Developmental Biology 452 (2019) 8–20

Contents lists available at ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/locate/developmentalbiology

A chromosome-scale genome assembly and dense genetic map for *Xenopus tropicalis*

Therese Mitros [a,1,2], Jessica B. Lyons [a,1,2], Adam M. Session [b,1,2], Jerry Jenkins [b,c], Shengquiang Shu [b], Taejoon Kwon [d], Maura Lane [e], Connie Ng [a], Timothy C. Grammer [a], Mustafa K. Khokha [e], Jane Grimwood [b,c], Jeremy Schmutz [b,c], Richard M. Harland [a,**], Daniel S. Rokhsar [a,b,f,*]

[a] University of California, Berkeley, Department of Molecular and Cell Biology, Life Sciences Addition, Berkeley, CA 94720-3200, USA
[b] Joint Genome Institute, 2800 Mitchell Dr # 100, Walnut Creek, CA 94598, USA
[c] HudsonAlpha Institute of Biotechnology, 601 Genome Way, Huntsville, AL 35806, USA
[d] Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea
[e] Pediatric Genomics Discovery Program, Department of Pediatrics and Genetics, Yale University School of Medicine, FMP 410, 333 Cedar St./LCI 305, New Haven, CT 06520, USA
[f] Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 9040495, Japan

ARTICLE INFO

Keywords:
Genome assembly
Genetic mapping
Comparative genomics
Pigmentation
Sex determination
Gene expression analysis

ABSTRACT

The Western clawed frog *Xenopus tropicalis* is a diploid model system for both frog genetics and developmental biology, complementary to the paleotetraploid *X. laevis*. Here we report a chromosome-scale assembly of the *X. tropicalis* genome, improving the previously published draft genome assembly through the use of new assembly algorithms, additional sequence data, and the addition of a dense genetic map. The improved genome enables the mapping of specific traits (*e.g.*, the sex locus or Mendelian mutants) and the characterization of chromosome-scale synteny with other tetrapods. We also report an improved annotation of the genome that integrates deep transcriptome sequence from diverse tissues and stages. The exon-intron structures of these genes are highly conserved relative to both *X. laevis* and human, as are chromosomal linkages ("synteny") and local gene order. A network analysis of developmental gene expression will aid future studies.

1. Introduction

The draft genome of the Western clawed frog *Xenopus tropicalis*, encoding over 20,000 genes (Hellsten et al., 2010), has been an important resource for cell and developmental biology and for comparative genomics (Harland and Grainger, 2011; Khokha, 2012). Analysis of the sequence has led to insights into vertebrate genome evolution (Hellsten et al., 2010), and served as a reference enabling analysis of gene expression in development (Collart et al., 2014; Owens et al., 2016; Paranjpe et al., 2013; Tan et al., 2013; Yanai et al., 2011) and regeneration (Love et al., 2011), DNA methylation and chromatin remodeling (Bogdanović et al., 2012), genetic mapping (Wells et al., 2011), rapid evolution of sex determination (Bewick et al., 2011; Roco et al., 2015), and the design of genome editing reagents (Bhattacharya et al., 2015; Blitz et al., 2013; Guo et al., 2014; Ishibashi et al., 2012; Nakayama et al., 2013; Young et al., 2011). As a diploid counterpart to the paleotetraploid African clawed frog *Xenopus laevis*, *X. tropicalis* sequence has also served as a point of reference for evolutionary and developmental comparisons between the two frogs and exploration of impact of tetraploidy (Chain and Evans, 2006; Hellsten et al., 2007;

* Corresponding author. Rokhsar Lab, University of California, Berkeley, 142 Life Sciences Addition # 3200, Berkeley, CA 94720-3200, USA.
** Corresponding author. Harland Lab, University of California, Berkeley, 142 Life Sciences Addition # 3200, Berkeley, CA 94720-3200, USA.
E-mail addresses: tmitros@berkeley.edu (T. Mitros), jblyons@berkeley.edu (J.B. Lyons), amsession@lbl.gov (A.M. Session), jjenkins@hudsonalpha.org (J. Jenkins), sqshu@lbl.gov (S. Shu), tkwon@unist.ac.kr (T. Kwon), maura.lane@yale.edu (M. Lane), connie.ng@berkeley.edu (C. Ng), timgrammer@gmail.com (T.C. Grammer), mustafa.khokha@yale.edu (M.K. Khokha), jgrimwood@hudsonalpha.org (J. Grimwood), jschmutz@hudsonalpha.org (J. Schmutz), harland@berkeley.edu (R.M. Harland), dsrokhsar@gmail.com (D.S. Rokhsar).
[1] co-first authors.
[2] These authors have contributed equally to this work.

https://doi.org/10.1016/j.ydbio.2019.03.015
Received 8 November 2018; Received in revised form 12 March 2019; Accepted 22 March 2019
Available online 10 April 2019

Sémon and Wolfe, 2008; Session et al., 2016; Yanai et al., 2011).

The utility of the 2010 draft genome of *X. tropicalis* (known as v4.2) belies several of its limitations. Most notably, the draft genome is not organized into chromosomes, but comprises scaffolds of typical size ("N50 length") 1.57 Mb. (A scaffold is a reconstructed stretch of sequence that may contain sequence gaps, indicated by Ns in the sequence.) These scaffolds typically contain dozens of genes, but are limiting for applications that rely on genetic mapping or analysis of conserved synteny (*i.e.*, physical linkage). The typical (N50) length of contigs (*i.e.*, gap-free regions) is 17 kilobases. This is long enough that, with sequence gaps often confined to introns or repetitive intergenic regions, the exon-intron structure of most protein-coding genes is properly represented, but gaps still can occur in inconvenient positions relative to genes of interest. Finally, although misjoins (assembly errors) were rare, several have been noted (Roco et al., 2015; Spirhanzlova et al., 2017).

To improve the utility of the *X. tropicalis* genome sequence and annotation, and enable chromosome-scale analyses, we generated an improved version, v9, that reconstructs the genome as complete chromosomes with substantially longer contigs. Key methodological improvements include the use of newer algorithms (Jaffe et al., 2003) to assemble the original Sanger shotgun dataset (Hellsten et al., 2010); the application of new methods for characterizing long-range physical linkage (Putnam et al., 2016); and the construction of a new dense genetic map using genotyping-by-sequencing (Elshire et al., 2011; ICGMC, 2015) to anchor the assembly to linkage groups. The chromosome-scale assembly is augmented by an updated annotation of repetitive elements and protein-coding genes, taking advantage of rich transcriptome sequences from *X. tropicalis* and the related *X. laevis* across developmental time and from diverse adult tissues and organs.

With the revised assembly and annotation in hand, we report several updated and novel genomic and genetic analyses. We find extensive long-range synteny, often but not always with considerable colinearity, between the *X. tropicalis* and human genomes. Orthologous regions corresponding to several human chromosomes are wholly contained within *X. tropicalis* chromosomes. We show that gene structures of *X. tropicalis* and human are highly conserved, both in exon-intron structure and coding length. We find that the rate of sequence change on *X. tropicalis* chromosome 10 is increased relative to other chromosomes. The dense genetic map is used to identify the chromosomal locus associated with a pigmentation mutant, *gray*, and to find a locus associated with sex. The existence of a recessive lethal chromosome in our partially inbred Ivory Coast strain is also inferred. Finally, we integrate previously generated transcriptome datasets to update the atlas of developmental gene expression, inferring 19 co-expression groups by weighted gene coexpression network analysis (WGCNA).

2. Materials and methods

2.1. Shotgun assembly

2.1.1. Sequencing summary

The sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at the Department of Energy Joint Genome Institute in Walnut Creek, California as previously reported (Hellsten et al., 2010). BAC end sequence was collected using standard protocols at the HudsonAlpha Institute in Huntsville Alabama. The sequencing consisted of seven libraries of 3 kb pairs (3.15×), six libraries of 8 kb pairs (3.30×), five fosmid libraries (0.52×), and three BAC libraries (0.01×) on the Sanger platform for a total of 7.47× Sanger based coverage. More detail can be found in Hellsten et al. (2010).

2.1.2. Version 7 genome assembly

A total of 23, 213, 531 sequence reads (7.47× assembled sequence coverage) were assembled using our modified version of Arachne v.20071016HA (Jaffe et al., 2003) with parameters correct1_passes = 0 maxcliq1 = 100 maxcliq2 = 50 num_cpus = 20 BINGE_AND_PURGE = True remove_duplicate_reads = True. This produced 11,471 scaffolds (59,673 contigs) totaling 1387.3 Mb of sequence, with a scaffold N50 of 1.7 Mb, 1315 scaffolds larger than 100 kb (1314.5 Mb, 90%). Scaffolds were screened against bacterial proteins, organelle sequences, and GenBank nr, and removed if found to be a contaminant. Gaps account for 4.3% of the resulting scaffold length.

2.1.3. Completeness with respect to known genes

Completeness of the euchromatic portion of the genome assembly was assessed using 9630 full length cDNAs (FLcDNAs, from RefSeq v71). The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The cDNAs were aligned to the assembly using BLAST (Parameters: 1e-10 and alignments ≥90% base pair identity and ≥50% coverage) were retained. The screened alignments indicate that 9455 (98.1%) of the FLcDNAs aligned to the assembly. The cDNAs that failed to align were checked against the NCBI nucleotide repository (nr), using a BLASTX of all six open-reading frames with an e-value cutoff of 1e-10. Surprisingly 160/175 aligned to *Xenopus* species, indicating that they are either missing or fragmented in our assembly (many of these have shorter alignments to the assembly that were not counted, so we favor the fragmentation hypothesis). Six cDNAs had no alignment to nr, while nine more showed strong similarity (96% + peptide identity) amongst themselves, suggesting that 15 of these nominally *Xenopus* sequences may be contaminants. We also aligned the v7 *X. tropicalis* CDS to the v9 genome to ensure completeness relative to the previous assembly. Using a length cutoff of 90%, 40,865/43,445 (94%) of the v7 transcripts are present in the genome at full length. The remainder (i.e., present in v7 but absent in v9) have short alignments that may be interrupted by sequence gaps or alignment gaps too long to be considered introns by automated analyses (intron cutoff of 50 kb). These are generally spurious predictions of the earlier v7 assembly/annotation, and have been provided to Xenbase for manual curation.

2.2. Construction of a genetic map

2.2.1. Mapping population

An F2 mapping population was developed from two inbred grandparents: a ninth generation inbred green Nigerian (N) male and a *gray* ICB female (Igawa et al., 2015; Khokha et al., 2009). 192 F2s were produced from matings between two F1 mating couples. Of these, 96 of the DNAs (37 from one clutch and 59 from the other) were prepared from adults for which we had phenotype information (sex and color; see Fig. S4).

Grandparental and F1 DNAs were isolated from toe clips lysed for 16–120 h at 55 °C in lysis buffer (50 mM Tris pH 7.5, 62.5 mM NaCl, 5 mM EDTA, 0.5% SDS, 5% Chelex resin [Bio-Rad]) supplemented with Proteinase K to 250–500 μg/mL, then mixed and spun down. The supernatant was removed and DNA was extracted using Phenol/Chloroform and resuspended in TE or water.

The 96 F2 tadpole DNAs were isolated using the same method as the grandparental and F1 toe clips, and resuspended in 10 mM Tris pH 8.5.

The 96 adult F2 DNAs were isolated from erythrocytes using a protocol modified from Sive et al. (2000). Importantly, we found that due to nuclease present in the serum of *X. tropicalis*, the standard method leads to degraded DNA, so we use a cell lysis buffer containing SDS and EDTA. Briefly, blood was collected into 0.9× SSC from a frog anesthetized with 0.05% benzocaine and without the use of heparin. After collection, the cells were dispersed by inverting the tube, and cells were gently pelleted at 4 °C, 240 *g*, for 3 min. Nearly all the supernatant was decanted, the cells were resuspended in the remaining supernatant by gentle vortexing, and lysed overnight at 55 °C in 1% SDS, 20 mM EDTA, 100 mM NaCl, 20 mM Tris pH 7.5–8.0 with 200 μg/mL Proteinase K. To precipitate DNA, 0.25 vol 10 M ammonium acetate was added, the solution mixed thoroughly, then 0.6 vol isopropanol added and the solution again mixed thoroughly. DNA was spooled with a glass rod into a clean tube of 70%

EtOH for washing, then spooled again into a clean tube, allowed to dry, and resuspended in TE pH 8.0 for at least a day. Most of these DNAs were further diluted 1:5 in water before use. All DNAs were stored at 4 °C.

2.2.2. Genotyping-by-sequencing library preparation and sequencing

Single nucleotide polymorphisms (SNPs) were genotyped using genotyping-by-sequencing (GBS), a reduced representation Illumina sequencing approach (Elshire et al., 2011). The adapter and library preparation approach were as described by the International Cassava Genetic Map Consortium (ICGMC, 2015). Briefly, genomic DNA was digested with the restriction enzyme *Nsp*I, which has recognition site RCATGY, where R represents A or G and Y represents C or T. Digests were performed at 37 °C for 2–9 h or overnight, with 1–2 U *Nsp*I (New England Biolabs) per 100 ng DNA in NEB buffer #2 with BSA, followed by a heat-inactivation step of 65 °C for 20 min.

"Forward" adapter oligos had sequences

5′ ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxCATG 3′ and "reverse" oligos the sequence 5′ Phos-yyyyAGATCGGAAGAGCGGTT CAGCAGGAATGCCGAG 3′, where xxxx is the 4–8 bp barcode, yyyy its reverse complement, and CATG the overhang facilitating ligation to *Nsp*I-digested DNA. Adapter oligos were made and HPSF purified by Operon. Forty-eight distinct barcodes were chosen from among those described in Elshire et al. (2011). One barcode was modified to remove an *Nsp*I recognition site (see Note S1). Ligation of DNA to pre-annealed adapters was performed at 22 °C for 1–2 h in 1× NEB T4 ligase buffer with 0.3–1 μL NEB T4 ligase per 100 ng genomic DNA and 0.1–0.5 μM pre-annealed adapter, followed by a heat-inactivation step of 65 °C for 30 min.

Pooled ligated DNA was size selected on an agarose gel, and ligated fragments ~450–650 bp long (grandparents and one F1 mating pair) or ~500–600 bp long (one F1 mating pair and all F2s) were excised and PCR amplified (98 °C for 30 s; 15 cycles of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s; 72 °C for 5 min) to produce libraries suitable for Illumina sequencing. PCR primer sequences were 5′ CAAGCAGAAGACGGCA-TACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT 3′, and 5′ CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCT GAACCGCTCTTCCGATCT 3′. Sequencing was performed on Illumina HiSeq 2000 with single- or paired-end 100 bp read lengths.

2.2.3. Construction of genetic map with JoinMap

2.2.3.1. Genotyping of GBS reads. Barcodes were clipped and Illumina GBS reads were sorted by barcode using pairedBarcode.pl (https://bitbucket.org/mitros/f2mapping/src/master/). Reads with barcodes recovered on both pairs were aligned to v7 scaffolds using bwa aln -q 15 (Li and Durbin, 2009) and further processed to an mpileup using samtools with a mapping quality cutoff of 25 (Li et al., 2009). To identify single nucleotide polymorphisms suitable for F2 mapping, we required aligned positions that were (a) apparently homozygous in each grandparent, but with different alleles and (b) not homozygous in any F1. Bases were filtered for minimum base call quality of 25 and mapping quality 25. A depth of 10 was required in each grandparent (F0). In the F1 and F2 offspring, allele counts were made on the grandparentally defined SNPs. Since the median depth at such sites in the F2 progeny is 1, individual sites could not be reliably genotyped, and genotype calls (AA, AB, or BB) were made in 400 kb bins ("marker bins"). Genotype calls were made for marker bins with a minimum total depth of 10 and a minor allele frequency between 0.3 and 0.7 for heterozygotes. Marker bins with segregation distortion worse than 1e-3 were not used in the construction of the genetic map. Scaffolds shorter than 400 kb that met these criteria (total depth, minor allele frequency, and segregation distortion) were treated as their own marker bins.

2.2.3.2. Construction of genetic linkage map. JoinMap was used to create a genetic map from the 2219 markers meeting the above conditions. Within JoinMap we used regression mapping with Kosambi mapping function for an F2 population. Ten linkage groups formed at LOD8 or better. An additional map, including distorted markers, was made in an attempt to place as many scaffolds as possible into the linkage groups. This accounts for 2682 400 kb marker bins.

2.2.4. Integration of genetic map and additional BAC end sequences

The combination of an NGS-based genetic map (43,052 SNPs), a BAC library provided by Nicolas Pollet's group at the University of Evry-Val-d'Essonne (Spirhanzlova et al., 2017), and *X. laevis* synteny (Session et al., 2016) was used to identify 55 misjoins in the overall assembly. Misjoins were identified as linkage group discontiguity coincident with an area of low BAC/fosmid coverage. A total of 996 scaffolds were oriented, ordered and joined using 986 joins to form the final assembly containing 10 pseudomolecule chromosomes. Each chromosome join is padded with 10,000 Ns. The intermediate v8 assembly (used as input to the chromatin-based scaffolding [**Methods 2.3**] and genetic map integration [**Methods 2.2.3**]) contains 8128 scaffolds (55,914 contigs) that cover 1369.9 Mb (including gaps) of the genome with a contig L50 of 71.9 kb and a scaffold L50 of 129.5 Mb.

2.3. Refinement of chromosomes via in vitro chromatin-based linkage

DNA was isolated from a Nigerian strain inbred F13 male named George, essentially as described above in "Mapping Population," and phenol extracted.

A Chicago library was prepared as described previously (Putnam et al., 2016). Briefly, 5.5 μg of high molecular weight genomic DNA was reconstituted into chromatin *in vitro*, and fixed with formaldehyde. Fixed chromatin was then digested with *Mbo*I, the 5′ overhangs were filled in with biotinylated and thiolated nucleotides, and then free blunt ends were ligated. After ligation, formaldehyde cross links were reversed and the DNA was purified to remove biotin not internal to ligated fragments. The DNA was then pulled down with streptavidin beads to enrich for biotin-containing fragments and sequencing libraries were generated using established protocols (Meyer and Kircher, 2010).

Re-scaffolding of a chromosome-scale assembly with HiRISE: *X. tropicalis* draft chromosome sequences in FASTA format (version 8, described above), and Chicago library sequencing reads in FASTQ format were used as input data for HiRISE, a software pipeline designed specifically for using Chicago library sequence data to correct misassemblies and scaffold genomes (Putnam et al., 2016). At the core of the HiRISE pipeline is a likelihood model that predicts the unique distribution of Chicago read pair separations.

Shotgun and Chicago library sequences were aligned to the draft input assembly using the SNAP read mapper (http://snap.cs.berkeley.edu) and marked as PCR duplicates using Novosort (http://www.novocraft.com/products/novosort/). Chicago read pairs were aligned with a modified SNAP such that reads were not penalized for having unusual orientation, insert size, or for mapping to different scaffolds. Additionally, when a read contained the restriction site junction "GATCGATC," the sequence after the first "GATC" was ignored for mapping purposes.

245.9 million of the 316.5 million shotgun reads were aligned with greater than 10 map quality and were not marked as PCR duplicates. 76.9 million of the 154.7 million Chicago read pairs were aligned such that both reads had greater than 20 map quality and were not marked as duplicates. Of these high-confidence read pairs, 60.9 million spanned between 0 and 2 kb, 1.37 million spanned between 2 kb and 10 kb, and 0.96 million spanned between 10 kb and 100 kb. The Chicago reads yield 5.9×, 7.3×, and 20.7× effective physical coverage between 0 and 2 kb, 2 kb–10 kb, and 10 kb–100 kb, respectively.

The mapped shotgun reads with greater than 10 map quality were used to flag highly-repetitive regions in the assembly to exclude from HiRISE.

The high-confidence Chicago read pairs were then scored with the

HiRISE likelihood model and these scores were used to identify and break misjoins in the input assembly. HiRISE was run and produced scaffolds. The scaffolds were re-joined into a chromosome-scale assembly guided by the v8 assembly as follows: unique 101-mer sequences "audit-mers" were identified in v8. The position of each of these audit-mers in the HiRISE assembly was tabulated. The longest run of audit-mers was used to assign the scaffold to a chromosome. Within each chromosome, the starting positions and orientations of these audit-mers were then used to order and orient the HiRise scaffolds. The chromosomes of the resulting assembly were reoriented as described in **Methods 2.4** to produce the v9 assembly.

2.4. Orientation of chromosomes

Extensive karyotyping has been performed on *X. tropicalis* and *X. laevis*, documenting the extensive chromosome-scale colinearity between the genomes for chromosomes 1–8 of *X. tropicalis* and their two co-orthologous chromosomes in the allotetraploid *X. laevis* (Matsuda et al., 2015; Uno et al., 2013). The co-orthologs of chromosomes 9 and 10 of *X. tropicalis* are fused, q-to-q, in *X. laevis*. Our observations of colinearity amongst *Xenopus* chromosomes are consistent with conserved linkages between *Xenopus* and *Rana* (Rodrigues et al., 2016) and previous observations that amphibian karyotypes are remarkably stable (Voss et al., 2011). To facilitate direct comparisons among *Xenopus* genomes, we oriented the *X. tropicalis* chromosomes from p to q, and confirmed conserved p-to-q orientation with *X. laevis* for chromosomes *1–8*. We note that this orientation differs from earlier releases, but is consistent with traditional cytogenetic conventions and is preferred by the *Xenopus* community. The v8 mapped assembly, corrected by *in vitro* chromatin and reoriented to conform to p to q convention, defines the *X. tropicalis* version 9 (v9) genome.

2.5. Annotation methods summary

We annotated the protein-coding genes of the *X. tropicalis* genome using a modified version of the DOE Joint Genome Institute annotation pipeline, which integrates transcriptome data, homology, and *ab initio* methods and has been previously applied to numerous metazoan genomes (see (Simakov et al., 2013), for example). Prior to annotating genes, *X. tropicalis* genome sequences were repeat-masked by RepeatMasker using a custom *X. tropicalis* transposable element database. Details are provided below.

First, we used RepeatScout to detect all fragments of the frog genome with repetitive sequences and classify them as coding for proteins similar to catalytic cores of transposases, reverse transcriptases, and DNA polymerases representing all known classes of transposable elements (TEs) collected in Repbase. The detected DNA sequences were clustered based on their pairwise identities by using the BLASTclust algorithm from the NCBI BLAST package (the pairwise DNA identity threshold was equal to 80%). Each cluster was then treated as a candidate TE family, described by its consensus sequence.

The consensus sequences were built automatically based on multiple alignments of the cluster sequences expanded in both directions and manually modified based on structural characteristics of known TEs. We then produced a TE library by merging these consensus sequences with tetrapod TE sequences reported previously in literature and collected in Repbase. Using RepeatModeler, we identified genomic copies of TEs similar to the library sequences. These were clustered based on their pairwise DNA identities using BLASTclust. In each cluster, a consensus sequence was derived based on multiple alignment of the cluster sequences.

After refinements of these consensus sequences, the identified families of TEs were classified based on their structural hallmarks, including target site duplications, terminal repeats, encoded proteins and similarities to TEs classified previously. Identified TEs are deposited in Repbase. These repeats mask 39.7% of the assembled genome sequence.

The JGI protein-coding gene annotation pipeline utilizes transcriptome support, similarity to genes in related species, and *ab initio* methods to predict protein-coding genes. The span of gene loci was identified as overlapping regions of aligned transcriptome and homology data on the shotgun assembly:

- ESTs and cDNAs: In support of gene annotation we aligned 35,238 *X. tropicalis* ESTs and cDNAs from NCBI to the chromosome-scale *X. tropicalis* v9 genome assembly, requiring a minimum of 98% identity and 50% coverage (*X. tropicalis* PASA).
- Homology: Peptide sequences from *X. laevis*, human, mouse, and chicken (UniProt) were used.

Briefly, gene locus spans were defined by the overlap of BLAT alignments of EST and cDNA data and BLASTX alignments of both homology and RNA-seq transcript assembly peptides, with an added extension of 500 bp at both ends of each locus. At each such locus, *X. laevis*, human, mouse, and chicken peptides, and RNA-seq transcript assembly ORFs were used as protein templates for both GenomeScan and Fgenesh + gene predictions. These predictions were then merged with EST and cDNA data using PASA, which corrects exon-intron boundaries and adds untranslated regions (UTRs) based on transcriptome alignments. The longest ORF prediction at each locus was retained, along with alternate splice isoforms accepted if supported by PASA. This defined the JGIv9.0 annotation (note that the 4.2 assembly has the same nucleotide sequence as the 9.0 chromosome scale assembly, and differs only by the organization of the sequence into chromosomes).

To assess functional categories we used several informatic methods to assign putative functions to genes.

- PfamScan (Pfam v27.0) was used to assign Pfam domains to gene loci (Finn et al., 2015).
- InterPro2GO was used to map pfam assignments to GO terms (Finn et al., 2016).
- *X. tropicalis* KEGG assignments were extracted from the KEGG database via the REST API.

2.6. Conservation of gene structure and synteny

The large scale synteny between *X. tropicalis* and *X. laevis* has been shown previously (Session et al., 2016; Uno et al., 2013). The (co)-orthologs between *Xenopus* species were obtained from Session et al. (2016). For comparison to other tetrapods, we extracted proteomes for human, mouse, dog, chicken, and lizard from Ensembl (v83) to compare to *X. tropicalis*. We aligned proteins via BLASTP in the BLAST + package, using an e-value cutoff of 1E-10 (Camacho et al., 2009), requiring a minimum 30% peptide identity across 50% of the length. Mutual best hits between *X. tropicalis* and human could be assigned for 14,864 proteins from the v9.0 annotation based on BLAST bit score. Fig. 3 shows that this method identifies frog and human orthologs that are similar in length and exon number. Peptide lengths and exon numbers were extracted from the Ensembl v83 annotation for human. Similar analysis was done to compare frog to chicken, mouse, and lizard to similar results (data not shown). For bar and dot plots in Fig. 4, at least 3 genes in a row with alignment between *X. tropicalis* and human must be found for a stretch of synteny or dot to be drawn. Triplets of orthologs were identified by MCScanX (Wang et al., 2012).

2.7. Gray phenotype

2.7.1. Animal husbandry and matings

X. tropicalis tadpoles and frogs were raised in a recirculating water facility according to established protocols (Grammer et al., 2005). Embryos were obtained by inducing natural matings of adult frogs with hCG (human chorionic gonadotropin).

2.7.2. Scoring adult wild type and gray frogs

Adult frogs were scored as wild type or *gray* after Stage 65, at which time the difference between wild type and gray coloration was obvious. Any un-metamorphosed tadpoles were eliminated (usually 1–3 tadpoles per clutch). Froglets were housed as mixed wild type/*gray* populations, and scored again 1–2 weeks later to confirm results.

2.7.3. Tissue preparation for electron microscopy

Freshly excised sections of dorsal skin were fixed in 2% glutaraldehyde (1.5% paraformaldehyde in 0.1 M cacodylate buffer [pH 7.2]) for 15 min. 1.5 mm diameter samples were excised with a 1.5 mm biopsy punch (Acuderm) and incubated overnight in a fresh solution of the same fixative. The samples were washed twice for 10 min in 0.1 M Na-cacodylate buffer (pH 7.2) + 50 mM glycine, then washed for 10 min in 0.1 M Na-cacodylate buffer (pH 7.2). The skin was post-fixed in 1% OsO4 in 0.1 M Na-cacodylate buffer (pH 7.2) for 30 min (in the dark, 4 °C).

The tissues were washed three times for 10 min in 0.1 M Na-cacodylate buffer (pH 7.2), followed by two 10 min rinses in distilled water. The biopsies were stained overnight with 0.5% aqueous uranyl acetate (in the dark, 4 °C), and then washed twice for 10 min in distilled water. The samples were dehydrated in acetone, and then embedded in resin. After curing, the resin-embedded skin samples were cut into 60–70 nm thin sections using an ultramicrotome. These sections were collected on formvar and carbon coated grids and contrast stained with 0.5% uranyl acetate. The sections were imaged and photographs taken with the FEI Tecnai 12,120 KV transmission electron microscope.

2.8. Genetic mapping of sex and gray

Quantitative trait locus (QTL) mapping was performed based on 95 post-metamorphic juvenile frogs from our F2 mapping population that were phenotyped for skin color (see **Methods 2.7**), and sexed after dissection (Hayes et al., 2002). The scanone algorithm of R/qtl (Broman et al., 2003) version 1.42–8 used the EM method with a binary model for *gray* and sex QTL analysis.

More information regarding the genetic map, *gray* mapping/candidates, and the sex locus, may be found in Notes S1, S3, and S4, respectively.

2.9. Analysis of gene expression data

We analyzed gene expression of the RNAseq data described in Note S4 for a developmental time series and selected adult tissues. After filtering (1) reads with no call ('N') and (2) reads with low complexity (not having all of 'A', 'C', 'G, and 'T') from raw Tan et al. (2013) RNA-seq reads, we mapped them to primary transcript sequences using bwa mem (version 0.7.10) with paired-end option (Li, 2013). We quantified the expression of each transcript using "Transcripts Per Million" (TPM) values estimated by RSEM (version 1.2.19) (Li and Dewey, 2011). All scripts used in this analysis are available at https://github.com/taejoonlab/HTseq-toolbox/.

Out of 26,550 predicted genes, 16,747 showed any evidence of expression in the Tan et al. developmental data sets (TPM > 0). Prior to network building we reduced all TPM values less than or equal to 0.5 to zero, then log-transformed the data according to $\log_{10}$ (TPM+0.1).

Only the 10,703 genes whose maximum TPM was at least 0.5 were used in the analysis. WGCNA was performed using the R package (Langfelder and Horvath, 2008). 29/10,703 genes could not be classified into co-expression networks confidently.

Significant enrichments were assessed by a 2 × 2 chi-squared test in R. List of homeologous and singleton (co)-orthologs in *X. laevis* were obtained from (Session et al., 2016). PFAM, GO, and KEGG classifications were assigned during the annotation process by InterPro2GO (Finn et al., 2016).

3. Results and discussion

3.1. Improving the reference genome

In order to improve the quality of the *X. tropicalis* reference genome sequence at short and long scales, we:

1. reassembled the original ~7.5-fold redundant Sanger dideoxy whole genome shotgun sequence with newer shotgun assembly methods (Jaffe et al., 2003),
2. developed and integrated a 2219 marker genetic map by constructing and genotyping an F2 mapping population of 191 individuals, and
3. incorporated new long-range paired-end data based on proximity ligation of *in vitro* reconstituted chromatin (Putnam et al., 2016). The result is a chromosome-scale assembly whose long range accuracy is corroborated by multiple lines of evidence.

3.2. Revised shotgun assembly

Advances in genome assembly algorithms allowed us to reassemble the original Sanger dideoxy sequences into longer and more contiguous sequences. Using a customized version of Arachne2 (Jaffe et al., 2003), the v7 assembly was produced. Relative to the 2010 draft (Hellsten et al., 2010) the N50 contig length improved from 17 kb to 71.2 kb, and the N50 scaffold length rose from 1.6 Mb to 1.7 Mb (**Methods 2.1**). The L50 for contigs and scaffolds improved from 22,312 to 5613 and 272 to 244 respectively. The v7 assembly total length is 1449.1 Mb, and 95.8% of this assembly is in scaffolds longer than 50 kb (improved from 89% in v4.2). Gaps account for 4.3% of the scaffold length, improved from 10% in the v4.2 assembly of Hellsten et al. (2010). The genetic map was then built on this v7 assembly, as detailed below.

3.3. Dense genetic map

For linkage mapping, we developed an F2 mapping population from two inbred grandparents: a 9th generation inbred green Nigerian (N) male and an F3 inbred ICB female with characteristic axanthic *gray* phenotype (Khokha et al., 2009) (Fig. S4). Two F1 mating pairs produced 192 F2 progeny. Of these, 96 were raised to adulthood and phenotyped for color (gray or green) and sex (**Methods 2.7 & 2.8**). The remaining 96 were genotyped as tadpoles; one of these was excluded from further analysis due to insufficient depth. We genotyped these individuals using reduced-representation multiplex Illumina sequencing, often called "genotyping-by-sequencing," or "GBS" (Elshire et al., 2011). In GBS, a reduced representation of the genome is produced by end-sequencing a defined size fraction from a restriction digest, which reproducibly samples a collection of genomic loci. Samples were multiplexed at up to 48 per lane with an in-line barcoded adapter (ICGMC, 2015) (**Methods 2.2.3**).

We identified 47,106 single nucleotide fixed (*i.e.* homozygous) differences between the grandparents at GBS-sampled positions, and confirmed that these sites were heterozygous in the F1 progeny. For simplicity, we focused only on these sites which are homozygous but different in the N and ICB grandparents. Heterozygous sites in the F0 generation were not considered further. While these sites could have been used, the density of fixed differences is so high that these markers would not have improved the resolution of the map. Using the v7.1 assembled scaffolds, we defined 2219–400 kb genomic regions without segregation distortion and created a high density genetic map of 10 linkage groups with JoinMap (Van Ooijen, 2006), matching the haploid chromosome number in *X. tropicalis* (Tymowska, 1973) (Fig. 1). (Note that there is a ~2 Mb gap separating the first three segregating markers of chromosome 8 with the remainder of the markers on this chromosome, with an estimated ~40 cM gap in map position. The scarcity of markers

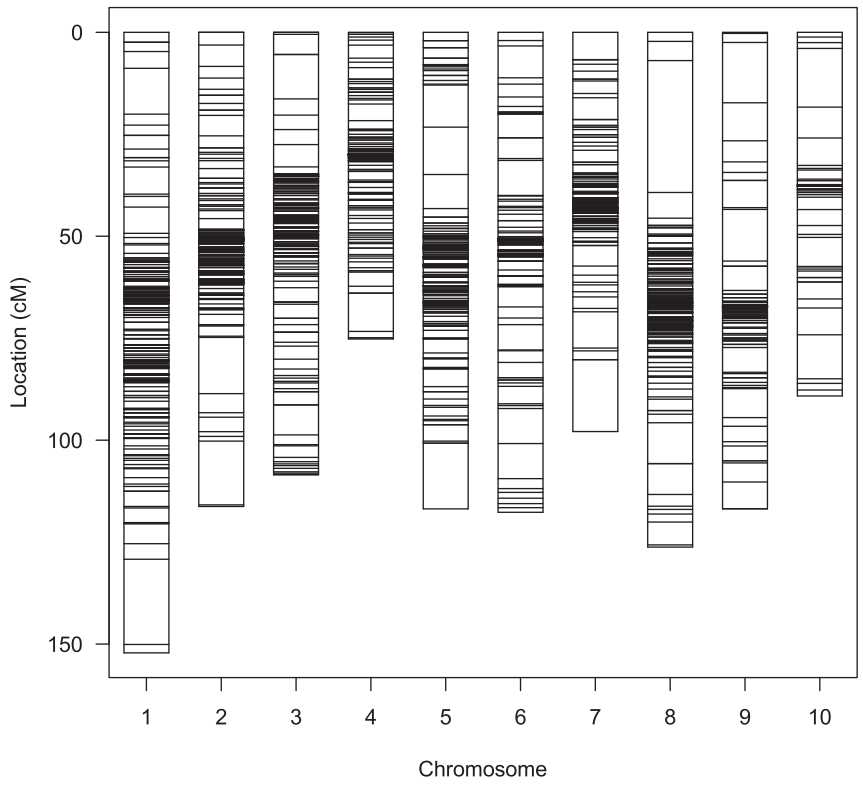


Fig. 1. *X. tropicalis* genetic map. Ten linkage groups were characterized by the analysis of markers in 2219 windows of the *X. tropicalis* genome assembly. Composite markers were formed for each window. See Supplementary File 2 for a complete list of markers.

in this region may be due to its repetitiveness and/or the acrocentric nature of chromosome 8. The lack of a continuous chain of markers in this region makes the map position estimates dependent on the genotyping of the three markers at the start of the chromosome.) The basic statistics of the map are included in Table 1.

This map was integrated with the *de novo* shotgun assembly to produce an initial chromosome-scale assembly (v8). In this process, 55 misjoins in the Arachne assembly were identified based on discrepancies with map and/or BAC end-pair sequences, including an additional BAC end set from (Spirhanzlova et al., 2017), and curated based on conserved synteny with *X. laevis*. These BAC ends were then used to make 986 additional linkages between adjacent scaffolds (see Note S2).

3.4. In vitro chromatin assembly

The resolution of a genetic map is limited by the number of crossover events in the mapping population. In particular, if no recombination occurs between two markers, then those markers are assigned the same estimated map position. Scaffolds within which no recombination occurs cannot be oriented on a genetic map, and pairs or larger groups of linked scaffolds that co-segregate in the population without recombination cannot be ordered relative to one another. This is particularly problematic in the vicinity of centromeres, which have both low recombination rates per unit physical length, and high repetitive content.

To infer fine scale order and orientation of scaffolds, we used *in vitro* chromatin conformation capture (**Methods 2.3**) (Putnam et al., 2016) that allows the inference of linkages and orientation on scales up to 1–200 kb. Purified high molecular weight DNA can be reconstituted into chromatin *in vitro* by the addition of histones and chromatin assembly factors. Sequences that are nearby in three dimensions in this compact chromatin polymer are typically nearby along the DNA sequence, but can be tens or even hundreds of kilobases apart. Restriction digestion of cross-linked chromatin followed by proximity ligation therefore provides long-range linkages that can be used for assembly validation and

Table 1
Statistics of genetic map built using v7 assembly.

F2 Individuals	191
markers	2219
Markers genotyped	88.4%
Genotypes (%):	AA:25.5 AB:50.9 BB:23.7
Centimorgans	1116.75
Linkage groups at LOD 8	10
Average inter-marker spacing	0.5 cM
Scaffolds placed in map	968
Scaffold sequence placed in map	1.2 Gb

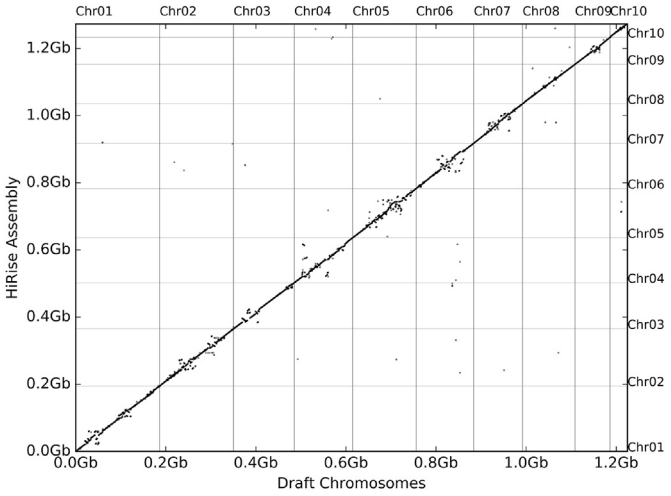


Fig. 2. Refinement of *X. tropicalis* genome assembly using HiRISE. The X-axis is the draft chromosome-scale assembly v7. The Y-axis is the post-HiRISE, rejoined, v9 chromosome-scale assembly. 101-mers that are unique in both assemblies are represented by points in the figure.

correction.

Since we already had a chromosome-scale sequence through the integration of the shotgun assembly and initial genetic map, we used this data to (1) validate our assembly, and (2) reorder and orient sequences within regions of the genetic map that showed little or no recombination in our population. The comparison between the final chromosome-scale assembly and the roughly integrated assembly is shown in Fig. 2. Notable features include (1) the general colinearity, showing that the chromatin data confirmed local linkage; (2) rearrangements around centromeres, consistent with these regions being unresolved in the genetic map due to limited recombination, and (3) one gross misjoin, identified and corrected with the Chicago data. This misjoin was at a finer scale than revealed by the relatively coarse genetic map.

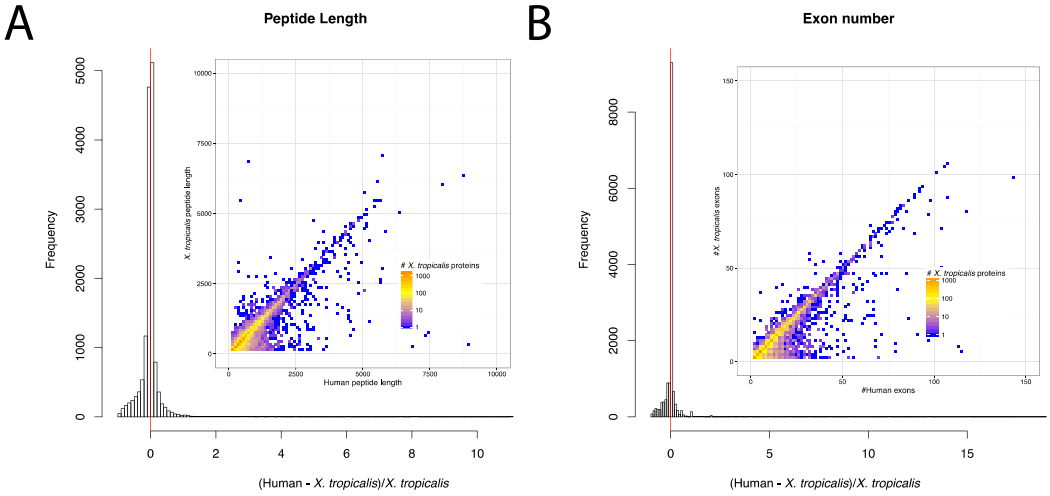
The final assembly has total length 1441.7 Mb (retaining only scaffolds longer than 1 kb), and 1384.7 Mb (96%) of this assembly is in scaffolds longer than 100 kb. 1272.9 Mb (88.3%) are in the 10 *X. tropicalis* chromosomes. 71.68 Mb (4.9%) of the assembly consists of gaps.

3.5. Genome annotation

We annotated protein-coding genes of the v9 assembly using the same pipeline we applied to *X. laevis* (Session et al., 2016). This annotation incorporated additional *X. tropicalis* transcriptomes (Love et al., 2011; Tan et al., 2013), and additional ESTs and cDNAs from NCBI (O'Leary et al., 2016), homology modeling with *X. laevis* predicted peptides, and peptides from other tetrapods. To exclude transposons and other repetitive sequences from the protein-coding gene annotation, we identified repetitive elements using *de novo* RepeatModeler (Smit and Hubley, 2008-2015) and added these elements to those annotated in Hellsten et al. (2010). This increased the estimated repeat density from 34.5% to 39.7%, closer to the 40–50% seen in mammalian genomes.

We predicted 26,550 protein-coding loci, and a total of 30,258 transcripts; 23,907 loci had "complete" open reading frames with both predicted start and stop codons for one or more splice isoform (see Table S5). Over 42% of loci (11,329) have *Xenopus* EST support (from *X. tropicalis* and/or *X. laevis* ESTs) over 90% of the lengths of their predicted spliced transcripts. For 21,858 predicted *X. tropicalis* peptides (82%), alignment to known tetrapod peptides covers more than 90% of their sequence, and there is extensive support for coding sequences from PFAM (Finn et al., 2015) (19,669; 74% of predicted loci) and PANTHER (Mi et al., 2015) (21,048; 79%) analyses.

We further assessed the completeness and accuracy of the *X. tropicalis* annotation by evaluating its similarity to the well-annotated human



	v4	v9
Median Human/*X.tropicalis* peptide length	1.02	1.002
Human = *X.tropicalis* peptide length	377	377
Human > *X.tropicalis* peptide length	7,623	6,156
Human < *X.tropicalis* peptide length	3,928	5,395
Median Human/*X.tropicalis* exon number	1	1
Human = *X.tropicalis* exon number	5,224	7,495
Human > *X.tropicalis* exon number	3,828	2,643
Human < *X.tropicalis* exon number	2,876	1,790
Unique	3,182	2,936

Fig. 3. Exon number distribution and conservation among *Xenopus*. (A) Histogram of (Human - *X. tropicalis*)/*X. tropicalis* peptide length (v9.0 *X. tropicalis* annotation, v83 Ensembl annotation for human). Most genes are about equal (median = 0.005), with a few outliers longer in human presumably due to annotation errors in *X. tropicalis*. The distribution around 0 may be interesting to investigate for biologically significant differences between the two protein sets. (B) Histogram of (Human - *X. tropicalis*)/*X. tropicalis* exon number (v9.0 *X. tropicalis* annotation, v83 Ensembl annotation for human). Most genes are about equal (median = 0), with a few outliers longer in human presumably due to annotation errors in *X. tropicalis*. The distribution around 0 may be interesting to investigate for biologically significant differences between the two protein sets. (C) Table illustrating the improved nature of the v9 *X. tropicalis* annotation compared to v4.

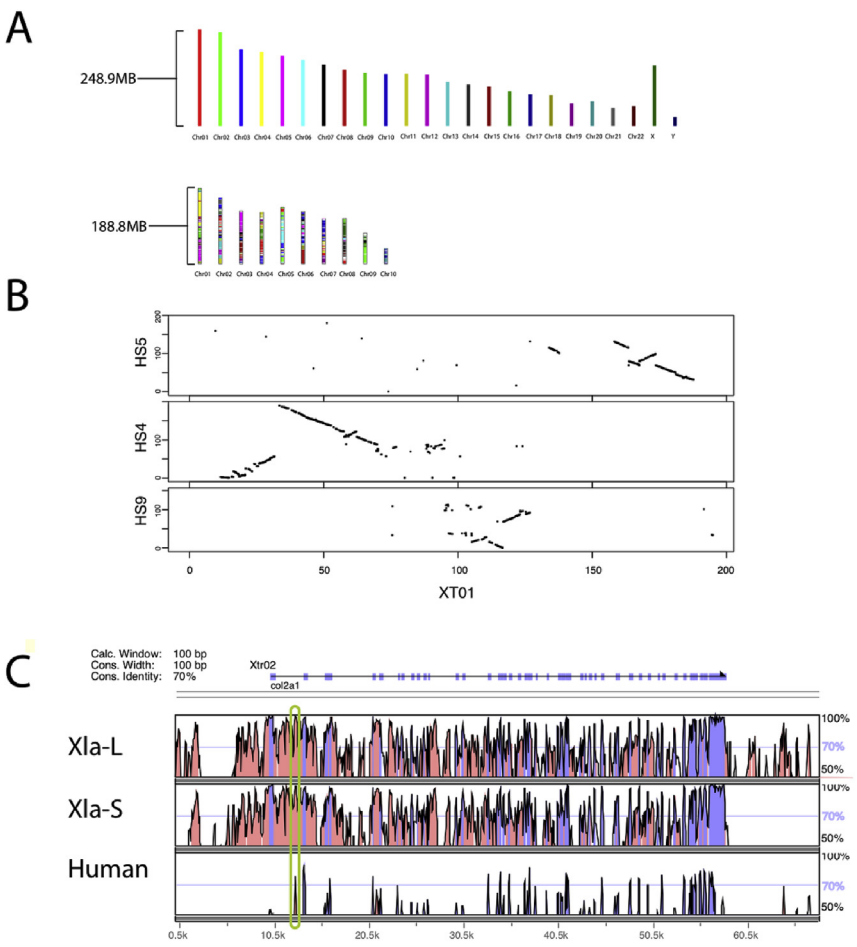


Fig. 4. Global conserved synteny between *X. tropicalis* and human. (A) Barplot showing large scale chromosome synteny between human (HS) and *X. tropicalis* (XT). Stretches of synteny were assigned based on the v9.0 *X. tropicalis* protein annotation alignment to the human proteome. Ungapped lengths of human and *X. tropicalis* chromosome 1 are shown to the left of each row. Chromosome lengths are scaled within each row, but not between rows. (B) Dotplot showing synteny of *X. tropicalis* chromosome 1 with the human proteome. Each dot corresponds to an orthologous pair of protein coding regions. *X. tropicalis* chromosome 1 (XT01) is shown on the x-axis; the corresponding orthologous human chromosomes (HS) are on the y. Tick marks on both axes measure space in Mb. The y-axes span the entire lengths of HS9, 4, and 5. Note that the syntenic orthologs of HS4 are completely mapped to a large block of XT01. (C) Vista alignment of *Xenopus* sequences surrounding *col2a1* mRNA, using *X. tropicalis* v9.0 as the reference sequence. Within *Xenopus* exon sequences are highly conserved (highlighted in blue). The latter exons and conserved noncoding sequence in the first intron are conserved even with human, and have been shown to drive expression in frogs (bottom row, circled) (Kerney et al., 2010). The circled flanking genomic sequence is unannotated, but conserved across (sub)-genomes and could potentially contain more interesting regulatory sequences.

proteome (Ensembl 83). 14,864 chromosomal proteins from the v9.1 annotation have a clear 1:1 ortholog (defined as mutual-best-BLAST hit, see **Methods 2.5**; including scaffold sequences increases the total to 15,020) in human, a slight improvement over the 14,796 1:1 orthologs found with v4.2. The predicted peptide lengths of v9.1 are more similar to the lengths of their human orthologs than those of v4.2, improving from 98.8% of the human peptide length to 99.5% (Fig. 3A). 1:1 orthologs between *X. tropicalis* and human have nearly identical exon numbers (Fig. 3B). The median difference in exon number is 0 for both annotations, with 2413 (3842) *X. tropicalis* genes having more (fewer) exons than their human orthologs in v9.1. Of course, some of these differences are *bona fide* evolutionary changes in one or both lineages during the ~360 million years since their divergence.

We assigned descriptive names to 69% of *X. tropicalis* genes using the official *Xenopus* gene nomenclature. Following the v4.2 genome release and publication, the Xenbase team (James-Zorn et al., 2012; Karpinka et al., 2014) and the *Xenopus* community manually annotated 4700 *X. tropicalis* protein-coding genes. At the same time GenBank analyzed the v4.2 genome with their own annotation and naming pipeline and found 28,829 genes with some similarity to a known mammalian gene, including large families whose 1:1 orthology is less certain. These genes aligned to a total of 18,214 loci in v9.1, with many alternate splice isoforms contributing to the redundancy. We gave priority to the manually annotated genes from Xenbase (Karpinka et al., 2014) and combined the data sets for v9.1 to assign descriptive names to 18,367 of our 26,550 primary loci. The details of the v9.1 names are described in Session et al. (2016). Briefly, priority was given to the manually curated community names provided by Xenbase, then to NCBI names, then further names based on the *Xenopus* alignments discussed above.

3.6. Comparative genome analysis of *X. tropicalis* with other tetrapods

The improved *X. tropicalis* genome and annotation is a platform for both comparative genomics and genetics, and for detailed evolutionary biology. We aligned the v9 proteome to the Ensembl 83 annotations for human, dog, and chicken. Significant rearrangements disrupt the tetrapod chromosomes, but we find large blocks of synteny scattered throughout the genome (Hellsten et al., 2010), including between the mammalian X chromosome, the *X. tropicalis* chromosome 2 (XT02), and two chicken (*Gallus gallus*) chromosomes (GGA01 and GGA23), consistent with previous findings showing deep ancestry of these regions (Mácha et al., 2012). Fig. 4A shows a barplot of the global conserved synteny between *X. tropicalis* and human. The dotplots in Fig. 4B show that the local synteny can be disrupted (as in the region orthologous to human chromosome 9), but are often maintained, despite hundreds of millions of years of divergence between the genomes. Fig. 4C illustrates conserved exon sequence in *col2a1* from *Xenopus* to human and in the non-coding sequence in the UTRs and introns within *Xenopus*.

3.6.1. Chromosome-specific sequence characteristics

The chromosome-scale assembly also allows us to ask questions about the evolutionary differences between *X. tropicalis* chromosomes. The synonymous substitution rate (K_S) of genes on *X. tropicalis* chromosome 10 is elevated when compared to *X. laevis* (Fig. 5A), despite *X. tropicalis* chromosomes 9 and 10 sharing the same chromosomal history throughout the polyploid evolution, due to a q-q fusion between chromosomes 9 and 10 in the *X. laevis* ancestor prior to the divergence of the progenitor species (Fig. 5B) (Session et al., 2016; Uno et al., 2013). Fig. 5C shows the 3rd codon position GC% between *X. tropicalis* chromosomes. *X. tropicalis* chromosome 10 has an elevated 3rd codon GC%.

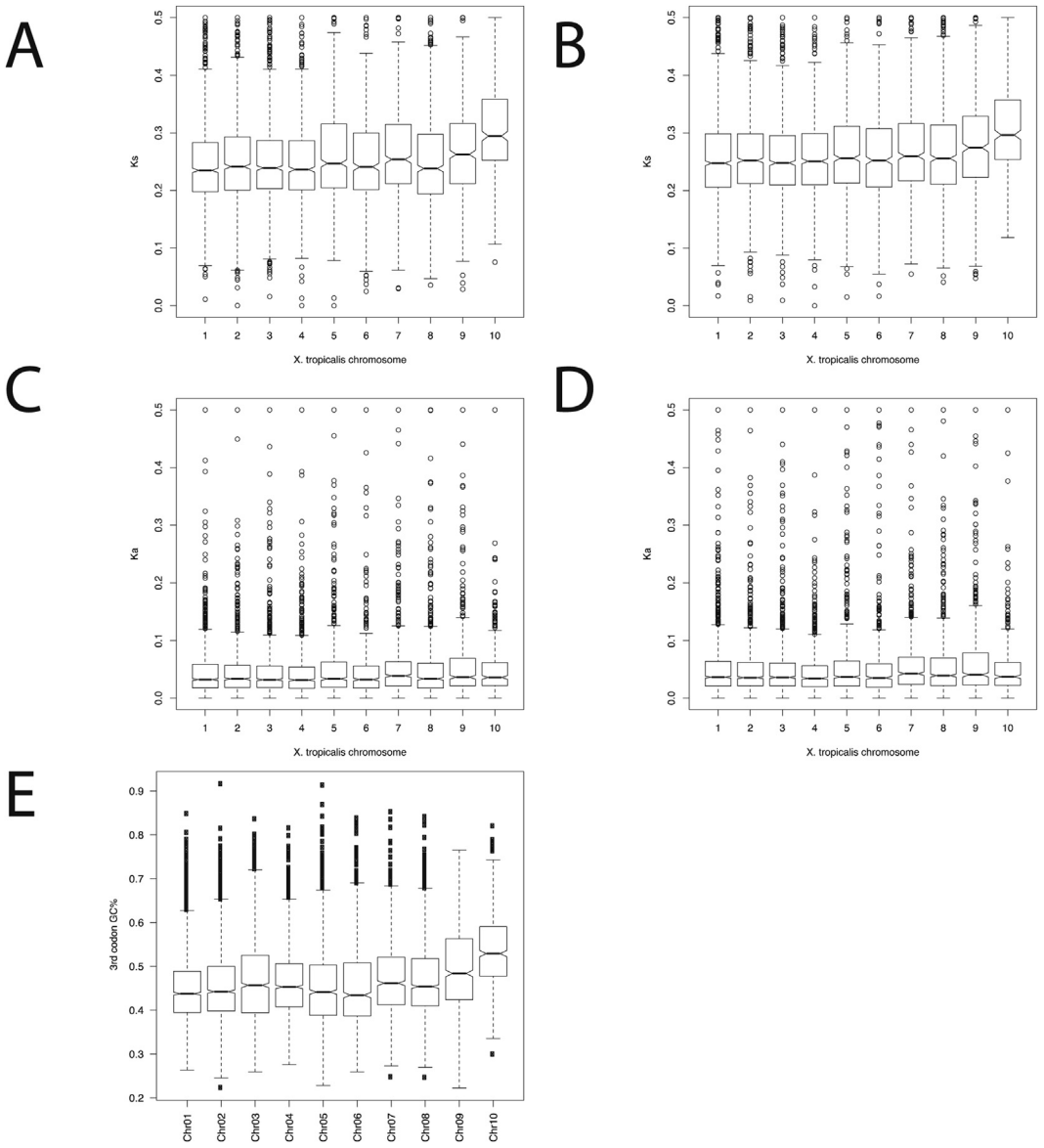


Fig. 5. Chromosome-specific characteristics. (A, B) Boxplots showing synonymous substitution rate (Ks) between *X. tropicalis* and *X. laevis* subgenomes (Xla-L in A, Xla-S in B), grouped by *X. tropicalis* chromosome (x-axis). The Ks of genes on *X. tropicalis* chromosome 10 is significantly higher than *X. tropicalis* chromosomes 1–9. ($p = 7.1E-39$ for Xtr-Xla-L Ks, $p = 1.78E-28$ for Xtr-Xla-S Ks). (C, D) Boxplots showing nonsynonymous substitution rate (Ka) between *X. tropicalis* and *X. laevis* subgenomes (Xla-L in C, Xla-S in D). The Ka of genes on *X. tropicalis* chromosome 10 is not significantly different from *X. tropicalis* chromosomes 1–9 ($p = 0.10$ for Xtr-Xla-L Ka, $p = 0.43$ for Xtr-Xla-S Ka). (E) Boxplots of 3rd codon GC% by *X. tropicalis* chromosome (*X. tropicalis* on the left, *X. laevis-L* on the right, *X. laevis-S* not shown). The 3rd codon GC% is significantly higher on *X. tropicalis* chromosome 10 than on *X. tropicalis* chromosomes 1–9 ($p = 2.3E-88$).

These observations support a model that the small size of *X. tropicalis* chromosome 10 causes a higher rate of recombinations/nucleotide (assuming at least one recombination/chromosome arm/gamete (Khokha et al., 2009)) compared to the other 9 chromosomes. Gene conversion that occurs during recombination is often GC-biased (Duret and Galtier, 2009). Therefore the small *X. tropicalis* chromosome 10 may be subject to an elevated rate of converting gene loci to the higher GC%. We hypothesize that this bias was relaxed when the chromosome fusion happened in the *X. laevis* ancestor.

3.7. Genetic mapping of pigmentation and sex

The combination of chromosome-scale assembly and dense linkage map provides a toolkit for mapping genetic determinants of mutant and natural phenotypes. To show the feasibility of mapping a natural phenotype to an area of the genome, we recorded the color (green or gray) and sex for 95 genotyped F2's (Fig. S4, Supplementary File 1) and correlated the allelic state of 2019 markers with these traits (**Methods**).

3.7.1. Genetic mapping of gray mutation

The recessive mutant *gray* emerged during the inbreeding of the ICB line of *X. tropicalis*. The *gray* phenotype displays a Mendelian inheritance pattern, consistent with a single *gray* locus (see Note S3). The variant coloration of *gray* frogs is a consequence of defects in both number and morphology of the yellow xanthophores as well as the iridescent iridophores in the adult dorsal skin (see Note S3 and Fig. S5). We find strong linkage (LOD = 14.8) to a single mapping window at the beginning of chromosome 8 (Fig. 6) with $\alpha \leq 0.001$. The highest LOD score for the binary *gray* QTL was assigned by R/qtl to the 1 cM interval at 3 cM on chromosome 8. The nearest genotyped marker bin, at 2.2 cM, is super_1372. Marker bin super_1372 is an 80 kb region comprising an underlying version 7 scaffold, and corresponds to 30,690–111,302 on

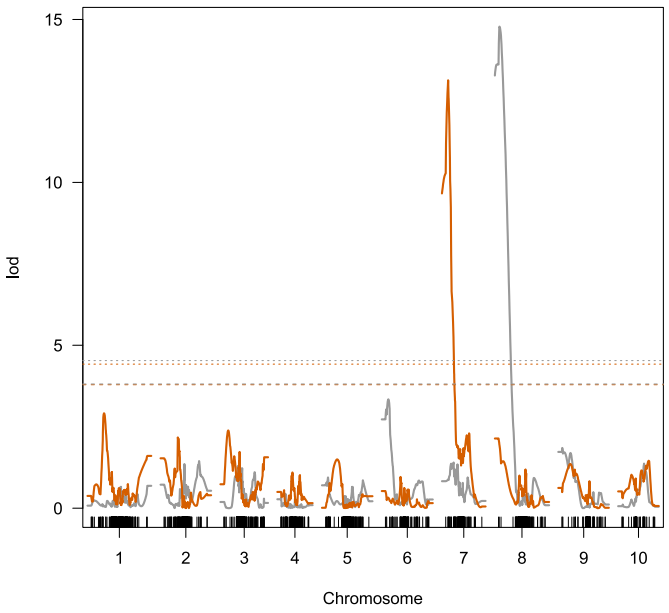


Fig. 6. Linkage mapping for sex and pigmentation. LOD scores by binary trait mapping for the *gray* trait (in gray) and sex (in orange). The *gray* locus mapped to within a 1 cM window centered at 3 cM on chromosome 8. The sex locus maps to a marker bin at 6.69 cM on chromosome 7. Horizontal lines represent $\alpha = .01$ (dotted) and $\alpha = 0.05$ (dashed) with a test of 10,000 permutations.

chromosome 8 in assembly v9. The 95% Bayes credible interval for the *gray* phenotype goes from 0 to 10 cM in chromosome 8. The 35 genes in this expanded window are listed in Supplementary File 3; none are obvious candidates for *gray*. Further discussion of candidate genes for *gray* can be found in Note S3.

3.7.2. Genetic mapping of sex determining locus

Sex determination in amphibians includes both genetic and environmental components (Nakamura, 2009). In *Xenopus*, sex is genetically determined via homomorphic sex chromosomes (Tymowska, 1991). *X. laevis* uses a ZZ/ZW mode of sex determination, via the female-specific DM-W gene (Yoshimoto et al., 2010). DM-W originated from a partial duplication of DMRT1 in the lineage leading to *X. laevis*, after its divergence from that leading to *X. tropicalis*—thus, we cannot assume that *X. tropicalis* uses the same mechanism (Bewick et al., 2011). According to Roco et al. (2015), W, Z, and Y chromosomes are present in *X. tropicalis*, and some combinations of parental sex chromosomes produce unisex offspring and other distorted sex ratios in *X. tropicalis*. In the common frog *Rana temporaria*, attempts to detect a genetic component to sex determination using high-density linkage maps were unsuccessful (Brelsford et al., 2016), and *Rana* shows a high degree of conserved synteny with *X. tropicalis* (Palomar et al., 2017).

We find strong sex linkage (LOD = 13.1) to marker super_547:0 (Fig. 6), which corresponds to the interval 1,731,784–2,045,519 on chromosome 7. At this marker, males either have the same genotype as the Nigerian male grandparent, or are hybrid N/ICB; females either have the same genotype as the ICB female grandparent or are hybrid N/ICB (Table S7). The observation that either sex can be hybrid N/ICB at this sex-correlated locus is consistent with several scenarios for the sex determination alleles present in the P0 grandparental frogs (see Note S4), and with the findings of Roco et al. (2015). There are 49 genes annotated in this 95% confidence interval; these are listed in Supplementary File 3. These scenarios cannot be resolved with our F2 mapping strategy, since the markers we use are homozygous for both the male and female grandparents, and heterozygous in their F1 progeny. We have, however, identified markers that are tightly linked to the sex-determining locus. The 95% Bayes credible interval for our sex determination marker, from 0 to 9.5 cM (corresponding to bases 1–3,906,563) on chromosome 7, is consistent with the AFLP genotyping-derived marker used by Roco et al. (2,190,623–2,191,015 of chromosome 7) (Roco et al., 2015). Genes that control sexual development tend to contain DM DNA-binding domains, for example, *Drosophila doublesex* and *C. elegans mab-3,* for which the DM domain is named; medaka *DMY/Dmrt1bY,* chicken *DMRT1,* and *X. laevis DMRT1* and *DM-W* all contain DM domains (Yoshimoto et al., 2008). No DM domains are annotated, however, for any of the genes in the mapping region. The genes present in the expanded mapping window are included in Supplementary File 3.

Based on the work of Roco et al. (2015), we infer that W, Z, and possibly Y alleles are segregating in our mapcross (see Note S4). Depending on the genotypes of the grandparents, and of the F1 parents (randomly) chosen to make the F2 clutches, eight out of ten possible crossing scenarios result in both homozygous N and hybrid males, and homozygous IC and hybrid females, present in a given F2 clutch (see Fig. S6).

3.7.3. A recessive lethal

We also note that most of chromosome 2 has an under-representation of ICB homozygotes among adults, but not tadpoles (Fig. S2). Presumably this is because the ICB grandparent was heterozygous for a recessive lethal allele on chromosome 2 which was passed on to the F1 parents used in this study, allowing homozygous genotypes of the recessive lethal to appear in the F2 generation.

3.8. Weighted gene co-expression network analysis

In order to facilitate evolutionary analysis of developmental networks we used published RNA-seq data (Tan et al., 2013) for *X. tropicalis* to construct weighted gene co-expression networks (WGCNA) (Langfelder and Horvath, 2008). 10,703 genes had sufficiently high gene expression in the Tan et al. data to be used for network construction (**Methods 2.9**). In a WGCNA network, a "module" represents a set of co-expressed genes. The expression level of each module is summarized by a single "eigengene" that represents a weighted average (more properly, the first principal component) of the expression level of genes belonging to that module. The expression pattern of the WGCNA eigengenes (compared with scaled expression profiles for genes belonging to the corresponding module) are shown in Fig. 7 and Fig. S7. The network identity of individual genes is available in Supplementary File 4.

Supplementary File 5 shows the retention of *X. laevis* (co)-orthologs in each of the *X. tropicalis* WGCNA groups. We found no relationship between group membership (indicated by a distinguishing color) and genome location. Interestingly the ME-1 "low in Stage 16" group, and ME-2 "expressed in Stage 9–44" groups show significantly low (55.7%) and high (77.1%) retention in *X. laevis*, respectively, compared with the genome-wide retention rate of 66.7%. The latter group is interesting because if we assume that expression in *X. tropicalis* reflects that of the *Xenopus* ancestor, it would imply zygotically expressed genes in the ancestor have a higher chance of being retained as two copies in the tetraploid. This was not observed in *X. laevis* expression WGCNA groups based on the sum of homeologous gene expression patterns (Session et al., 2016). Interestingly, the highly retained ME-2 group is enriched in genes related to the *fibroblast growth factor* (*fgf*) pathway, which was found to be retained at a higher rate in the *X. laevis* duplication. 15 of the 16 homeolog pairs annotated as involved in the fgf pathway are retained in *X. laevis* and 9 of their *X. tropicalis* orthologs are classified in co-expression networks. Six out of nine of these are clustered into the ME-2 expression group (the remaining 3 are in clusters that differ in their maternal expression pattern). The six ME-2 *fgf* loci can only contribute a small amount to the increased retention seen for the 306 genes being grouped together in that coexpression cluster.

The ME-1 group shows an increased number of classifications associated with mitochondrial gene function such as mitochondrial outer membrane, mitochondrial import receptor, oxidation reduction, and FeS

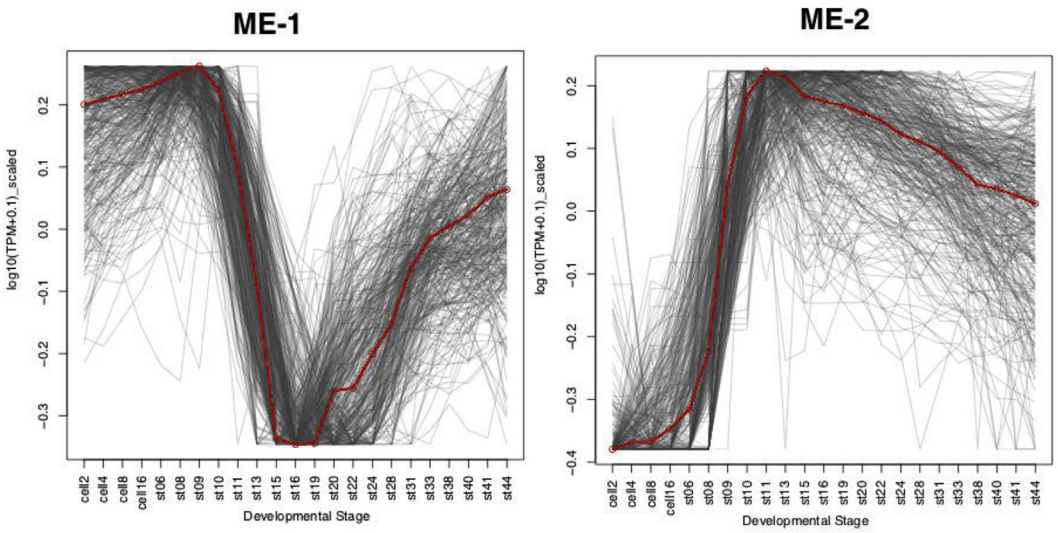


Fig. 7. Weighted Gene Coexpression Network Eigengenes. Out of 19 WGCNA co-expression clusters, one (ME-1: expressed in oocyte and early cleavage, lost after MZT; minimum at stage 15–19; rising at later stages) showed reduced retention after genome duplication in *X. laevis*. Another (ME-2: sharp gene expression onset after MZT) showed increased retention. Other co-expression clusters are shown in Fig. S7. For individual genes, expression values (y-axis) are scaled to match the range of expression of the corresponding eigengene.

cluster binding. While we did not find a link between mitochondrial gene localization and retention when analyzing the orthologs of genes known to localize to the mitochondria (Session et al., 2016), these results suggest specific pathways within the mitochondria may be subject to reduction to single copy. These analyses provide an example of the advanced genomics questions that can be answered given the new advances in *Xenopus* genomics.

4. Conclusions

The improvements made to the *X. tropicalis* genome and annotation described here provide key data sets to the *Xenopus* research community that allow for accurate comparative genomics. The new assembly contributed to better understanding of the evolving HoxB cluster in *X. laevis* (Kondo et al., 2017) as well as structural genome evolution following polyploidy (Session et al., 2016). Here we show its usefulness in allowing for mapping of a mutant locus and the *X. tropicalis* sex locus, as well as the ability of the complete annotation to investigate the syntenic relationship between tetraploid chromosomes. While there are still improvements to be made to bring the frog assemblies to the level of the model mammals such as human and mouse, the upgrades outlined here enable a large number of previously impossible experiments in the model amphibian genomes.

Data availability

The *Xenopus tropicalis* assembly is deposited in RefSeq under accession GCF_000004195.3. The RNAseq reads are published (Tan et al., 2013), and are available at the Gene Expression Omnibus under GSE37452. The XENLAv9 genome assembly and annotation used for comparison are deposited at NCBI (accession LYTH00000000). Demultiplexed GBS reads are deposited in SRA (BioProject ID PRJNA526297).

Author contributions

Therese Mitros: Genome assembly, analyzed GBS data, built genetic map, mapped traits, wrote the paper.

Jessica B. Lyons: GBS protocol design and library preparation, phenotyping for sex, participated in analysis, wrote the paper.

Adam M. Session: Annotation, analysis of gene expression, orientation of chromosomes, analysis of gene structure, wrote the paper.

Jerry Jenkins: Map integration, chromosome assembly, analysis, contributed text.

Shengquiang Shu: Annotation pipeline and refinement, contributed text.

Taejoon Kwon: RNA-seq alignment, gene expression calculation, contributed text.

Maura Lane: *gray* project execution and analysis, contributed text.

Connie Ng: *gray* project assistance.

Timothy C. Grammer: Leadership and experimental design for *gray* project.

Mustafa K. Khokha: *gray* analysis, input on GBS adapter design, contributed text.

Jane Grimwood: Sequencing of BAC-ends; QC projects at HudsonAlpha.

Jeremy Schmutz: ARACHNE assembly and BAC-end sequencing lead at HudsonAlpha.

Richard M. Harland: Project leadership.

Daniel S. Rokhsar: Project leadership, wrote the paper.

Acknowledgements

This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303.

Work at UC Berkeley was supported by NIH grant GM66684 “Comparative genetics and genomics of *Xenopus*” to DSR and RMH, and NIH grant HD08070 “Systematic Improvement of *Xenopus* Gene Annotations and Reference Genomes” to DSR. MKK was supported with a K08-HD42550 award from the NICHD/NIH. TK was supported with the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning (NRF-2016R1C1B2009302), and the UNIST research fund (1.180063 and 1.180040). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors would like to thank Nicolas Pollet who kindly provided us with IC TGA frogs, the Harland Laboratory Trop Team for their assistance with animal husbandry, and Ajeet Pratap Singh and Tetsuaki Kimura for advice on the *gray* candidate gene list. We also thank Reena

Zalpuri from the Electron Microscope lab at UC Berkeley for training and assistance with electron microscopy, as well as Dr. Alan M. Kuzirian at the Marine Biological Laboratory, Woods Hole, MA for comments on electron microscopic images.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ydbio.2019.03.015.

References

Bewick, A.J., Anderson, D.W., Evans, B.J., 2011. Evolution of the closely related, sex-related genes *DM-W* and *DMRT1* in African clawed frogs (*Xenopus*). Evolution 65, 698–712.

Bhattacharya, D., Marfo, C.A., Li, D., Lane, M., Khokha, M.K., 2015. CRISPR/Cas9: an inexpensive, efficient loss of function tool to screen human disease genes in *Xenopus*. Dev. Biol. 408, 196–204.

Blitz, I.L., Biesinger, J., Xie, X., Cho, K.W.Y., 2013. Biallelic genome modification in F0 *Xenopus tropicalis* embryos using the CRISPR/Cas system, 51, 827–834.

Bogdanović, O., van Heeringen, S.J., Veenstra, G.J.C., 2012. The epigenome in early vertebrate development. Genesis 50, 192–206.

Brelsford, A., Rodrigues, N., Perrin, N., 2016. High-density linkage maps fail to detect any genetic component to sex determination in a *Rana temporaria* family. J. Evol. Biol. 29, 220–225.

Broman, K.W., Wu, H., Sen, Ś., Churchill, G.A., 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19, 889–890.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinf. 10, 421.

Chain, F., Evans, B.J., 2006. Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. PLoS Genet. 2, e56.

Collart, C., Owens, N.D.L., Bhaw-Rosun, L., Cooper, B., De Domenico, E., Patrushev, I., Sesay, A.K., Smith, J.N., Smith, J.C., Gilchrist, M.J., 2014. High-resolution analysis of gene activity during the *Xenopus* mid-blastula transition, 141, 1927–1939.

Duret, L., Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genom. Hum. Genet. 10, 285–311.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379.

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., 2016. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Res. 45, D190–D199.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., 2015. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285.

Grammer, T.C., Khokha, M.K., Lane, M.A., Lam, K., Harland, R.M., 2005. Identification of mutants in inbred *Xenopus tropicalis*. Mech. Dev. 122, 263–272.

Guo, X., Zhang, T., Hu, Z., Zhang, Y., Shi, Z., Wang, Q., Cui, Y., Wang, F., Zhao, H., Chen, Y., 2014. Efficient RNA/Cas9-mediated genome editing in *Xenopus tropicalis*. Development 141, 707–714.

Harland, R.M., Grainger, R.M., 2011. *Xenopus* research: metamorphosed by genetics and genomics. Trends Genet. 27, 507–515.

Hayes, T.B., Collins, A., Lee, M., Mendoza, M., Noriega, N., Stuart, A.A., Vonk, A., 2002. Hermaphroditic, demasculinized frogs after exposure to the herbicide atrazine at low ecologically relevant doses. Proc. Natl. Acad. Sci. Unit. States Am. 99, 5476–5480.

Hellsten, U., Harland, R.M., Gilchrist, M.J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N.H., Shu, S., Taher, L., 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science 328, 633–636.

Hellsten, U., Khokha, M.K., Grammer, T.C., Harland, R.M., Richardson, P., Rokhsar, D.S., 2007. Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. BMC Biol. 5, 31.

International Cassava Genetic Map Consortium (ICGMC), 2015. High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from 10 populations. G3 Genes, Genomes, Genet. 5, 133–144.

Igawa, T., Watanabe, A., Suzuki, A., Kashiwagi, A., Kashiwagi, K., Noble, A., Guille, M., Simpson, D.E., Horb, M.E., Fujii, T., Sumida, M., 2015. Inbreeding ratio and genetic relationships among strains of the Western clawed frog, *Xenopus tropicalis*. PloS One 10, e0133963.

Ishibashi, S., Cliffe, R., Amaya, E., 2012. Highly efficient bi-allelic mutation rates using TALENs in *Xenopus tropicalis*. Biol. Open 1, 1273–1276. BIO20123228.

Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., Lander, E.S., 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 13, 91–96.

James-Zorn, C., Ponferrada, V.G., Jarabek, C.J., Burns, K.A., Segerdell, E.J., Lee, J., Snyder, K., Bhattacharyya, B., Karpinka, J.B., Fortriede, J., Bowes, J.B., Zorn, A.M., Vize, P.D., 2012. Xenbase: expansion and updates of the *Xenopus* model organism database. Nucleic Acids Res. 41, D865–D870.

Karpinka, J.B., Fortriede, J.D., Burns, K.A., James-Zorn, C., Ponferrada, V.G., Lee, J., Karimi, K., Zorn, A.M., Vize, P.D., 2014. Xenbase, the *Xenopus* model organism database; new virtualized system, data types and genomes. Nucleic Acids Res. 43, D756–D763.

Kerney, R., Hall, B.K., Hanken, J., 2010. Regulatory elements of *Xenopus col2a1* drive cartilaginous gene expression in transgenic frogs. Int. J. Dev. Biol. 54, 141–150.

Khokha, M.K., 2012. *Xenopus* white papers and resources: folding functional genomics and genetics into the frog. Genesis 50, 133–142.

Khokha, M.K., Krylov, V., Reilly, M.J., Gall, J.G., Bhattacharya, D., Cheung, Y.J.C., Kaufman, S., Lam, D.K., Macha, J., Ngo, C., Prakash, N., Schmidt, P., Tlapakova, T., Trivedi, T., Tumova, L., Abu-Daya, A., Geach, T., Vendrell, E., Ironfield, H., Sinzelle, L., Sater, A.K., Wells, D.E., Harland, R.M., Zimmerman, L.B., 2009. Rapid gynogenetic mapping of *Xenopus tropicalis* mutations to chromosomes. Dev. Dynam. 238, 1398-1346.

Kondo, M., Yamamoto, T., Takahashi, S., Taira, M., 2017. Comprehensive analyses of *hox* gene expression in *Xenopus laevis* embryos and adult tissues. Dev. Growth Differ. 59, 526–539.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 9, 559.

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. 12, 323.

Li, H., 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM arXiv preprint arXiv:1303.3997.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Love, N.R., Chen, Y., Bonev, B., Gilchrist, M.J., Fairclough, L., Lea, R., Mohun, T.J., Paredes, R., Zeef, L.A.H., Amaya, E., 2011. Genome-wide analysis of gene expression during *Xenopus tropicalis* tadpole tail regeneration. BMC Dev. Biol. 11, 70.

Mácha, J., Teichmanová, R., Sater, A.K., Wells, D.E., Tlapáková, T., Zimmerman, L.B., Krylov, V., 2012. Deep ancestry of mammalian X chromosome revealed by comparison with the basal tetrapod *Xenopus tropicalis*. BMC Genomics 13, 315.

Matsuda, Y., Uno, Y., Kondo, M., Gilchrist, M.J., Zorn, A.M., Rokhsar, D.S., Schmid, M., Taira, M., 2015. A new nomenclature of *Xenopus laevis* chromosomes based on the phylogenetic relationship to *Silurana/Xenopus tropicalis*. Cytogenet. Genome Res. 145, 187–191.

Meyer, M., Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010 pdb. prot5448.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2015. PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Res. 44, D336–D342.

Nakamura, M., 2009. Sex Determination in Amphibians, Seminars in Cell & Developmental Biology. Elsevier, pp. 271–282.

Nakayama, T., Fish, M.B., Fisher, M., Oomen-Hajagos, J., Thomsen, G.H., Grainger, R.M., 2013. Simple and efficient CRISPR/Cas9-mediated targeted mutagenesis in *Xenopus tropicalis*, 51, 835–843.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., …, Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745.

Owens, N.D., Blitz, I.L., Lane, M.A., Patrushev, I., Overton, J.D., Gilchrist, M.J., Cho, K.W.Y., Khokha, M.K., 2016. Measuring absolute RNA copy numbers at high temporal resolution reveals transcriptome kinetics in development, 14, 632–647.

Palomar, G., Ahmad, F., Vasemägi, A., Matsuba, C., Nicieza, A.G., Cano, J.M., 2017. Comparative high-density linkage mapping reveals conserved genome structure but variation in levels of heterochiasmy and location of recombination cold spots in the common frog. G3 Genes, Genomes, Genet. 7, 637–645.

Paranjpe, S.S., Jacobi, U.G., van Heeringen, S.J., Veenstra, G.J.C., 2013. A genome-wide survey of maternal and embryonic transcripts during *Xenopus tropicalis* development. BMC Genom. 14, 762.

Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J., Fields, A., Hartley, P.D., Sugnet, C.W., Haussler, D., Rokhsar, D.S., Green, R.E., 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26, 342–350.

Roco, Á.S., Olmstead, A.W., Degitz, S.J., Amano, T., Zimmerman, L.B., Bullejos, M., 2015. Coexistence of Y, W, and Z sex chromosomes in *Xenopus tropicalis*. Proc. Natl. Acad. Sci. Unit. States Am. 112, E4752–E4761.

Rodrigues, N., Vuille, Y., Brelsford, A., Merilä, J., Perrin, N., 2016. The genetic contribution to sex determination and number of sex chromosomes vary among populations of common frogs (*Rana temporaria*). Heredity 117, 25.

Sémon, M., Wolfe, K.H., 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc. Natl. Acad. Sci. Unit. States Am. 105, 8333–8338.

Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., …, Rokhsar, D.S., 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature 538, 336.

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., …, Rokhsar, D.S., 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493, 526.

Sive, H.L., Grainger, R.M., Harland, R.M., 2000. Early Development of *Xenopus laevis*: a Laboratory Manual. CSHL Press.

Smit, A., Hubley, R., 2008-2015. RepeatModeler Open-1.0. http://www.repeatmasker.org.

Spirhanzlova, P., Dhorne-Pollet, S., Fellah, J., Da Silva, C., Tlapakova, T., Labadie, K., Weissenbach, J., Poulain, J., Jaffredo, T., Wincker, P., Krylov V. and Pollet N., 2017. Construction and characterization of a BAC library for functional genomics in *Xenopus tropicalis*. Dev. Biol. 426, 255–260.

Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., Li, J.B., 2013. RNA sequencing reveals a diverse and dynamic repertoire of the *Xenopus tropicalis* transcriptome over development. Genome Res. 23, 201–216.

Tymowska, J., 1973. Karyotype analysis of *Xenopus tropicalis* gray, pipidae. Cytogenet. Genome Res. 12, 297–304.

Tymowska, J., 1991. Polyploidy and cytogenetic variation in frogs of the genus *Xenopus*. Amphib. Cytogenet. Evolut. 259, 297.

Uno, Y., Nishida, C., Takagi, C., Ueno, N., Matsuda, Y., 2013. Homoeologous chromosomes of *Xenopus laevis* are highly conserved after whole-genome duplication. Heredity 111, 430.

Van Ooijen, J., 2006. JoinMap 4: Software for the Calculation of Genetic Linkage Maps in Experimental Populations of Diploid Species. Plant Research International BV and Kyazma BV, Wageningen, Netherlands.

Voss, S.R., Kump, D.K., Putta, S., Pauly, N., Reynolds, A., Henry, R.J., Basa, S., Walker, J.A., Smith, J.J., 2011. Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. Genome Res. 21, 1306–1312.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.-h., Jin, H., Marler, B., Guo, H., Kissinger, J.C., Paterson, A.H., 2012. *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40 e49-e49.

Wells, D.E., Gutierrez, L., Xu, Z., Krylov, V., Macha, J., Blankenburg, K.P., Hitchens, M., Bellot, L.J., Spivey, M., Stemple, D.L., …, Sater, A.E., 2011. A genetic map of *Xenopus tropicalis*. Dev. Biol. 354, 1–8.

Yanai, I., Peshkin, L., Jorgensen, P., Kirschner, M.W., 2011. Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility. Dev. Cell 20, 483–496.

Yoshimoto, S., Ikeda, N., Izutsu, Y., Shiba, T., Takamatsu, N., Ito, M., 2010. Opposite roles of *DMRT1* and its W-linked paralogue, *DM-W*, in sexual dimorphism of *Xenopus laevis*: implications of a ZZ/ZW-type sex-determining system dev. 048751.

Yoshimoto, S., Okada, E., Umemoto, H., Tamura, K., Uno, Y., Nishida-Umehara, C., Matsuda, Y., Takamatsu, N., Shiba, T., Ito, M., 2008. A W-linked DM-domain gene, DM-W, participates in primary ovary development in *Xenopus laevis*. Proc. Natl. Acad. Sci. Unit. States Am. 105, 2469–2474.

Young, J.J., Cherone, J.M., Doyon, Y., Ankoudinova, I., Faraji, F.M., Lee, A.H., Ngo, C., Guschin, D.Y., Paschon, D.E., Miller, J.C., Zhang, L., Rebar, E.J., Gregory, P.D., Urnov, F.D., Harland, R.M., Zeitler, B., 2011. Efficient targeted gene disruption in the soma and germ line of the frog Xenopus tropicalis using engineered zinc-finger nucleases. Proc. Natl. Acad. Sci. Unit. States Am. 108, 7052–7057.