly. Ab initio gene prediction using AUGUSTUS [7] required no extrinsic data. Fathead minnow expressed sequence tags (EST) and zebrafish coding DNA sequences (CDS) were aligned to the genome assembly using the BLAST [12] and Exonerate [13] tools in conjunction. Fathead minnow RNA Sequencing reads (RNA-seq) were aligned to the genome assembly using HISAT2 [18]. The resulting annotation files were loaded into a genome browser and are available to explore or download at the Fathead minnow genome project [32].

Figure 2.

Illustration of 2-step process used to align expressed sequence tags and coding DNA sequences. Nucleotide BLAST search was used to quickly find scaffolds with similarity to the query sequence. Then the query sequence was aligned with each of the high-scoring scaffolds using Exonerate [13], a program with splice-site-aware alignment methods useful for mapping transcripts to genomic DNA sequences.

Figure 3.

Screenshot of the genome browser in use. This region of the genome appears to encode the ptges3 gene, as evidenced by the gene prediction and aligned zebrafish homolog, fathead minnow expressed sequence tag sequences, and RNA-seq coverage. CDS = coding DNA sequence; FHM = fathead minnow; EST = expressed sequence tag.

AUGUSTUS [7] training comparison. A known subset of fathead minnow genes was used as a reference to evaluate the sensitivity and specificity of gene predictions resulting from the use of different models.

Table 1

Model	Nucleotide Sensitivity	Nucleotide Specificity	Exon Sensitivity	Exon Specificity	Gene Sensitivity	Gene Specificity
Fathead	0.74	0.84	0.28	0.40	0	0
Generic	0.91	0.88	0.58	0.60	0	0
Zebrafish	0.88	0.87	0.66	0.73	0.28	0.28

Table 2

Summary of Evidence Alignment.

	Alignments Attempted	Successfully Aligned	Num. Scaffolds Aligned To ^a
ZF CDSs	44,888	36,911 (82%) ¹	9,228 (13%)
FHM ESTs	240,280	197,408 (82%) ²	11,225 (15%)
FHM RNAseq Reads	469,447,382	361,905,245 (77%) ³	50,917 (70%)

¹ Alignments of >90% query identity and query length to a single genomic scaffold.

² Alignments of >70% identity and >50% query length to a single scaffold.

³ RNA-sequencing read alignments using HISAT2 [18] default settings. These include all reads aligned, including concordant pairs, discordant pairs, and aligned singles.

^a The FHM_SOAPdenovo assembly consists of 73,057 scaffolds (GenBank assembly accession GCA_000700825.1); 25,530 (35%) of the scaffolds contain at least one gene model.

Table 3

Data overlap. Using the genomic coordinates of the gene predictions and evidence alignments, the overlap between the different datasets were determined using the BEDtools [21] software suite. Each combination of datasets was assessed for intersection with 100 base pairs of overlap. The amount and percent of the queries which overlapped with at least one of the subjects is reported.

		Subjects			
		Gene Predictions	Zebrafish CDSs	FHM ESTs	FHM RNAseq Reads
Queries	Gene Predictions	-	16,775 (39%)	16,043 (37%)	37,538 (87%)
	Zebrafish CDSs	33,954 (92%)	-	23,746 (64%)	35,656 (97%)
	FHM ESTs	156,991 (80%)	118,863 (60%)	-	192,762 (98%)