Xenopus tropicalis v10 reference genome assembly release

Sequencing, assembly, and annotation

Whole-genome shotgun datasets: Pacific Biosciences (PacBio) Sequel long-read and 10X Genomics Chromium linked-read sequencing data were generated at the HudsonAlpha Institute of Biotechnology (HudsonAlpha) from an F17 Nigerian strain female, contributed by Mustafa K. Khokha and Maura Lane at the Yale School of Medicine. This resulted in 65.4 Gb (38× depth) of long reads and 114.7 Gb (67× depth) of linked-reads. In addition, an Illumina short-insert library was constructed by the Functional Genomics Laboratory at the University of California Berkeley (UCB) and sequenced 2×251 bp paired-end by the QB3 Vincent J. Coates Genomics Sequencing Laboratory (VCGSL) to 95.0 Gb (55× depth). Using a blood sample from a sister of the above frog, Hi-C libraries were constructed by Dovetail Genomics LLC and sequenced 2×151 bp paired-end by the VCGSL, yielding 108 million chromatin contacts.

Genome assembly: Contigs were assembled from PacBio long-reads using both *de novo* and hybrid strategies. *De novo* contigs were assembled with Canu v1.6-132-gf9284f8 (Koren 2017). Hybrid contigs were constructed from Supernova-derived (v1.1.5, assembled by HudsonAlpha) (Weisenfeld 2017) contigs integrated with long-reads using DBG2OLC (commit 1f7e752) (Ye 2016). The two assemblies were merged with quickmerge (Chakraborty 2016) and scaffolded with 3 kb, 8 kb, 40 kb, and 140 kb insert Sanger paired-ends, fosmids, and BAC-ends (Mitros 2019) using SSPACE3 (Boetzer 2010). Chromosomal scaffolds were constructed with Hi-C using Juicer and 3D-DNA (Durand 2016; Dudchenko 2017) and then manually curated in JuiceBox (Dudchenko 2018). Synteny information with other unpublished *Xenopus* assemblies was also incorporated. The resulting scaffolds were gap-filled with PBJelly v15.8.24 (English 2012) and polished using Arrow (Chin 2013), Pilon (Walker 2014), and custom scripts. The genome assembly was performed by Jessen V. Bredeson, Sanjit S. Batra, and Austin B. Mudd in the Rokhsar Lab at UCB.

Generation of amphibian repeat library: Repeat Modeler v1.0.11 was run on an intermediate, unpolished version of the assembled *X. tropicalis* contigs. Identified repeats were manually curated to exclude false positives and recover false negatives, leading to a total of 973 repeats. Repeats from seven other Anuran genome assemblies (*A. truei* [n=1,769], *E. coqui* [n=1,441], *E. pustulosus* [n=1,146], *H. boettgeri* [n=1,160], *P. adspersus* [n=908], *X. borealis* [n=1,026], and *X. laevis* [n=913]) were also identified and included into a pan-Anuran repeat library. We combined the frog and ancestral RepBase v23.12 dataset (n=934) with the curated repeats above to create the final repeat library (n=10,270) used for the *Xenopus tropicalis* v10 repeat annotation (RepeatMasker v4.0.7).

EST assembly: We obtained 1,271,375 *Xenopus tropicalis* EST sequences from NCBI. ESTs sequenced from the forward and reverse orientations were assembled using PEAR v0.9.8 (Zhang 2013). Successfully assembled ESTs and single ESTs with a minimum sequence length of 250bp were mapped against the v10 genome assembly using STARlong v2.7.0e (Dobin 2013).

Genome-guided transcriptome assembly: Stranded RNA-seq data from adult tissues (Marin 2017) and non-stranded RNA-seq data from different developmental stages (Owens 2016) were obtained from the Sequence Read Archive (SRA). We only considered samples that were poly-A selected. We pooled samples from equivalent stage or tissue type to increase sequencing read depth. The RNA-seq reads were aligned against the unmasked version of the *Xenopus tropicalis* genome assembly v10 using STAR v2.7.0e (Dobin 2013). Alignments became the inputs for the Trinity transcriptome assembler v2.5.1 (Grabherr 2011). We obtained a total of 2.4 billion transcript models, which were subsequently evaluated after being mapped to the genome using STARlong.

Filtering transcript model predictions from Trinity and ESTs: The EST and transcript model predictions mapped by STARlong, together with a splice junctions database obtained from a second round of RNA-seq alignments, were used to evaluate the quality of the transcript models. Transcripts and ESTs were discarded if they showed inconsistencies in splice junction support. This step was critical to minimize the number of gene fusions, as these often exhibit significantly fewer reads connecting the first and last exons from two adjacent genes. Transcripts and ESTs were also discarded if they presented abnormal mapping features (*e.g.*, low mapping quality, short intron lengths, short first or last exon lengths).

Genome annotation: Filtered EST and transcript models were used as mRNA evidence by the JGI Integrated Gene Call (IGC) pipeline (Shengqiang Shu unpublished) for genome annotation. *X. tropicalis* clones (n=8,980) from the Mammalian Gene Collection (MGC) (Klein 2002) were used as the set of confident and full-length mRNA evidence. MGC clones from *X. laevis* (n=11,515) were utilized as sister transcripts. Human, mouse, chicken, and zebrafish proteins were used for peptide homology evidence. Several rounds of genome annotation and evaluation were implemented to assess the completeness of the gene predictions.

The transcriptome assembly, repeat and gene annotations were performed by Sofia Medina Ruiz in the Rokhsar Lab at UCB. Excluding *P. adspersus* (Denton 2018), the remaining six unpublished Anuran genomes used for identifying pan-Anuran repeats were assembled by Austin B. Mudd, Jessen V. Bredeson, Sanjit S. Batra, and Kodiak C. Berkoff in the Rokhsar Lab at UCB.

Assembly statistics

Scaffold sequence total / count	1,451.3 Mb	166
Scaffold L50 / N50	154.0 Mb	5
Scaffold L90 / N90	91.2 Mb	9
Contig sequence total / count	1,448.4 Mb	850
Contig L50 / N50	14.6 Mb	32
Contig L90 / N90	1.9 Mb	133

Annotation statistics

Annotation version	10.7
Taxonomy ID	8364
Primary transcripts (loci)	25,016
Alternate transcripts	37,983
Total transcripts	62,999

Primary transcripts:

Average number of exons	8.7
Median exon length	137
Median intron length	1,075
Number of complete genes	24,167
Number of incomplete genes with start codon	276
Number of incomplete genes with stop codon	360

Gene model support:

Number of genes with Pfam annotation	18,895
Number of genes with Panther annotation	20,727
Number of genes with KOG annotation	12,109
Number of genes with KEGG Orthology annotation	10,252
Number of genes with E.C. number annotation	5,175

Data use policy

We are committed to early data release and, as a service to the research community, we are therefore making the chromosome-scale *Xenopus tropicalis* genome assembly version 10 available prior to publication. We (the data producers) encourage the pre-publication use of these data in accordance with the Toronto data sharing principles described at https://www.nature.com/articles/461168a. A manuscript describing the analysis and comparisons with other species in under preparation.

Questions about data use should be directed to Jessen V. Bredeson or Sofia Medina Ruiz (see contact information below).

Files

- ***Xentr10.fasta.gz*** **:** The haploid assembly sequence in FASTA format, block-compressed with bgzip. This file includes chromosomal scaffolds 1–10 and 156 unplaced scaffolds. Chromosome and scaffold names are prefixed with “Chr” and “Sca”, respectively, and lack zero-padding (*e.g.*, “Chr1”).
- ***Xentr10.repeatMasked.fasta.gz*** **:** Genome assembly sequence as in *Xentr10.fasta.gz* above, but soft-masked with RepeatMasker v4.0.7 using a custom repeat library derived

from RepBase_23.12 and other amphibian repeat sequences identified by Repeat Modeler. See *Xentr10.7.repeatMasked.gff* for the repeat annotations.

- ***Xentr10.lowqual.bed*** **:** A BED-formatted file highlighting regions of the assembly that have low base-level accuracy due to lack of read coverage in sequence polishing.
- ***Xentr10.from9.chain*** **:** UCSC chain-formatted lift-over table for mapping genome-anchored features/annotations from the v9 assembly to the v10 assembly.
- ***Xentr10.to9.chain*** **:** UCSC chain-formatted lift-over table for mapping genome-anchored features/annotations from the v10 assembly to the v9 assembly.
- ***Xentr10.7.repeatMasked.gff*** **:** Repeat annotation file in GFF format.
- ***Xentr10.7.allTrs.cds.fasta*** **:** Nucleotide sequences of the coding regions from all transcript model predictions (excludes non-coding regions).
- ***Xentr10.7.allTrs.pep.fasta*** **:** Peptide sequences from all transcript model predictions.
- ***Xentr10.7.allTrs.nuc.fa*** **:** Complete mRNA sequences from all transcript model predictions (includes 5' and 3' UTRs).
- ***Xentr10.7.primaryTrs.cds.fasta*** **:** Nucleotide sequences of all primary transcript model. A primary transcript is defined as the longest coding CDS at a locus.
- ***Xentr10.7.primaryTrs.pep.fasta*** **:** Peptide sequences of all primary transcript model. A primary transcript is defined as the longest coding CDS at a locus.
- ***Xentr10.7.altTrs.gff3*** **:** Nucleotide sequence sequences from alternative gene isoforms.
- ***Xentr10.7.gene.gff3*** **:** Gene annotation file containing primary and alternative isoforms.
- ***Xentr10.7.gene_exons.gff3*** **:** Gene annotation file containing the primary and alternative isoforms. Same as *Xentr10.7.gene.gff3* but with exon features.
- ***Xentr10.7.primaryTrs.gff3*** **:** Gene annotations of primary transcripts only.
- ***Xentr10.7.annot.summary*** **:** Gene annotation and evidence summary in a human-readable table. Includes EST, peptide, and functional evidence types.
- ***Xentr10.7.JGI_gene_EST.ovlps.txt*** **:** A text table summarizing annotation evidence from the different sources used by the annotation pipeline.
- ***Xentr10.7.transcript.functions.txt*** **:**
  1. transcriptName: transcript ID associated to one of the isoforms from a gene locus
  2. Id: ID of the database ID
  3. IdType: Database type (PANTHER, GO, PANTHER, SIGNALP, KEGGORTH, PFAM, or EC)
  4. Description: If available for the database ID
- ***Xentr10.7.gene.functions.txt*** **:**
  1. locusName: geneID associated to a single gene locus (primary transcripts)

2. Id: ID of the database ID
3. IdType: Database type (PANTHER, GO, PANTHER, SIGNALP, KEGGORTH, PFAM, or EC)
4. Description: If available for the database ID

- **Xentr10.7.annot_hgnc.tab :** Master table summarizing the functional annotation of primary transcripts. The table includes results from BLASTP run between *X. tropicalis* and Human peptides under the following columns: HGNC id, HGNCdescription, HGNC blastp_evalue, HGNC cscore (BLASTP score ratio), genes with same HGNC id. For convenience, peptide sequence associated to the primary transcript is also provided.

Contributors

UC Berkeley
Daniel S. Rokhsar (UCB, DOE JGI)
Richard M. Harland
Sanjit S. Batra
Kodiak C. Berkoff
Jessen V. Bredeson
Sofia Medina Ruiz
Therese Mitros
Austin B. Mudd

DOE Joint Genome Institute
Shengqiang Shu

Yale University
Mustafa K. Khokha
Maura Lane

Contacts

Principal Investigator: Dan S. Rokhsar (UCB), email: dsrokhsar -AT- gmail -DOT- com
Assembly: Jessen V. Bredeson (UCB), email: jessenbredeson -AT- berkeley -DOT- edu
Annotation: Sofia Medina Ruiz (UCB), email: sofiamr -AT- berkeley -DOT- edu

References

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*. 2010 Dec 12; 27(4):578-9.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic acids research*. 2016 Jul 25; 44(19):e147-e147.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*. 2013 Jun; 10(6):563.

Denton RD, Kudra RS, Malcom JW, Du Preez L, Malone JH. The African Bullfrog (*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making vertebrate sex chromosomes. *bioRxiv*. 2018 Jan 1:329847.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013 Jan 1; 29(1):15-21.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017 Apr 7; 356(6333):92-5.

Dudchenko O, Shamim MS, Batra S, Durand NC, Musial NT, Mostofa R, Pham M, St Hilaire BG, Yao W, Stamenova E, Hoeger M. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv.* 2018 Jan 1:254797.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*. 2016 Jul 27; 3(1):95-8.

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one*. 2012 Nov 21; 7(11):e47768.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*. 2011 Jul; 29(7):644.

Klein SL, Strausberg RL, Wagner L, Pontius J, Clifton SW, Richardson P. Genetic and genomic tools for *Xenopus* research: The NIH *Xenopus* initiative: A peer reviewed forum. *Developmental dynamics*: an official publication of the American Association of Anatomists. 2002 Dec; 225(4):384-91.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*. 2017 May 1; 27(5):722-36.

Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert J, Brüning T, Mössinger K, Trefzer T, Conrad C, Kerver HN, Wade J, Tschopp P, Kaessmann H. Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage. *Genome Res*. 2017 Dec; 27(12):1974-1987.

Mitros T, Lyons JB, Session AM, Jenkins J, Shu S, Kwon T, Lane M, Ng C, Grammer TC, Khokha MK, Grimwood J. A chromosome-scale genome assembly and dense genetic map for Xenopus tropicalis. *Developmental biology.* 2019 Apr 10.

Owens NDL, Blitz IL, Lane MA, Patrushev I, Overton JD, Gilchrist MJ, Cho KWY, Khokha MK. Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. *Cell Rep*. 2016 Jan 26; 14(3):632-647.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*. 2014 Nov 19; 9(11):e112963.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome research*. 2017 May 1; 27(5):757-67.

Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific reports*. 2016 Aug 30; 6:31900.

Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*. 2013 Oct 18; 30(5):614-20.