The Genome of the Self-Fertilizing Mangrove Rivulus Fish, *Kryptolebias marmoratus*: A Model for Studying Phenotypic Plasticity and Adaptations to Extreme Environments

Joanna L. Kelley¹,*, Muh-Ching Yee², Anthony P. Brown¹, Rhea R. Richardson³, Andrey Tatarenkov⁴, Clarence C. Lee⁵, Timothy T. Harkins⁵, Carlos D. Bustamante³, and Ryan L. Earley⁶

¹School of Biological Sciences and Center for Reproductive Biology, Washington State University, Pullman, Washington
²Department of Plant Biology, Carnegie Institution for Science, Stanford, California
³Department of Genetics, Stanford University, Stanford, California
⁴Department of Ecology and Evolutionary Biology, University of California, Irvine, California
⁵Thermo Fisher Scientific, Carlsbad, California
⁶Department of Biological Sciences, University of Alabama, Tuscaloosa, Alabama

*Corresponding author: E-mail: joanna.l.kelley@wsu.edu.
Accepted: June 8, 2016
Data deposition: Sequence data have been deposited under NCBI BioProject PRJNA290522. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession LHSH00000000. The version described in this paper is version LHSH01000000.

Abstract

The mangrove rivulus (*Kryptolebias marmoratus*) is one of two preferentially self-fertilizing hermaphroditic vertebrates. This mode of reproduction makes mangrove rivulus an important model for evolutionary and biomedical studies because long periods of self-fertilization result in naturally homozygous genotypes that can produce isogenic lineages without significant limitations associated with inbreeding depression. Over 400 isogenic lineages currently held in laboratories across the globe show considerable among-lineage variation in physiology, behavior, and life history traits that is maintained under common garden conditions. Temperature mediates the development of primary males and also sex change between hermaphrodites and secondary males, which makes the system ideal for the study of sex determination and sexual plasticity. Mangrove rivulus also exhibit remarkable adaptations to living in extreme environments, and the system has great promise to shed light on the evolution of terrestrial locomotion, aerial respiration, and broad tolerances to hypoxia, salinity, temperature, and environmental pollutants. Genome assembly of the mangrove rivulus allows the study of genes and gene families associated with the traits described above. Here we present a de novo assembled reference genome for the mangrove rivulus, with an approximately 900 Mb genome, including 27,328 annotated, predicted, protein-coding genes. Moreover, we are able to place more than 50% of the assembled genome onto a recently published linkage map. The genome provides an important addition to the linkage map and transcriptomic tools recently developed for this species that together provide critical resources for epigenetic, transcriptomic, and proteomic analyses. Moreover, the genome will serve as the foundation for addressing key questions in behavior, physiology, toxicology, and evolutionary biology.

Key words: genome, fish, mangrove rivulus, hermaphrodite, isogenic.
primary males developing from embryos incubated at low temperatures (18–20 °C; [Harrington 1967; Harrington and Kallman 1968; Ellison et al. 2015]), and remains sexually labile throughout adulthood (Harrington 1971, 1975; Turner et al. 2006; Garcia et al. 2016) (fig. 1). No functional females have been identified in the wild or in captivity (Harrington 1961, 1975; Farmer and Orlando 2012) leading to their classification as an androdioecious vertebrate; however, histological studies have shown that some animals possess only ovarian tissue when young (Cole and Noakes 1997), leading to their classification as an androdioecious vertebrate. Male abundance varies among populations from 0% to 24% (Turner, Davis, et al. 1992) and there is evidence for correspondingly variable outcrossing rates in some locales (e.g., Central America; Tatarenkov et al. 2015). While males have not been observed in all populations, microsatellite data from wild-caught individuals successfully identified highly homozygous lineages as well as individuals with heterozygosity expected under random mating, suggesting that outcrossing between hermaphrodites and males is common in some populations (Taylor et al. 2001; Mackiewicz, Tatarenkov, Taylor, et al. 2006; Earley et al. 2012); hermaphrodites do not mate with one another (Furness et al. 2015).

These features of mangrove rivulus biology have motivated significant efforts into understanding the environmental (Harrington 1967, 1968), and physiological/molecular (e.g., [Kanamori et al. 2006; Orlando et al. 2006; Park et al. 2013]) mechanisms underlying sex determination. The coupling of these features with multiple genetically distinct lineages for many populations, evidence for strong population genetic structure (Tatarenkov et al. 2007, 2015), and variation among lineages and populations in the extent to which environmental factors induce male development (Harrington 1968, 1975; Kristensen 1970; Ellison et al. 2015) position mangrove rivulus as a strong model for understanding sex ratio evolution and the evolutionary ecology of sexual plasticity (e.g., [Earley et al. 2012; Ellison et al. 2015]).

Mangrove rivulus have been maintained in laboratory settings for over 50 years (Harrington 1961). There are currently over 400 distinct lineages being maintained in laboratories throughout the globe (Earley R, personal communication), and genetic divergence among lineages has been assessed primarily via microsatellite typing and restriction-site associated DNA sequencing approaches (Mackiewicz, Tatarenkov, Turner et al. 2006; Tatarenkov et al. 2010, 2012). These lineages are easily reared, reproduce readily via self-fertilization when individually housed, and have a relatively short generation time (ca. 3–5 months). Wild-caught lineages that are highly heterozygous can, if given the opportunity only to self, produce offspring that are homozygous at all microsatellite loci after approximately 5–10 generations (Mackiewicz, Tatarenkov, Perry, et al. 2006), which effectively generates many isogenic lineages with different allele combinations of the progenitor. Completely homozygous genotypes occur naturally as well (Kallman and Harrington 1964; Taylor et al. 2001; Mackiewicz, Tatarenkov, Taylor, et al. 2006; Mackiewicz, Tatarenkov, Turner, et al. 2006; Tatarenkov et al. 2007, 2012, 2015), which allows for the immediate production of isogenic lineages if those animals exclusively self-fertilize in the laboratory. The reproductive biology of mangrove rivulus provides a unique opportunity to study, with extraordinary resolution, the contribution of genes, environment, and gene-by-environment interactions to the developing phenotype and simplifies quantitative genetic approaches to investigating phenotypic evolution. Genetically identical animals can be maintained in different environments to examine phenotypically plastic responses to an arsenal of ecologically relevant contexts (e.g., salinity, temperature: Lin and Dunson 1999; air exposure: Wells et al. 2015) and to construct lineage-specific reaction norms.

**Fig. 1.**—Photograph of hermaphrodite (top) and secondary male (bottom). Secondary males lack ocellus and have a pinker coloration than hermaphrodites. Photo credits to D. Scott Taylor.

Indeed, virtually all studies that have employed different lineages demonstrate, under common garden conditions, significant among-lineage variation for phenotypic characteristics. These range from sensitivity of the sexual phenotype to early-life or adult temperature regimes (Harrington 1967, 1968, 1971, 1975; Kristensen 1970; Turner et al. 2006; Ellison et al. 2015), some aspects of endocrine function (e.g., [Earley and Hsu 2008; Earley et al. 2013]), risk-taking and exploratory behavior (Edenbrow and Croft 2011, 2012), performance in aggressive contests (Hsu et al. 2008), the propensity to voluntarily jettison from the water (Turko et al. 2011), and life history traits (Garcia et al. 2015). These data indicate that there is considerable heritable variation among genotypes that persists for generations under laboratory conditions.

Another key aspect of mangrove rivulus biology is that they exhibit a number of remarkable adaptations to living in the notoriously variable, sometimes noxious, upland mangrove habitat. These environments are tidal, which can leave the fish stranded on moist land or in stagnant crab burrows for significant periods of time, and can expose the fish to salinities ranging from 0 to 60 parts per thousand (ppt), low dissolved oxygen levels, high hydrogen sulfide levels, and extreme temperatures (Taylor 2012). Mangrove rivulus can live for extended periods of time on land (up to 2 months), often occupying rotting logs and leaf litter (Taylor et al. 2008), and will jettison from water in response to changes in their aquatic environment, which they perceive via a number of specialized mechanisms (e.g., neuroepithelial cells; Robertson et al. 2015). These fish can navigate on land using a variety of intricate movements (Pronko et al. 2013), some that involve the fish effectively turning itself into a projectile to launch off of the substrate (Ashley-Ross et al. 2014).

Emersion onto land also can function in adult thermoregulation (Gibson et al. 2015), and mangrove rivulus will deposit fertilized eggs on land, which appears to reduce oxygen consumption without detrimental consequences to the adult phenotype (Wells et al. 2015). When on land, mangrove rivulus deploy a number of physiological changes that promote aerial respiration, osmoregulation, and ionoregulation (Wright 2012; Turko et al. 2014) and they remodel their gills in ways that reduce surface area for gas exchange (Wright 2012).

The coastal habitats of mangrove rivulus, which are being developed at an extraordinary rate, are subject to significant anthropogenic disturbance, particularly with respect to the influx of environmental pollutants (e.g., endocrine disruptors) from urban run-off and wastewater treatment plants. Research into the effects of such pollutants has demonstrated staggering, often sex-specific, effects on gene expression, endocrine function, reproduction, and survival (Lee et al. 2006, 2007; Rhee et al. 2008, 2010, 2011; Rhee, Lee, et al. 2009; Rhee, Kang, et al. 2009; Farmer and Orlando 2012; Johnson et al. 2016), leading to rivulus being promoted as an important model system for ecotoxicology (Davis 1986; Lee et al. 2008). Most studies on rivulus’ physiological and locomotory adaptations to extreme environments and on their responses to environmental pollutants use one or very few lineages, which limits our ability to understand the full scope of variation on which selection can act to drive genetic and phenotypic divergence among populations. Another limitation is access to the genetic and genomic resources necessary to provide a framework for our understanding of both the mechanisms underlying and the evolution of the whole-organism phenotype.

A linkage map is now available (Kanamori et al. 2016) and N-ethyl N-nitrosourea (ENU) mutagenesis screens have been successful in this species (Moore et al. 2012;Sucar et al. 2016). A transcriptome was recently made available, and shed light on genes that are up- and downregulated during embryonic diapause (the embryo is fully developed but hatching is delayed; [Mesak et al. 2015]). In addition, there has been some success with producing crosses between phenotypically distinct rivulus lineages (Harrington 1971; Mackiewicz, Tatarenkov, Perry, et al. 2006; Nakamura et al. 2008). One cross between lineages with divergent growth patterns showed that F2s have intermediate phenotypes (Nakamura et al. 2008). Although outcrossing frequencies are low, even under controlled laboratory conditions (e.g., 6–8% Mackiewicz, Tatarenkov, Perry, et al. 2006; in vitro: Nakamura et al. 2008), there is great promise for generating crosses between phenotypically dissimilar lineages that could produce an arsenal of recombinant inbred lines that would rival the resources available for Arabidopsis thaliana, Caenorhabditis elegans, and Drosophila melanogaster (Abzhanov et al. 2008; Mackay et al. 2009, 2014; Lehner 2013; Rankin 2015). Also, our growing understanding of embryogenesis (Mourabit et al. 2011; Mourabit and Kudoh 2012) will facilitate genetic manipulation during early development.

Here, we present the annotated genome sequence for the mangrove rivulus from a lab-reared population of the Reckley Hill Lake (RHL) lineage from San Salvador, Bahamas. We annotated 27,328 protein-coding genes. We compared the K. marmoratus reference genome to genomic data generated for the sister species K. hemaphroditus (see fig. 2 for the relationship of Kryptolebias to several model fish species). The availability of a genome assembly for K. marmoratus provides the basis for future comparative genomic and phenotypic studies and, given increased availability of wild-caught isogenic lineages and information about population genetic structure, for highly controlled studies on environmental sex variation in reaction norm characteristics (Earley et al. 2012).
determination, the evolution of phenotypic plasticity, and adaptations to extreme environments.

**Results**

**Genome Sequencing, Assembly, and Annotation**

The *K. marmoratus* genome is composed of 24 haploid chromosomes (Scheel 1972), and the genome size was estimated to be 0.936 picograms, based on flow cytometry (Kiehart R, Goddard K, and Turner B, unpublished data) which translates to approximately 915 megabase pairs (Mb). The genome size estimate was consistent with the k-mer based estimate from the raw genomic reads, which estimates the genome size to be 830 Mb. A total of 289,349,170 paired-end (2 × 101 bp) reads were included in the short read assembly. Additionally, mate-pair reads were trimmed and filtered for the adapter using Ion Torrent PGM software for a total of 15,457,550 mate-pair reads postfiltering with an average read length of 87 bp (supplementary table S1, Supplementary Material online). The reads were assembled and the resulting draft genome consists of 40,758 scaffolds, for a total of 654 Mb assembled with 3.77% ambiguous bases (table 1). The assembled genome size is similar to the genome size estimated from the k-mer based estimate from the raw genomic reads, which estimates the genome size to be 830 Mb. A total of 289,349,170 paired-end (2 × 101 bp) reads were included in the short read assembly. Additionally, mate-pair reads were trimmed and filtered for the adapter using Ion Torrent PGM software for a total of 15,457,550 mate-pair reads postfiltering with an average read length of 87 bp (supplementary table S1, Supplementary Material online). The reads were assembled and the resulting draft genome consists of 40,758 scaffolds, for a total of 654 Mb assembled with 3.77% ambiguous bases (table 1). The assembled genome size is similar to the genome size estimated from the sum of contigs from two other de novo genome assemblies (642 Mb [Mesak et al. 2015] and 633 Mb [Rhee and Lee 2014]). The longest scaffold is 1,759,740 bp and the N50 scaffold length is 111,539 bp. The de novo genome assembly was combined with the recently published linkage map (Kanamori et al. 2016). Of the 9,904 markers used to construct the linkage map, 9,855 markers map uniquely (99.2% of markers) to 4,871 scaffolds, 21 do not map to the genome assembly, and 28 do not map uniquely (map to two positions). More than 396.5 Mb of the assembled genome is placed in linkage map (supplementary table S2, Supplementary Material online).

**Table 1**

<table>
<thead>
<tr>
<th>Genome Assembly Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genome</strong></td>
</tr>
<tr>
<td>Size (1n)</td>
</tr>
<tr>
<td>Karyotype</td>
</tr>
<tr>
<td>GC content</td>
</tr>
<tr>
<td><strong>Assembly</strong></td>
</tr>
<tr>
<td>Size in scaffolds &gt; 500bp</td>
</tr>
<tr>
<td>Number of scaffolds &gt;500bp</td>
</tr>
<tr>
<td>Number of scaffolds &gt; 10kb</td>
</tr>
<tr>
<td>Base pairs placed in linkage map</td>
</tr>
<tr>
<td>Number of scaffolds placed in linkage map</td>
</tr>
<tr>
<td>N50 scaffold</td>
</tr>
<tr>
<td>N50 contig</td>
</tr>
<tr>
<td><strong>Annotation</strong></td>
</tr>
<tr>
<td>Coding loci</td>
</tr>
<tr>
<td>Noncoding loci</td>
</tr>
</tbody>
</table>

Evidence-guided annotation of the genome assembly resulted in 27,328 protein-coding genes, 26 ribosomal RNAs, 536 transfer RNAs, and 814 other noncoding RNAs including microRNAs and small nucleolar RNAs (supplementary tables S3 and S4, Supplementary Material online). Of the protein-coding genes, 84% share sequence similarity to 16,919 proteins in the SWISS-PROT database (e-value 10 ^ -10 ) (Boeckmann et al. 2003). The mitochondrial genome is absent from our assembly due to the high raw read coverage.

The gene regions for *K. marmoratus* in this assembly are relatively well assembled. Of the highly conserved core eukaryotic genes determined by CEGMA, 208 complete proteins and an additional 28 partial proteins were identified in the genome assembly, which, in total, correspond to a 95% of CEGMA proteins (supplementary table S5, Supplementary Material online). Moreover, 76.5% complete and 10.2% fragmented single-copy vertebrate orthologs were identified using Benchmarking Universal Single-Copy Orthologs (BUSCO) (supplementary table S6, Supplementary Material online). This completeness of BUSCO genes is similar to other widely used and well-assembled fish genomes (supplementary table S7, Supplementary Material online) To support the completeness of the genome, one of the short-insert libraries used for the genome assembly maps back to the reference genome with a mapping rate of 98.7%. Moreover, 90.5% of transcriptome reads from an unrelated study (Mesak et al. 2015), map to the reference genome. More than 30% of the unmapped reads map to the mitochondrial genome sequence available on GenBank (NC_003290, Lee et al. 2001).

**Comparison to Sister Species, *K. hermaphroditus***

To compare *K. marmoratus* to *K. hermaphroditus*, we sequenced a lab-reared *K. hermaphroditus* individual to approximately 7-fold coverage and compared the resulting sequence data to our genome assembly. For this analysis, we appended the existing mitochondrial genome to our genome assembly. The reads map to the reference assembly with a 97.8% mapping rate. Of sites that met our coverage threshold for comparison between the reference and *K. hermaphroditus*, 0.4% (1,739,544/421,508,114) were homozygous for an alternate allele and 0.1% (432,064/421,508,114) were heterozygous for an alternate allele. All other alleles were identical between *K. marmoratus* and *K. hermaphroditus*.

As *K. marmoratus* and *K. hermaphroditus* are sister species, we were interested in characterizing loci that were highly differentiated between the two species. We identified highly differentiated genes and overrepresented gene ontology (GO) terms for the comparison. We identified 23,598 genes that had at least one site where the RHL individual was homozygous for one allele and the *K. hermaphroditus* individual was homozygous for another allele. We ranked genes based on the number of differences between *K. marmoratus* and *K. hermaphroditus* and considered any genes that were in the
top 1% of distribution as highly differentiated (241 genes). Of the 241 highly differentiated genes, 218 had a significant hit in the BLAST search and, of those, 204 had a human homolog and were associated with at least one biological process GO (see Materials and Methods) term. Our resulting list included 1,091 GO terms. We tested for enrichment of biological process, molecular function, and cellular component GO terms in our highly differentiated genes using Webgestalt (Wang et al. 2013). The most highly enriched terms in biological process were nervous system development (GO:0007399), multicellular organismal process (GO:0032501), and single-multicellular organism process (GO:0044707). For molecular function, transmembrane signaling receptor activity (GO:0004888), receptor activity (GO:0004872), and signaling receptor activity (GO:0038023), were the most significant terms. Finally, terms associated with cellular component that were most enriched were membrane related (GO:0031224, GO:0016021, GO:0044425).

Sex Determination Loci

Due to the environmental sex determination (hermaphrodite vs. primary male) and sexual plasticity (hermaphrodite to secondary male transition) (fig. 1), we focused on annotating and comparing known sex determination loci from other systems. We identified major known sex-determining loci and report their gene names along with the scaffold position and linkage group, when applicable (supplementary table S8, Supplementary Material online). We identified two sox9 genes, whose homologs have been identified in zebrafish and appear to have arisen during the teleost whole-genome duplication event (Chiang et al. 2001). None of the sex determining loci was in the set of highly differentiated loci, which was expected because both species are hermaphrodites. We analyzed the sex determining genes for elevated rates of nonsynonymous to synonymous substitutions (dN/dS) using both a sites model (Yang 2007) and a branch-sites model (Yang and dos Reis 2011) with the branch leading to K. marmoratus and K. hermaphroditus as the foreground branch. None of the loci had significant evidence for positive selection using either of these approaches with P < 0.01. Some of the genes that have been implicated in sex determination have complex roles in development and we speculate that gene expression changes rather than structural changes may be driving differences in these species.

Comparison to Other Fish Species

We compared our protein-coding gene annotation set with annotated proteins from zebrafish (Danio rerio) (Howe et al. 2013), stickleback (Gasterosteus aculeatus) (Jones et al. 2012), medaka (Oryzias latipes) (Kasahara et al. 2007), fugu (Takifugu rubripes) (Aparicio et al. 2002), Amazon molly (Poecilia formosa) (W. Warren and The Genome Institute, Washington University School of Medicine, GenBank Assembly ID GCA_000485575.1), mummichog (Fundulus heteroclitus) (GenBank Assembly ID GCA_000826765.1), and turquoise killfish (Nothobranchius furzeri) (Dario et al. 2015) (fig. 2). In a comparison among the eight species, 4,827 genes were unique to mangrove rivulus. It is important to note that defining unique genes is a difficult endeavor, especially in teleost fish, and these are preliminary results. Of the 4,827 genes, 2,465 of them (~51%) have a BLAST hit in the SwissProt database and human orthologs exist for 2,213 of them (~90%) and can be used for GO enrichment analyses. The enriched GO terms include many terms relating to the nervous system and ion channels.

**FIG. 2.**—Cladogram to represent the relationship among K. marmoratus, K. hermaphroditus and several well-known fish species. The origin of hermaphroditism is noted on the cladogram.
Materials and Methods

Sample Collection and DNA/RNA Extraction

The *K. marmoratus* genome was sequenced and assembled using DNA from a lineage RHL derived from San Salvador, Bahamas; the progenitor was caught in the field by Dr. Scott Taylor in 1997 and the animals used here were 10–11 generations removed from the field. DNA was extracted from muscle tissue from multiple *K. marmoratus* RHL specimens from the Earley laboratory. Genetic homogeneity of the RHL lineage was verified by genotyping the specimens at 32 microsatellite loci developed for *K. marmoratus* (Mackiewicz, Tatarenkov, Perry, et al. 2006). For each locus, one polymerase chain reaction (PCR) primer in each pair was labeled with a fluorescent dye (HEX, 6-FAM, or NED) and DNA was amplified in several multiplex reactions, as described previously (Tatarenkov et al. 2012). Multiplexed PCR products were diluted 10–20 fold and 1 µl of the diluted product was mixed in several multiplex reactions, as described previously (Tatarenkov et al. 2012). Multiplexed PCR products were denatured for 4 min at 95 °C, and electrophoresed on an Agilent 3100 using 50 cm capillaries filled with Pop4 (Applied Biosystems). Alleles were scored using Genemapper 4.0 (Applied Biosystems). All specimens were fully homozygous, genetically identical, and matched microsatellite profile of the RHL lineage described in Tatarenkov et al. (2010). DNA was extracted using the Gentra Puregene Tissue Kit (Qiagen) according to the manufacturer’s instructions. Purified DNA was sheared using a Covaris E220 (Woburn, MA) (duty cycle 10, intensity 5, cycles/burst 200, time 180 s) to approximately 400 bp. Sheared genomic DNA was gel purified and used as input for sequencing library preparation. Sheared genomic DNA was end repaired with NEBNext end-repair kit (E6050L), and A-tailing was accomplished using Taq polymerase. Ultrapure ligase (L603-HC-L) from Enzymatics and iProof (BioRad) were used for ligation of the Illumina paired-end adapters (PE-102-1003) and amplification, respectively. Agencourt Ampure XP beads were used for clean-up at each step. Three libraries of the RHL lineage were sequenced using Illumina HiSeq2000 sequencing technology with 101 bp paired-end reads on an Illumina HiSeq2000 with 2 × 101 bp paired-end reads at the Stanford Center for Genomics and Personalized Medicine (Raw reads have been deposited under NCBI BioProject PRJNA290522).

Genome Size Estimation

Genome size was estimated from the raw reads using the k-mer frequency distribution for k = 17 and k = 23, as in Li et al. (2010). K-mer frequencies were counted using JELLYFISH (Marcais and Kingsford 2011).

Assembly Strategy

De novo genome assembly was accomplished using a multi-step process. Assembly of the genome using the short read Illumina data included error correction, preprocessing, and assembly with the String Graph Assembler (SGA) version 0.10.1 (Simpson and Durbin 2012). Scaffolding with the mate pair data was accomplished using SSPACE Basic v2.0 (Boetzer et al. 2011). The final draft assembly contains scaffolds at least 500 bp. The final assembly is deposited under NCBI BioProject PRJNA290522. Genome assembly statistics were computed using a Perl script, provided by Assemblathon 2 (Bradnam et al. 2013), and CEGMA v.2.5 (Parra et al. 2007). CEGMA required dependencies included geneid v.1.4.4 (Blanco et al. 2002), geneewe v.2.4.1 (Birney et al. 2004), and HHMER 3.1b2 (Mistry et al. 2013). CEGMA was used to determine the percent of a defined core set of 248 highly conserved eukaryotic genes is present in the assembly. BUSCO (version 1.1b1) was also used to assess assembly completeness, using the lineage vertebrata. BUSCO dependencies included HHMER 3.1b2 (Mistry et al. 2013) and augustus2.5.5 (Stanke et al. 2008).
To perform higher-order scaffolding, the linkage map from (Kanamori et al. 2016) was used. Markers were mapped in fasta format to the genome assembly using Burrows–Wheeler Aligner (BWA; Li and Durbin 2009). Information regarding the scaffold and the mapped markers was parsed from alignment file. Markers with multiple mappings were discarded. Information about orientation is not available for the linkage map, therefore the mapping of scaffolds to linkage groups is included in supplementary table S2, Supplementary Material online.

Annotation Strategy

The MAKER2 annotation pipeline (Holt and Yandell 2011) was utilized for gene annotation. Homo sapiens UniProt (UniProt Consortium 2012) and Austrofundulus limnaeus Annotation Release 100 (GenBank GCF_001266775.1) protein databases were used for protein homology data. RepeatMasker was used to mask low complexity regions in the genome assembly (Smit et al. 2013–2015; supplementary table S9, Supplementary Material online). Filtered RNA-sequencing data was mapped to the genome using Bowtie (Langmead et al. 2009), assembled into transcriptomes using Cufflinks (Roberts et al. 2011) and used as EST evidence in MAKER2. Ab initio gene prediction was accomplished using two rounds of training with SNAP (Korf 2004). Annotation of the predicted protein coding genes was accomplished with homology searching using BLASTP (Altschul et al. 1990) with a threshold cutoff e-value of 10^-10 against the Swiss-Prot database (Boeckmann et al. 2003). Ribosomal RNAs and other noncoding RNAs were annotated using Rfam (Navrocki et al. 2015); tRNAs were additional annotated using tRNAscan-SE 1.3.1 (options -H and -y) (Lowe and Eddy 1997) and Aragorn 1.2.34 (option -w -t -i116 -l -d) (Laslett and Canback 2004).

Comparison to Other Data Sets

Transcriptome reads from Mesak et al. (2015) were downloaded from the Short Read Archive (SRR2001227) and mapped to the reference genome using BWA (Li and Durbin 2009). Raw reads generated in this study, from RHL and K. hermaphroditus, were separately mapped to the reference genome using BWA (Li and Durbin 2009). File manipulation and summary statistics were generated using SAMtools (Li et al. 2009) and BamTools (Barnett et al. 2011). The mitochondrial reference sequence (Genbank, NC_003290) was included for the K. hermaphroditus mapping. Genotypes were called separately for each individual using the UnifiedGenotyper tool in the Genome Analysis Toolkit (GATK) (v. 3.5) with the EMIT_ALL_CONFIDENT_SITES option. (McKenna et al. 2010) Individual vcf files were then filtered following the GATK recommended best practices (DePristo et al. 2011; Van der Auwera et al. 2013). Vcf files were filtered using vcftools (v. 0.1.12b) such that only sites with at least 6x coverage (~minDP 6) were kept for each individual (Danecek et al. 2011). The individual vcf files were combined using the CombineVariants tool in GATK (McKenna et al. 2010) and only sites with genotypes in both individuals were retained with –max-missing 1.0 in vcftools (Danecek et al. 2011). Sites where the two individuals were homozygous for different alleles were pulled out for further analysis. We identified which of these sites were located in genes using the intersectbed utility in BEDtools (v. 2.17.0) (Quinlan and Hall 2010).

Sex Determination Loci

To identify genes implicated in sex determination processes, we queried our annotation set for the orthologous loci and report additional details about linkage group and relative position across contigs.

GO Enrichment

To characterize putatively differentiated genes, we used bioDBnet (Mudunuri et al. 2009) to identify biological process GO terms that were associated with the highly differentiated genes. We also tested for enrichment of biological process, molecular function, and cellular component GO terms in our highly differentiated genes. bioDBnet (Mudunuri et al. 2009) was used to convert gene names into human Entrez IDs. We used Webgestalt (Wang et al. 2013) to test for enrichment of GO terms compared to a reference set composed of human orthologs of the annotated genes in the K. marmoratus genome. Significantly overrepresented GO terms were identified at a false discovery rate of 0.01 (Benjamini and Hochberg 1995). At least two genes had to be associated with a GO term for it to be considered enriched.

To determine sets of orthologous genes, we compared protein data sets among closely related fish species. We included protein data from D. rerio (43,153), G. aculeatus (27,576), O. latipes (24,674), T. rubripes (47,841), P. formosa (30,898), F. heteroclitus (33,705), and N. furzeri (30,028) in our comparative analysis. Clusters of orthologous genes were identified using OrthoFinder (Emms and Kelly 2015) with MCL (Enright et al. 2002).

Supplementary Material

Supplementary tables S1–S9 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org).

Acknowledgments

This work was in part supported by a National Institutes of Health National Research Service Award postdoctoral fellowship [GM087069 to J.L.K.]. The authors thank Craig Cummings, Elizabeth Levandowsky, Minita Shah from Life Technologies for their assistance in library preparation and
sequencing. The authors would like to thank Jared Simpson for useful feedback on genome assembly with SGA.

**Literature Cited**


Costa WJEM. 2006. Redescription of *Kryptolebias ocellatus* (Hensel) and a new name for a hermaphroditic species of *Kryptolebias* from south-eastern Brazil. Ichthyol Biol. 15:51–60.


Lewsey DR, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11–16.


