SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 35, No. 1, pp. 212–217, 2016
© 2015 SETAC
Printed in the USA

SEQUENCING AND DE NOVO DRAFT ASSEMBLIES OF A FATHEAD MINNOW (*PIMEPHALES PROMELAS*) REFERENCE GENOME

FRANK R. BURNS,† L. AMARIN COGBURN,† GERALD T. ANKLEY,‡ DANIEL L. VILLENEUVE,‡ ERIC WAITS,§ YUN-JUAN CHANG,|| VICTOR LLACA,# STEPHANE D. DESCHAMPS,# RAYMOND E. JACKSON,†† and ROBERT ALAN HOKE*†

†Haskell Global Centers for Health and Environmental Sciences, E.I. du Pont de Nemours, Newark, Delaware, USA
‡Mid-Continent Ecology Division, US Environmental Protection Agency, Duluth, Minnesota, USA
§US Environmental Protection Agency, Cincinnati, Ohio, USA
||High-Performance Biological Computing, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
#Agricultural Biotechnology, E.I. du Pont de Nemours, Wilmington, Delaware, USA
††Central Research and Development Biotechnology, E.I. du Pont de Nemours, Wilmington, Delaware, USA

(Submitted 10 December 2014; Returned for Revision 11 February 2015; Accepted 27 July 2015)

Abstract: The present study was undertaken to provide the foundation for development of genome-scale resources for the fathead minnow (*Pimephales promelas*), an important model organism widely used in both aquatic toxicology research and regulatory testing. The authors report on the first sequencing and 2 draft assemblies for the reference genome of this species. Approximately 120× sequence coverage was achieved via Illumina sequencing of a combination of paired-end, mate-pair, and fosmid libraries. Evaluation and comparison of these assemblies demonstrate that they are of sufficient quality to be useful for genome-enabled studies, with 418 of 458 (91%) conserved eukaryotic genes mapping to at least 1 of the assemblies. In addition to its immediate utility, the present work provides a strong foundation on which to build further refinements of a reference genome for the fathead minnow. *Environ Toxicol Chem* 2016;35:212–217. © 2015 SETAC

Keywords: Fathead minnow Genome Genetics Ecotoxicogenomics

INTRODUCTION

The fathead minnow, *Pimephales promelas*, has been the most widely used small fish model for regulatory ecotoxicology in North America since the early 1950s [1]. With the establishment of the US Environmental Protection Agency (USEPA) in 1970, fathead minnows were adopted as a primary model organism for standardized regulatory ecotoxicity testing, and numerous test guidelines employing the fathead minnow have been developed and applied both nationally and internationally [2–5]. Because of its prominence as an aquatic test organism and the availability of well-established methods for its culture and husbandry [6], the fathead minnow is also widely used as a research model for studying mechanisms of toxicity [1].

Nonetheless, in the evolving era of toxicogenomic research and subsequent regulatory applications, the fathead minnow has lagged somewhat behind other small fish models, such as the zebrafish (*Danio rerio*) and Japanese medaka (*Oryzias latipes*), in terms of development of genome-scale resources. The zebrafish genome project was initiated by the Sanger Institute in 2001, and the ninth version of the zebrafish genome assembly and annotation was released in 2013 [7]. A draft medaka genome was published in 2007 [8]. Availability of these genome-scale resources greatly enhanced the ability to design and develop molecular research tools (e.g., real-time polymerase chain reaction [PCR] primers, probes for in situ hybridization, morpholinos, small interfering RNAs) and quickly led to commercially available high-density microarray platforms for the zebrafish and Japanese medaka [9]. For example, commercially available zebrafish microarrays were available on both Agilent and Affymetrix platforms as early as 2005 (e.g., Gene Expression Omnibus accessions GPL1319, GPL2878 [10]).

In contrast, as late as 2007, a 2000-gene microarray was state-of-the-art for the fathead minnow [11]. Release of 250 000 fathead minnow expressed sequence tags deposited in the National Center for Biotechnology Information's GenBank, Department of Energy's Joint Genome Institute, in 2005 has since stimulated development of high-density microarray tools for the fathead minnow (e.g., Gene Expression Omnibus accessions GPL10259, GPL10260, GPL10277, GPL9248, GPL7351, GPL7342). More recently, Wiseman et al. [12] constructed a reference hepatic transcriptome for the fathead minnow. However, neither expressed sequence tags nor the reference transcriptome provides complete coverage of expressed sequence and unexpressed regions of the fathead minnow genome (e.g., gene regulatory regions). The fathead minnow was not represented in the 160 fish genomes being sequenced as of 2012 or identified for sequencing under the Genome 10K project [13]. Consequently, to exploit the full potential of this well-established aquatic ecotoxicological model using 21st-century approaches, there remained a need to develop genome-scale resources for the fathead minnow. The lack of a complete genome sequence for the fathead minnow currently limits the dissection of complex traits, genetic marker discovery, identification of gene regulatory domains (e.g., promoters), and the elucidation of biological networks, all of which are critical to fully utilizing this species

This article includes online-only Supplemental Data.

* Address correspondence to robert.a.hoke@dupont.com

Published online 29 October 2015 in Wiley Online Library (wileyonlinelibrary.com).

DOI: 10.1002/etc.3186

for toxicogenomic applications. These are all important components of fathead minnow systems biology and modeling that can help lay the scientific foundation for greater use of predictive approaches in ecotoxicology and ecological risk assessment.

The present study was designed to help address the need for a defined fathead minnow genome by employing massively parallel Illumina® sequencing to generate draft genome sequence information. In contrast to long-read Sanger sequencing employed for the Human Genome Project [14] as well as zebrafish sequencing, the next-generation sequencing (NGS) method employed in the present study offers greater throughput at lower cost [15,16]. However, the increased throughput and reduced cost come at the expense of read length (typically $\leq$100 bp for Illumina sequencing) and the associated difficulties of assembling short reads, particularly for organisms with large repeat-rich regions of their genome, which increases the computational complexity of assembly activities [15–17]. The approach employed in the present study, which involved sequencing of paired-end reads of varying insert sizes as well as fosmid libraries, has been employed previously for sequencing and assembly of draft genome sequences for other vertebrates [18,19]. Although there are no gold-standard methods for determining the ultimate accuracy and completeness of the resulting assemblies [20], common metrics used to evaluate the assemblies include the numbers and sizes of the contigs (continuous sequences assembled from overlapping fragments) and scaffolds (assemblies of contigs in which the order of multiple contigs with gaps between them are known), as well as N50 (the length of contigs and scaffolds at which 50% of the assembly can be found [21]). The present study reports on Illumina-based sequence generation for the fathead minnow and subsequent evaluation of 2 draft assemblies using these metrics. The sequences and assemblies reported are intended as a resource that can be further developed by the community to advance the utility of the fathead minnow as a model organism for predictive, pathway-based ecotoxicology.

MATERIALS AND METHODS

Inbreeding and genetic analysis

Fathead minnows (F0) from an existing culture at the USEPA laboratory in Duluth, Minnesota, were inbred via sibling intercrosses for 6 generations to reduce heterozygosity. Fish representing generations F0, F3, and F6 ($n = 94$) were genotyped at the USEPA laboratory in Cincinnati, Ohio, using 8 polymorphic microsatellite markers—Ppr101, Ppr103-Ppr107 [22], Ppro48, and Ppro126 [23]—according to methods described elsewhere [24]. Briefly, a 3-primer probe system was used to fluorescently label amplicons generated by PCR. Sizes of the amplified microsatellites were visualized using an ABI 3730 genetic analyzer (Applied Biosystems), and GeneMarker software (Ver 1.85; Softgenetics) was used to score polymorphisms. Allelic diversity, mean number of alleles, and Nei's unbiased heterozygosity [25] were determined for each generation. Adult male fish from the F6 generation were shipped overnight from the Duluth USEPA lab to DuPont's Haskell Global Centers laboratory (Newark, DE, USA) for DNA extraction and genome sequencing. Upon receipt, the fish were held in a flow-through culture system at 20 °C with daily feeding until they were processed for sequencing. Fish were euthanized using tricaine methanesulfonate, the skin was dissected away from the lateral tail muscle, and muscle tissue was excised for DNA extraction (see Supplemental Data). Extracted DNA samples were transferred to the DuPont Pioneer sequencing facility at the DuPont Experimental Station (Wilmington, DE, USA) for sequencing.

Next-generation library construction and sequence generation

Four DNA NGS libraries with varying average DNA fragment sizes were generated: 1) a paired-end NGS library with an average fragment size of 180 bases; 2) 2 "mate-pair" NGS libraries with average fragment sizes of 3 kb and 6 kb, respectively; and 3) a fosmid library with an average fragment size of 40 kb (see Supplemental Data for details).

Cluster generation and DNA sequencing

Cluster generation and paired-end sequencing were performed on an Illumina cBot and a HiSeq 2000, respectively, according to protocols developed by Illumina [26]. Sequencing was performed at both ends of the clustered DNA fragments using paired-end sequencing primers for Read1 and Read2 (Illumina) for the paired-end and mate-pair libraries. The fosmid next-generation sequencing library was sequenced using the following custom sequencing primers: read 1, ACACTCTTTCCCTA-CACGACGCTCTTCCGATCTCAC; and read 2, CGGTCTCG GCATTCCTGCTGAACCGCTCTTCCGATCTCAC. For all libraries, the resulting read 1 and read 2 sequences were grouped into "read pairs" according to the x and y coordinates of the corresponding DNA cluster on the flow cell. Sequencing reads and quality scores were generated in a real-time fashion with the Illumina Data Collection Software RTA 1.12. After initial base calling, additional custom filtering was performed using calibrated quality scores generated by the Illumina pipeline. Reads generated from both ends of DNA fragments were trimmed by removing from the 3′ end's bases with a PHRED-equivalent quality score below 10. A length threshold of 24 was applied to filtering, indicating that all bases <24 bases in length after trimming were removed from further analysis.

Jump library construction and sequence generation

DNA was fragmented using a Covaris S220 Ultrasonicator. Fragmented DNA was characterized for size distribution using a BioAnalyzer 2100. Fragment sizes were determined to be 600 bp to 1000 bp. Fragmented DNA was then purified, and adapters for sequencing were ligated per the instructions for the TruSeq DNA PCR-Free Sample Prep kit (Illumina). The resultant sample library was sequenced in 2 runs on the Illumina MiSeq using 2 × 250 base, paired-end settings.

Genome assembly and analysis

Prior to assembly, the quality of the raw sequence reads for each library was assessed using FastQC [27]. Sequences of poor quality were subsequently trimmed from the 3′ end of all reads using the FASTX-toolkit [28]. Two independent assembly pipelines were used. One assembly was prepared using the Short Oligonucleotide Analysis Package (SOAPdenovo) software [29,30], and the other assembly was performed using String Graph Assembler software [31,32]. Assembly details are presented in the Supplemental Data. The SOAPdenovo2 program was able to utilize both the 2 × 100 180-base short reads and the 2 × 250 reads from the PCR-free MiSeq reads for initial contig assembly. For the String Graph Assembler software, initial contig assembly was limited to using only the 2 × 100 data from the 180-base insert sequences. The

2 × 250 reads from the MiSeq 600-base to 1000-base library were alternatively used as a jump library for scaffolding during the String Graph Assembler assembly. Both assembly programs utilized the 3-kb, 6-kb, and 40-kb mate-pair libraries for scaffolding of the contigs. The completeness of the assemblies was assessed for the inclusion of 458 core proteins using the Core Eukaryotic Genes Mapping Approach [33]. Additionally, the assemblies were analyzed for presence, completeness, and presence of orthologs for a set of 248 highly conserved, eukaryotic genes [34].

RESULTS

Six generations of sibling intercrosses were performed to reduce allelic heterozygosity of fish chosen for sequencing. Fish representing generations F0, F3, and F6 ($n = 94$) were analyzed. Heterozygosity decreased from 0.62 (F0) to 0.24 (F6), with the mean number of alleles at each locus reduced from 4.0 to 1.63 (Table 1). The inbreeding coefficient of the resultant F6 fish ($F_{is} = 0.56$) indicates a successful high degree of inbreeding. The relatively homogeneous genomic structure of these fish helped simplify the task of genomic sequencing.

Sequence coverage

Based on a survey of genome sizes among 20 species of North American cyprinids, the overall size of the fathead minnow genome was expected to be in the range of 1.06 Gb to 1.11 Gb [35]. Four sequence libraries derived from different-sized DNA fragments were generated. Sequence coverage was determined using an estimated genome size of 1.1 Gb. Sequence coverage ranged from approximately 74.4× for the 180-bp library to 3.5× for the 6-kb mate-pair library (Table 2). Overall, when the 5 libraries were combined, sequence coverage of approximately 120× was achieved. Raw sequence reads for all libraries were submitted to the National Center for Biotechnology Information Short Read Archive under accession number SRA123892.

Draft genome assemblies

The SOAPdenovo2 assembly generated contigs totaling 811 Mb. The GC content of the contigs was 37.9%. Using the jump libraries, the contigs were linked into scaffolds totaling 1219 Mb with an N content of 33.48%. The longest scaffold generated was 580 kb. The scaffold N50 was 60 380 bases, whereas contig N50 was 7468 bases. This draft genome assembly was submitted to the National Center for Biotechnology Information, and the SOAPdenovo2 assembly from this Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under accession JNCD00000000 (i.e., the version described in the present study).

The String Graph Assembler assembly generated contigs totaling 807 Mb. The GC content of the contigs was 38.0%. Using the jump libraries, the contigs were assembled into scaffolds totaling 958 Mb with an N content of 15.07%. The longest scaffold generated was 811 kb. Scaffold N50 was 15 414 bases, while contig N50 was 1668 bases. This draft genome assembly was submitted to the National Center for Biotechnology Information, and the String Graph Assembler assembly from this Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under accession JNCE00000000 (i.e., the version described in the present study).

Table 1. Reduction in genetic diversity following 6 generations of inbreeding

Generation	n	Heterozygosity (± standard deviation)	Mean number of alleles (± standard deviation)
F0	31	0.62 ± 0.01	4.0 ± 1.07
F3	31	0.35 ± 0.09	2.0 ± 0.93
F6	32	0.24 ± 0.1	1.63 ± 0.74

Table 2. Summary of fathead minnow genome sequence coverage obtained through Illumina sequencing of 4 libraries derived from DNA fragments of varying length

Library	Reads (in millions)	No. of bases (in Gb)	Sequence coverage
180-bp paired end	485.4	81.8	74.4
250-bp paired end	13.9	6.9	6.3
3-kb mate pair	282.3	25.8	23.5
6-kb mate pair	45	3.8	3.5
40-kb fosmid	159.4	15.7	14.3
Overall	986.0	134.0	121.8

Estimation of draft genome completeness

Core Eukaryotic Genes Mapping Approach mapping of 458 core eukaryotic genes [33] to the developed assemblies successfully mapped 405 of 458 of the genes in the SOAPdenovo2 assembly and 310 of the 458 genes in the String Graph Assembler assembly. The majority of the genes (297) mapped in both assemblies. A more stringent analysis that distinguishes target genes from closely related orthologs and indicates whether the complete genes are contained within a single scaffold was also performed on the 248 most highly conserved genes (Table 3) [34]. This analysis indicated that 74% of the genes were complete and present on a single scaffold in the SOAPdenovo2 assembly compared with 51% in the String Graph Assembler assembly.

DISCUSSION

In the present study, we describe the successful sequencing and de novo assembly of 2 draft genomes for the fathead minnow, a fish species used extensively both for single-chemical testing and monitoring complex environmental mixtures of contaminants [1]. An important historical limitation of utilization of toxicogenomic approaches with the fathead minnow to support the development of mechanistic predictive and diagnostic tools has been the lack of genome-level information. The present study addresses that limitation. The nature of research and regulatory studies with the fathead minnow increasingly has included toxicogenomic approaches in an effort to address some of the challenges currently faced by environmental toxicologists, such as the need for the rapid generation of hazard data for a greater number of chemicals using fewer resources or for efficient monitoring of the potential impacts of complex mixtures. For example, Ankley et al. [36] describe research with the fathead minnow to develop mechanistic, predictive approaches to support mandated screening and testing programs for identifying endocrine-active chemicals. Analogously, Ekman et al. [37] propose the use of pathway-based techniques with the fathead minnow for effects-based monitoring of complex mixtures of environmental contaminants in a field setting.

The 120× genome coverage achieved in the present study provided a read coverage similar to or greater than a number

Table 3. Estimated completeness of the genome based on mapping 248 conserved eukaryotic genes to the assemblies

Assembly	Complete genes present in single scaffold, *n* (%)	Average of orthologs per complete gene in assembly, *n*	Genes mapped but not complete in single scaffold, *n* (%)	Average of orthologs per incomplete gene in assembly, *n*	Total genes mapped, *n* (%)
SOAPdenovo2	183 (74%)	1.4	43 (17%)	3.4	226 (91%)
String Graph Assembler	126 (51%)	1.24	40 (16%)	2	166 (67%)

SOAPdenovo2 = Short Oligonucleotide Analysis Package software.

of pioneering de novo, high-throughput vertebrate genome sequencing projects. For example, employing a similar approach to develop a draft sequence of the giant panda (*Ailuropoda melanoleuca*) genome, Li et al. [19] achieved 73× coverage of the whole genome. Assembly of a draft chicken genome using Illumina sequencing and SOAPdenovo 1.04 was based on 74× genome coverage [17]. A domestic turkey (*Meleagris gallopavo*) genome was assembled from approximately 5× coverage of 454 reads coupled with 25× coverage of Illumina sequencing reads [38]. Based on a comparison of N50 values for assemblies using various sequencing depths of Illumina reads, there is relatively little gain in overall length of contigs or scaffolds with additional sequencing depth beyond approximately 70× [39]. Consequently, the overall sequence coverage achieved through the present study was adequate to facilitate assembly using state-of-the-science de novo assembly algorithms, and there is likely little to be gained through additional depth of sequencing.

At present, the greatest challenge in de novo sequencing of vertebrate genomes is the assembly of contigs and scaffolds from the short reads associated with the Illumina sequencing [20].

The total scaffold size of the 2 scaffolded assemblies generated for the fathead minnow genome was 1.219 Gb for the SOAPdenovo2 assembly and 0.958 Gb for the String Graph Assembler assembly. A genome size of 1.1 Gb was used to estimate the sequence coverage of the generated libraries. Thus, both assemblies fall within 10% of the expected *P. promelas* genome size range. This suggests that the sequence assemblies encompass a large majority, if not all, of the fathead minnow genome. Nonetheless, there is likely room for improvement in the assemblies. The contig N50 of 7.5 kb for the SOAPdenovo2 assembly was similar to that of 9.8 kb for the medaka draft genome [8], 9.2 kb for the Japanese lamprey (*Lethenteron japonicum* [40]), and 7.1 kb for the Atlantic cod (*Gadus morhua*) Celera assembly [41]. In contrast, the contiguity of the reference fathead minnow genome assemblies remains less than that achieved for channel catfish (*Ictalurus punctatus*, 13.1 kb [39]), Atlantic salmon (*Salmo salar*, 35 kB; GenBank Assembly GCA_000233375.3), and zebrafish (1.6 Mb; GenBank Assembly GCA_000002035.2). That said, many existing fish genome sequencing projects have employed longer read technologies (e.g., Sanger sequencing, 454 sequencing) and have gone through multiple rounds of assembly and revision. Thus, it can be expected that the current level of contiguity can be improved by integrating results of multiple assemblies based on different algorithms with their respective strengths and weaknesses [42].

Concern has been raised that contiguity as measured via N50 or total contig and scaffold lengths are not necessarily the best indicators of assembly quality [20,33,42]. Consequently, the Core Eukaryotic Genes Mapping Approach was employed as an additional measure of assembly quality. On the basis of Core Eukaryotic Genes Mapping Approach mapping and contiguity metrics (e.g., N50), the SOAPdenovo2 algorithm appeared to provide a higher-quality assembly of the fathead minnow genome than the String Graph Assembler algorithm. However, as attested to by the mapping of the Core Eukaryotic Genes Mapping Approach gene set, the generation of 2 assemblies by different informatics pipelines provides for more comprehensive coverage than either of the assemblies would have individually and prompted our reporting and databank submission of both assemblies. In combination, the 2 assemblies collectively mapped 418 out of 458 eukaryotic core genes (91.2%). This value was within the range of 85% to 95% reported for the range of assembly algorithms tested in Assemblathon 2 [42]. This success rate in the mapping of core genes was achieved despite a greater proportion of genes in both assemblies with paralogs (1.7×) versus those without paralogs that were not complete on a single scaffold. The existence of large numbers of paralogs in fish is a recognized challenge in fish genome assembly, stemming from genome duplication in the ray-finned fish over 300 million yr ago [43]. Consistent with that complexity, of 3 vertebrate classes (birds, fish, reptiles) considered in Assemblathon 2, fish were identified as the most challenging to assemble [42]. Therefore, on the whole, the draft assembly appears to be of comparable quality to the state of the art in the field for Illumina sequencing, particularly given some of the recognized challenges with sequencing fish genomes.

This de novo sequencing project, and continued evolution and improvement of the sequence assembly, is expected to provide a variety of benefits to aquatic toxicology and ecotoxicogenomics research efforts, including increasing the feasibility and cost-effectiveness of assaying the entire fathead minnow transcriptome using microarrays; enhancing interpretation of transcriptomic, proteomic, and metabolomic data from fathead minnow studies from a systems biology/network perspective; identification of high-quality markers of fathead minnow genetic sex to support reproduction and sexual determination regulatory studies with fathead minnows; facilitating environmental monitoring of changes in genetic structure that could be indicative of system impairment or alteration in wild populations of fathead minnows (and potentially other species of freshwater fish); and facilitating genome resequencing projects and comparison of genetic structure among different populations of fathead minnows that may be useful in identifying genetic features/markers that confer resistance or sensitivity to particular types of pollutants or stressors, which could be useful for developing minimally invasive genetic screens for predicting population susceptibility to different exposure scenarios (e.g., spills, point source discharges, overspray events). The work of Head et al. [44] provides an example in avian species of the type of genetic markers of sensitivity/susceptibility that could be more readily developed with a fully sequenced and mapped genome for the fathead minnow.

Species extrapolation is another fundamental challenge in ecotoxicology and ecological risk assessment. It is not feasible to test all species that may be exposed in the environment, nor is it possible to develop species-specific batteries of pathway-based suborganismal bioassays. Therefore, predictive approaches to species extrapolation that utilize available molecular sequence data are being explored as a possible aid to prioritization and risk assessment (e.g., Gunnarsson et al. [45] and LaLone et al. [46]). Such efforts would be enhanced by the availability of whole-genome sequence information for a broader diversity of species, particularly for widely used aquatic toxicity test organisms such as the fathead minnow. Transcriptional network modeling also is being applied to elucidate adverse outcomes pathways in the fathead minnow (and by extension other vertebrates) and identify functional molecular modules that play key roles in toxicity [47,48]. Although these approaches have already shown preliminary utility, their application and potential will be considerably enhanced by genome sequencing and the associated identification of gene regulatory sequences (e.g., transcription factor binding sites).

CONCLUSION

Although design of targeted molecular tools for the fathead minnow has been enhanced by the availability of expressed sequence tags, coverage of relevant molecular targets was incomplete, and nonexpressed portions of the genome were largely uncharacterized in the species. Availability of a fully sequenced genome generates significant efficiencies to enhance the application of transcriptomic, proteomic, and many other focused molecular and genetic analyses across the field. The present study does not provide extensive annotation, comparative and evolutionary analysis, or targeted mapping of genome completeness relative to the previously reported fathead minnow transcriptome. Nonetheless, the effort represents a significant and critical step forward in the generation of genome-scale resources for the fathead minnow and provides a new resource of sequence information which is freely accessible to the broader scientific community for further development.

Supplemental Data—Supplemental Data are available on the Online Wiley Library at DOI: 10.1002/etc.3186.

Acknowledgment—We thank the University of Illinois HPCBio group for computing resources and guidance on the SOAPdenovo2 assembly, J. Simpson for assistance with the String Graph Assembler, and the Korf laboratory at the University of California at Davis for assistance with the Core Eukaryotic Genes Mapping Approach analysis. Resources to support fathead minnow inbreeding and genetic analysis and continued development of genome-scale informatics for the species were supported by a USEPA Office of Research and Development Pathfinder Innovation Project Award. The contents of the present study neither constitute nor necessarily reflect official USEPA policy. Mention of trade names or commercial products does not constitute USEPA endorsement or recommendation for use.

Data availability—Data are available from the National Center for Biotechnology Information database; access numbers are indicated in the manuscript.

REFERENCES

1. Ankley GT, Villeneuve DL. 2006. The fathead minnow in aquatic toxicology: Past, present and future. *Aquat Toxicol* 78:91–102.
2. US Environmental Protection Agency. 1989. Pesticide assessment guidelines. Subdivision E, hazard evaluation: Wildlife and aquatic organisms. Washington, DC.
3. US Environmental Protection Agency. 1994. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms, 3rd ed. Washington, DC.
4. US Environmental Protection Agency. 2009. Fish short-term reproduction assay. Endocrine Disruptor Screening Program Test Guidelines. OPPTS 890.1350, EPA 740/C09/007. Washington, DC.
5. Organisation for Economic Co-operation and Development. 1992. Test No. 210: Fish early life-stage toxicity test. *OECD Guidelines for the Testing of Chemicals*. Paris, France.
6. US Environmental Protection Agency. 1987. Guidelines for the culture of fathead minnows (*Pimephales promelas*) for use in toxicity tests. Duluth, MN.
7. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch GJ, White S, Chow W, Kilian B, Quintais LT, Guerra-Assunção JA, Zhou Y, Gu Y, Yen J, Vogel JH, Eyre T, Redmond S, Banerjee R, Chi J, Fu B, Langley E, Maguire SF, Laird GK, Lloyd D, Kenyon E, Donaldson S, Sehra H, Almeida-King J, Loveland J, Trevanion S, Jones M, Quail M, Willey D, Hunt A, Burton J, Sims S, McLay K, Plumb B, Davis J, Clee C, Oliver K, Clark R, Riddle C, Elliot D, Threadgold G, Harden G, Ware D, Begum S, Mortimore B, Kerry G, Heath P, Phillimore B, Tracey A, Corby N, Dunn M, Johnson C, Wood J, Clark S, Pelan S, Griffiths G, Smith M, Glithero R, Howden P, Barker N, Lloyd C, Stevens C, Harley J, Holt K, Panagiotidis G, Lovell J, Beasley H, Henderson C, Gordon D, Auger K, Wright D, Collins J, Raisen C, Dyer L, Leung K, Robertson L, Ambridge K, Leongamornlert D, McGuire S, Gilderthorp R, Griffiths C, Manthravadi D, Nichol S, Barker G, Whitehead S, Kay M, Brown J, Murnane C, Gray E, Humphries M, Sycamore N, Barker D, Saunders D, Wallis J, Babbage A, Hammond S, Mashreghi-Mohammadi M, Barr L, Martin S, Wray P, Ellington A, Matthews N, Ellwood M, Woodmansey R, Clark G, Cooper J, Tromans A, Grafham D, Skuce C, Pandian R, Andrews R, Harrison E, Kimberley A, Garnett J, Fosker N, Hall R, Garner P, Kelly D, Bird C, Palmer S, Gehring I, Berger A, Dooley CM, Ersan-Ürün Z, Eser C, Geiger H, Geisler M, Karotki L, Kirn A, Konantz J, Konantz M, Oberländer M, Rudolph-Geiger S, Teucke M, Lanz C, Raddatz G, Osoegawa K, Zhu B, Rapp A, Widaa S, Langford C, Yang F, Schuster SC, Carter NP, Harrow J, Ning Z, Herrero J, Searle SM, Enright A, Geisler R, Plasterk RH, Lee C, Westerfield M, de Jong PJ, Zon LI, Postlethwait JH, Nüsslein-Volhard C, Hubbard TJ, Roest Crollius H, Rogers J, Stemple DL. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496:498–503.
8. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, Jindo T, Kobayashi D, Shimada A, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto S, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin-I T, Takeda H, Morishita S, Kohara Y. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.
9. Douglas SE. 2006. Microarray studies of gene expression in fish. *Omics* 10:474–489.
10. National Center for Biotechnology Information. 2015. Gene Expression Omnibus. [cited 2014 December 4]. Available from: http://www.ncbi.nlm.nih.gov/geo/
11. Larkin P, Villeneuve DL, Knoebl I, Miracle AL, Carter BJ, Liu L, Denslow ND, Ankley GT. 2007. Development and validation of a 2,000-gene microarray for the fathead minnow (*Pimephales promelas*). *Environ Toxicol Chem* 26:1497–1506.
12. Wiseman SB, He Y, Gamal-El Din M, Martin JW, Jones PD, Hecker M, Giesy JP. 2013. Transcriptional responses of male fathead minnows exposed to oil sands process-affected water. *Comp Biochem Physiol C* 157:227–235.
13. Bernadi G, Wiley EO, Mansour H, Miller MR, Orti G, Haussler D, O'Brien SJ, Ryder OA, Venkatesh B. 2012. The fishes of Genome 10K. *Mar Genomics* 7:3–6.
14. International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945.
15. Henson J, Tischler G, Ning Z. 2012. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 13:901–915.
16. Metzker ML. 2010. Sequencing technologies—The next generation. *Nat Rev Genet* 11:31–46.
17. Ye L, Hillier LW, Minx P, Thane N, Locke DP, Martin JC, Chen L, Mitreva M, Miller JR, Haub KV, Dooling DJ, Mardis ER, Wilson RK, Weinstock GM, Warren WC. 2011. A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biol* 12:R31.
18. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello

M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108:1513–1518.

19. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* 463:311–317.
20. Baker M. 2012. De novo genome assembly: What every biologist should know. *Nat Methods* 9:333–337.
21. Michael TP, Jackson S. 2013. The first 50 plant genomes. *Plant Genome* 6:1–7.
22. Ardren WR, Miller LM, Kime JA, Kvitrud MA. 2002. Microsatellite loci for fathead minnow (*Pimephales promelas*). *Mol Ecol Notes* 2:226–227.
23. Bessert ML, Orti G. 2003. Microsatellite loci for paternity analysis in the fathead minnow, *Pimephales promelas* (Teleostei: Cyprinidae). *Mol Ecol Notes* 3:532–534.
24. Waits ER, Nebert DW. 2011. Genetic architecture of susceptibility to PCB126-induced developmental cardiotoxicity in zebrafish. *Toxicol Sci* 122:466–475.
25. Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590.
26. Illumina. 2012. *cBOT™ User Guide*. Part 15006165 Rev. K. San Diego, CA, USA.
27. Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
28. Hannon Lab. 2010. *FASTX-Toolkit*. [cited 2014 December 4]. Available from: http://hannonlab.cshl.edu/fastx_toolkit/
29. Beijing Genomics Institute. 2010. *Short Oligonucleotide Analysis Package (SOAP)*. [cited 2014 December 4]. Available from: http://soap.genomics.org.cn/soapdenovo.html
30. 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, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18.
31. GitHub. 2015. SGA – String Graph Assembler. [cited 2014 December 4]. Available from: http://github.com/jts/sga
32. Simpson JT, Durban R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 22:549–556.
33. Parra G, Bradnam K, Korf I . 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23: 1061–1067.
34. Parra G, Bradnam K, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Res* 37:289–297.
35. Gold JR, Amemiya CT. 1987. Genome size variation in North American minnows (Cyprinidae) II. Variation among 20 species. *Genome* 29:481–489.
36. Ankley GT, Bencic D, Breen M, Collette TW, Conolly R, Denslow ND, Edwards S, Ekman DR, Jensen KM, Lazorchak J, Martinovic D, Miller DH, Perkins EJ, Orlando EF, Garcia-Reyero N, Villeneuve DL, Wang R-L, Watanabe K. 2009. Endocrine disrupting chemicals in fish: Developing exposure indicators and predictive models of effects based on mechanisms of action. *Aquat Toxicol* 92:168–178.
37. Ekman DR, Ankley GT, Blazer VS, Collette TW, Garcia-Reyero N, Iwanowicz LR, Jorgenson ZG, Lee KE, Mazik PM, Miller DH, Perkins EJ, Smith ET, Tietge JE, Villeneuve DL. 2013. Biological effects-based tools for monitoring impacted surface waters in the Great Lakes: A multi-agency program in support of the GLRI. *Environ Pract* 15:409–426.
38. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le Ann, Bouffard P, Burt DW, Crasta O, Crooijmans RP, Cooper K, Coulombe RA, De S, Delany ME, Dodgson JB, Dong JJ, Evans C, Frederickson KM, Flicek P, Florea L, Folkerts O, Groenen MA, Harkins TT, Herrero J, Hoffmann S, Megens HJ, Jiang A, de Jong P, Kaiser P, Kim H, Kim KW, Kim S, Langenberger D, Lee MK, Lee T, Mane S, Marcais G, Marz M, McElroy AP, Modise T, Nefedov M, Notredame C, Paton IR, Payne WS, Pertea G, Prickett D, Puiu D, Qioa D, Raineri E, Ruffier M, Salzberg SL, Schatz MC, Scheuring C, Schmidt CJ, Schroeder S, Searle SM, Smith EJ, Smith J, Sonstegard TS, Stadler PF, Tafer H, Tu ZJ, Van Tassell CP, Vilella AJ, Williams KP, Yorke JA, Zhang L, Zhang HB, Zhang X, Zhang Y, Reed KM. 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): Genome assembly and analysis. *PLoS Biol* 8:9.
39. Jiang L, Lu J, Peatman E, Kucuktas H, Liu S, Wang S, Sun F, Liu Z. 2011. A pilot study for channel catfish whole genome sequencing by de novo assembly. *BMC Genomics* 12:269.
40. Mehta TK, Ravi V, Yamasaki S, Lee AP, Lian MM, Tay B-H, Tohari S, Yanai S, Tay A, Brenner S, Venkatesh B. 2013. Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*). *Proc Natl Acad Sci USA* 110:16044–16049.
41. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB. 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 477:207–210.
42. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam TW, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, Maccallum I, Macmanes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu SM, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF. 2013. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2:10.
43. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res* 13:382–390.
44. Head JA, Hahn ME, Kennedy SW. 2008. Key amino acids in the aryl hydrocarbon receptor predict dioxin sensitivity in avian species. *Environ Sci Technol* 42:7535–7541.
45. Gunnarsson L, Jauhiainen A, Kristiansson E, Nerman O, Larsson DG. 2008. Evolutionary conservation of human drug targets in organisms used for environmental risk assessments. *Environ Sci Technol* 42:5807–5813.
46. LaLone CA, Villeneuve DL, Burgoon LD, Russom CL, Helgen HW, Berninger JP, Tietge JE, Severson MN, Cavallin JE, Ankley GT. 2013. Molecular target sequence similarity as a basis for species extrapolation to assess the risk of chemicals with known modes of action. *Aquat Toxicol* 144/145:141–154.
47. Edwards SW, Preston RJ. 2008. Systems biology and mode of action based risk assessment. *Toxicol Sci* 106:312–318.
48. Perkins EJ, Chipman K, Edwards S, Habib T, Falciani F, Taylor R, Van Aggelen G, Vulpe C, Antczak P, Loguinov A. 2011. Reverse engineering adverse outcome pathways. *Environ Toxicol Chem* 30:22–38.