**Supplementary Information**

**Title:**

**First draft genome of loach (Orenectus shuilongensis; Cypriniformes: Nemacheilidae) provide insights into the evolution of cavefish**

Zhijin Liu1,2,3,4,†\*, Xuekun Qian3,4,†, Ziming Wang4,†, Huamei Wen2,†, Ling Han2,4, Jiang Zhou2,\*

1 College of Life Sciences, Capital Normal University, Beijing 100048, China.

2 School of Life Sciences, Guizhou Normal University, Guiyang, Guizhou, 550001, China.

3 Colleage of Life Sciences, Hebei University, Baoding, Hebei, 071002, China.

4 Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China.

† Contributed equally

\* Correspondence: Jiang Zhou, zhoujiang@ioz.ac.cn; Zhijin Liu, liuzj@ioz.ac.cn

**Short running title**: Loach genome

1. [*De novo* sequencing and assembly………………………………………………3](file:///E%3A%5CA%20work%5C%E5%8F%B6%E7%8C%B4%E6%96%87%E7%AB%A0%5Ctougao%5CLangur_genome_SI-submission_final_zhangly.docx#_Toc377658642)

1.1 Sample information………………………………………………………….3

1.2 Illumina sequencing………………………………………………………….3

1.3 Estimation of genome size using *K*-mer method……………………………..3

1.4 PacBio sequencing………………………………………………….………4

1.5 *De novo* Assembly……………………………………………………….…4

1.6 Evaluation of genome assembly…………………………………………….5

1. Genome annotation……………………………………………………………………..7

2.1 Annotation of repeats sequences……………………………………………..7

 2.2 RNA preparation and sequencing…………………………………….7

2.3 Annotation of protein coding genes………………………………………….7

2.4 Non-coding RNA annotation……………………………………………….9

1. Genome evolution………………………………………………………………10

3.1 Global gene family classification……………………………………………10

3.2 Phylogenetic relationship and genomic comparison………………………10

1. Re-sequencing analysis[………………………………………………………..12](file:///E%3A%5CA%20work%5C%E5%8F%B6%E7%8C%B4%E6%96%87%E7%AB%A0%5Ctougao%5CLangur_genome_SI-submission_final_zhangly.docx#_Toc377658664)

4.1 Sample collection and sequencing………………………………………….12

4.2 Sequence data pre-processing and variant calling………………………….12

4.3 Pseudogenization enrichment analysis……………………………………12

1. Reference……[…………………………………………………………………14](file:///E%3A%5CA%20work%5C%E5%8F%B6%E7%8C%B4%E6%96%87%E7%AB%A0%5Ctougao%5CLangur_genome_SI-submission_final_zhangly.docx#_Toc377658677)
2. ***De novo* sequencing and assembly**

**1.1 Sample information**

The present study was approved by the Animal Ethics Committee of Guizhou Normal University. The procedure of sample collection was in strict accordance with the Animal Ethics Procedures and Guidelines of the People's Republic of China. One individual of *Orenectus shuilongensis* used in *de novo* assembly analysis were obtained from Shuilong Township, Sandu County, Guizhou Province.

**1.2 Illumina sequencing**

Genomic DNA was extracted using DNeasy Blood &Tissue Kit (Qiagen) from muscle. Three small-insert libraries (270bp) were constructed by using Illumina’s paired-end kits according to the manufacturer’s instructions. The libraries were sequenced on Illumina Hiseq X Ten platforms. For the raw reads, sequencing adaptors were removed; contaminated reads (chloroplast, mitochondrial, bacterial and viral sequences, etc.) were screened by alignment to the NCBI-NR database using BWA v0.7.13[1](file:///C%3A%5CUsers%5CMr%20zhang%5CDesktop%5CCavefish%5C%E5%B2%AD%E9%B3%85Notes_Text.doc#_ENREF_1) with default parameters; the FastUniq v1.1[2](file:///C%3A%5CUsers%5CMr%20zhang%5CDesktop%5CCavefish%5C%E5%B2%AD%E9%B3%85Notes_Text.doc#_ENREF_2) was used to remove the duplicated read pairs; the low-quality reads were filtered satisfying the following conditions: 1) reads with ≥10% unidentified nucleotides (N), 2) reads with >10 nucleotides aligned to the adapter, allowing ≤10% mismatches, 3) reads with >50% bases having Phred quality <5. Finally, we generated a total of 120.94 Gb clean reads for paired-end (**Supplementary Table 1**).

**Supplementary Table 1. Statistics of Illumina sequencing data.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | #Library | Data (Gb) | Depth (X) | Q20 (%) | Q30 (%) |
| Paired-Ends | 270 bp\_1 | 43.3 | 83.97 | 97.53 | 94.02 |
| 270 bp\_2 | 41.73 | 80.92 | 97.33 | 93.6 |
| 270 bp\_3 | 35.92 | 69.66 | 97.44 | 93.83 |
| Total | - | 120.94 | 234.55 | - | - |

**1.3 Estimation of genome size using K-mer method**

Corrected Illumina reads were selected to perform genome size estimation. The distribution of 19-kmer showed a major peak at 199× (**Supplementary Fig. 1**). Based on the total number (102,673,535,298) and corresponding to a kmer depth of 199, the genome size was estimated to be 515.64 Mb using the formula: Genome size= kmer\_Number/Peak\_Depth.



**Supplementary Fig. 1.** Distribution of 19-kmer.

**1.4 PacBio sequencing**

Single-molecule sequencing was done on the PacBio Sequal platform. After removal of shorter than 500bp PacBio subreads we yielded 5,000,440 subreads with an average length of 10,187bp which were performed genome assembly (**Supplementary Table 2**).

**Supplementary Table 2. Statistics of PacBio raw data.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **Read Bases (bp)** | **Read Number** | **Mean Read Length (bp)** | **Read Quality** |
| Pacbio | 5,000,440 | 50,938,604,673 | 10,187 | 81,434 |

**1.5 *De novo* assembly**

The single-molecule sequencing (SMS) data are assembled through Canu13, then the draft assembly polished through Pilon24. Canu is a comprehensive and scalable pipeline for SMS data assembly (available at https://github.com/marbl/canu, v1.5). In the correction step, Canu first selects longer seed reads with the settings ‘genomeSize = 520M’ and ‘corOutCoverage = 90’, then detects raw reads overlapping through a highly sensitive overlapper MHAP (mhap-2.1.2, option ‘cor Mhap Sensitivity = normal’), and finally performs an error correction through the falcon sense method (option ‘corrected Error Rate = 0.045’). In the next step, with the default parameters, error-corrected reads are trimmed of unsupported bases and hairpin adapters to get their longest supported range. In the last step, Canu generates the draft assembly using trimmed reads (**Supplementary Table 3**).

The draft assembly is polished to obtain the final assembly. The polishing adopts pilon algorithm (v1.22, available at https://github.com/broadinstitute/pilon) using illumina data with the parameters ‘--mindepth 10 --changes --threads 4 --fix bases’. The final *O.* *shuilongensis* genome information summarized in **Supplementary Table 4**.

**Supplementary Table 3. Length distribution of PacBio subreads.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Length (bp)** | **Number** | **Total Length (bp)** | **Average Length (bp)** |
| 500~2000 | 688,917 | 827,673,911 | 1,201.41 |
| 2000~4000 | 691,407 | 2,037,289,224 | 2,946.58 |
| 4000~6000 | 565,220 | 2,811,371,742 | 4,973.94 |
| 6000~8000 | 487,074 | 3,397,735,959 | 6,975.81 |
| 8000~10000 | 436,548 | 3,925,820,704 | 8,992.87 |
| 10000~12000 | 437,807 | 4,813,712,393 | 10,995.06 |
| 12000~14000 | 378,521 | 4,906,442,552 | 12,962.14 |
| 14000~16000 | 298,656 | 4,467,568,349 | 14,958.91 |
| 16000~18000 | 232,229 | 3,938,078,331 | 16,957.74 |
| 18000~ | 784,061 | 19,812,911,508 | 25,269.6 |
| Total | 5,000,440 | 50,938,604,673 | 10,187 |

**Supplementary Table 4. Statistics of the genome assembly**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Contig Number** | **Contig Length (bp)** | **Contig N50 (bp)** | **Contig N90 (bp)** | **Contig Max (bp)** | **GC Content (%)** |
| 803 | 521,689,915 | 5,584,306 | 292,936 | 16,441,799 | 38.34 |

**1.6 Evaluation of genome assembly**

Completeness of the assembly was assessed respectively through CEGMA v2.55 and BUSCO v26. In total, 457 (99.78%) of the conserved Core Eukaryotic Genes (CGEs) and 248 (100%) of the highly CGEs were found to be present (**Supplementary Table 5**). Furthermore, 4,473 (97.58%) (**Supplementary Table 6**) of the vertebrate Benchmarking Universal Single-Copy Orthologs were found to be present, indicating that most genic sequences were present in the *O.* *shuilongensis* genome assembly. The draft assembly was evaluated by mapping the high-quality reads from short insert size PE libraries to the scaffolds using BWA1. Around 99.52% of the reads could be mapped to the assembly and 98.91% of the reads could be properlymapped to the assembly (**Supplementary Table 7**).Furthermore, genome synteny relationships between *O.* *shuilongensis* and *Danio rerio* were defined by MCscanX[7](#_ENREF_20) based on orthologous gene sets identified using BLAST[8](#_ENREF_21) (*E*-value ≤10-5; number of genes required to all syntenies ≥10) (**Supplementary Fig. 2**).

**Supplementary Table 5. Evaluation of genome assembly through CEGMA.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Number of 458 CEGs\* present in assembly** | **% of 458 CEGs present in assemblies** | **Number of 248 highly conserved CEGs present** | **% of 248 highly conserved CEGs present** |
| 457 | 99.78% | 248 | 100% |

**Supplementary Table 6. Evaluation of genome assembly through BUSCO.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Complete BUSCOs** | **Complete and single-copy BUSCOs** | **Complete and duplicated BUSCOs** | **Fragmented BUSCOs** | **Total** |
| 4,371 (95.35%) | 4,162 (90.79%) | 209 (4.56%) | 102 (2.23%) | 4,473 (97.58%) |

**Supplementary Table 7. The alignment information of reads mapping to the genome.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Total reads** | **Mapped reads** | **Mapped (%)** | **Properly mapped reads** | **Properly mapped (%)** |
| 289,292,979 | 287,910,918 | 99.52 | 285,869,522 | 98.91 |



**Supplementary Fig. 2.** Synteny between *O.* *shuilongensis* and *D. rerio.*

1. **Genome annotation**

**2.1 Annotation of repeats sequences**

The repeat composition of the assemblies was estimated by building a repeat library employing the *de novo* prediction programs LTR-FINDER9, MITE-Hunter10, RepeatScout11 and PILER-DF12. The database was classified using PASTEClassifier13 and was then combined with the Repbase database14 to create the final repeat library. Repeat sequences in the *O.* *shuilongensis* genome were identified and classified using the RepeatMasker program15. The LTR family classification criterion was that 5’ LTR sequences of the same family would share at least 80% identity over at least 80% of their length. The characteristics of repeat sequences summarized in **Table 3**.

**2.2** **RNA preparation and sequencing**

We also performed RNA-sequencing for the cDNA libraries from the same loach individual used for genome sequencing and assembly. Tissues of skin, muscle, intestinal, liver and kidney were collected and RNAs were extracted with TRIZOL Reagent (Invitrogen, USA). RNAs were then balanced mixed for the sequencing. The absorbance of 1.90 at 260 nm/280 nm and the RIN of 9.1 were obtained for the purified RNA sample by Nanodrop ND-1000 spectrophotometer (LabTech, USA) and 2100 Bioanalyzer (Agilent Technologies, USA), respectively. According to the protocol, one microgram of RNA was reverse transcribed using Clontech SMARTer cDNA synthesis kit, and was further fragmented using divalent cations for the sequencing. The paired-end library was prepared following the manual of the Paired-End Sample Preparation Kit (Illumina Inc., San Diego, CA, USA). The library with an insert length of 270 bp was sequenced by Illumina HiSeq X Ten in 150 bp paired-end mode (Illumina Inc., San Diego, CA, USA). Finally, a total of 11.7 Gb transcriptome data were obtained from RNA- sequencing (Table 1).

**2.3 Annotation of protein coding genes**

Protein-coding genes were predicted based on *de novo*, protein homology and RNA-Seq approaches. Genscan16, Augustus17, GlimmerHMM18, GeneID19 and SNAP20 were performed *de novo* gene prediction. The homologous peptides alignment to our assemblies were used to identify homologous genes with GeMoMa21; the RNA-Seq reads were assembled into contigs *de novo* into unigenes using Trinity and the resulting unigenes were aligned to the repeat-masked assemblies using BLAT22, and subsequently the gene structures of BLAT alignment results were modeled using PASA; additionally, the RNA-Seq reads were also assembled into transcripts through mapping to the assembled genome using Hisat2 v2.0.423 and Stringtie v1.3.024, and the protein-coding regions were identified with TransDecoder v3.0.125 and GeneMarkS-T26, respectively. Finally, these consensus gene models were generated by integrating the *de novo* predictions, protein alignments and transcripts data using EVidenceModeler27 (**Supplementary Table 8**).

**Supplementary Table 8. Statistics of each protein-coding gene set and integrated prediction.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Software** | **Species** | **Gene Number** |
| *Ab initio* | Genscan | - | 24,503 |
| Augustus | - | 31,176 |
| GlimmerHMM | - | 63,275 |
| GeneID | - | 26,766 |
| SNAP | - | 69,477 |
| Homology-based | GeMoMa | *Danio rerio* | 24,393 |
| *Astyanax mexicanus* | 28,134  |
| *Cyprinus carpio* | 24,422 |
| *Sinocyclocheilus rhinocerou*s | 24,603 |
| RNAseq | PASA | - | 20,418 |
| GeneMarkS-T | - | 42,495  |
| TransDecoder | - | 54,423 |
| Integration | EVM | - | 25,247 |

Annotation of the predicted genes were performed by blasting their sequences against a number of nucleotide and protein sequence databases, including GO28, KOG29, KEGG30, NCBI-NR31 and Swiss-Prot32 and with an *E*-value cutoff of 1e-5. The annotation information of protein-coding genes summarized in **Supplementary Table 9.**

**Supplementary Table 9. Functional annotation of protein-coding genes.**

|  |  |  |
| --- | --- | --- |
| **Database** | **Annotated Number** | **Percentage (%)** |
| GO Annotation | 14,284 | 56.58 |
| KEGG Annotation | 11,795 | 46.72 |
| KOG Annotation | 16,288 | 64.51 |
| TrEMBL Annotation | 23,960 | 94.90 |
| NR Annotation | 24,119 | 95.53 |
| All Annotated | 24,149 | 95.65 |

**2.4 Non-coding RNA annotation**

Non-coding RNAs play important roles in a great variety of processes, such as the rRNAs and tRNAs involved in mRNA translation. The rRNA fragments were identified by aligning the rRNA template sequences (Pfam database v31) using BLAST with *E*-value at 1e-10 and identity cutoff at 95% or more. The tRNAScan-SE33 algorithms with default parameters were applied to the prediction of tRNA genes. The miRNA genes were predicted by INFERNAL v1.1 software34 against the Rfam database35 with cutoff score at 30 or more. The minimum cutoff score was based on the settings which yield a false positive rate of 30 bits. The non-coding RNAs annotated information summarized in **Supplementary Table 10.**

**Supplementary Table 10. Non-coding RNAs annotation.**

|  |  |  |
| --- | --- | --- |
| **Classification** | **Number** | **Family** |
| miRNA | 947 | 238 |
| rRNA | 561 | 4 |
| tRNA | 417 | 25 |

1. Genome evolution

**3.1** **Global gene family classification**

In order to identify gene families among fish species in this work, proteins of the longest transcripts of each individual gene from *O. shuilongensis* and other sequenced species, including*Salmo salar*, *Ictalurus punctatus*, *A. mexicanus*, *C. carpio*, *S. rhinocerous, D. rerio*, *Larimichthys crocea* were analyzed. All data was downloaded from NCBI31. Gene family analysis based on the homolog of gene sequences in related species was initially implemented by the alignment of an “all against all” BLASTP36 with a cutoff of 1e-5 and subsequently followed by alignments with high-scoring segment pairs conjoined for each gene pair by Solar. To identify homologous gene pairs, we required more than 30% coverage of the aligned regions in both homologous genes. Finally, homologous genes were clustered into gene families by OrthoMCL37 with the inflation parameter set at 1.5. As a result, 16,708 gene families were constructed for the *O. shuilongensis*. Among the families, there were 144 families unique to *O. shuilongensis* (**Fig. 2 and Supplementary Table 11**).

**Supplementary Table 11. Gene family statistics.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Species** | **Total gene number** | **Cluster gene number** | **Total family number** | **Unique gene family number** |
| *L. crocea* | 24,623 | 22,688 | 15,157 | 182 |
| *O. shuilongensis* | 25,247 | 23,145 | 16,708 | 144 |
| *S. rhinocerous* | 42,217 | 37,958 | 17,579 | 86 |
| *S. salar* | 46,355 | 39,871 | 17,185 | 1,003 |
| *C. carpio* | 48,809 | 36,236 | 17,882 | 539 |
| 1. *mexicanus*
 | 24,981 | 23,421 | 16,374 | 187 |
| *D. rerio* | 32,258 | 31,006 | 16,891 | 267 |
| *I. punctatus* | 22,680 | 21,641 | 15,536 | 94 |

**3.2** **Phylogenetic relationship and genomic comparison**

Evolutionary analysis was performed using the single-copy protein-coding genes among all species. Amino acid and nucleotide sequences of the ortholog genes were aligned using the multiple alignment software MUSCLE38 with default parameters. A total number of 108 single-copy ortholog alignments were concatenated into a super alignment matrix of 242,085 nucleotides. A maximum likelihood method deduced tree was inferred based on the matrix of nucleotide sequences using PhyML39 package with the JTT+G+F model. Clade support was assessed using bootstrapping algorithm in the PhyML package with 100 alignment replicates. A molecular clock data from the divergence time between Cyprinidae and Salmonidae [230.4 million years ago (MYA), 95% Confidence interval (CI): 204.5-255.3] from the TimeTree database40. According to the phylogenetic analysis, *O. shuilongensis* were clustered together with Cyprinid fishes (*D. rerio*, *S. rhinocerous* and *C. carpio*), which was consistent with the fish species taxonomy. *O. shuilongensis* diverged from the common ancestor with Cyprinid fishes around 92.8 MYA (95% CI: 73.4 - 108.1) (**Fig. 3)**.

We determined the expansion and contraction of the orthologous gene families by comparing the cluster size differences between the ancestor and each of the *O. shuilongensis* and seven other fish species using the CAFÉ41 program. A random birth and death model were used to study changes of gene families along each lineage of the phylogenetic tree. A probabilistic graphical model (PGM) was introduced to calculate the probability of transitions in gene family size from parent to child nodes in the phylogeny. Using conditional likelihoods as the test statistics, we calculated the corresponding *P*-values in each lineage. A *P*-value of 0.05 was used to identify families that were significantly expanded in *O. shuilongensis* genome. When comparing with this other seven fish, the expansion and contraction of gene orthology clusters showed 77 gene families were expanded and 282 gene families contracted significantly in the *O. shuilongensis* (**Fig. 3)**.

1. **Re-sequencing analysis**

**4.1 Sample collection and sequencing**

The samples of three individuals of *O. jiarongensis*, two individuals of *O. daqikongensis* and one individual of *O.dongliangensis* were collected. Total genomic DNA was extracted from the tissue samples using the chloroform method. All experimental procedures and sample collections were conducted under the supervision of the Committee for Animal Experiments of the Institute of Zoology, CAS. For each individual, ~3μg of DNA was sheared into fragments of 270 bp with the Covaris v1.8 system. DNA fragments were then processed and sequenced using the Illumina HiSeq 4000 platform. The raw pair-end reads were trimmed to remove the adaptors and low-quality bases and after quality control by FastQC42. The raw reads were filtered with the following criteria: (1) reads with unidentified nucleotides (N) > 10% were discarded, (2) reads with the proportion of low-quality base (phred quality <=10) > 50% were discarded. (**Table 4**)

**4.2 Sequence data pre-processing and variant calling**

Filtered sequence reads were mapped to the langur reference genome using BWA-MEM with default parameters (0.7.10-r789)1. Alignment bam files were imported to SAMtools (v0.1.19)43 for sorting and removing duplicated reads and Picard (http://broadinstitute.github.io/picard/, version 1.92) was used to assign read group information containing library, lane and sample ID. Following mapping, we performed variant calling using the GATK44 package with default parameters on individual-scale for all samples. Only mapped reads without gaps and with less than five mismatches were included in the identification of SNPs. The variants were filtered unless the minimum root-mean-square (RMS) mapping quality was 20. Variants were then removed if their average Phred scaled base quality was lower than 30 or the distance between the SNP and a gap was less than 5 bp, the Indel and a gap was less than 10 bp. Furthermore, only variants with a coverage of at least 5 were further investigated. For each individual, the Ti/Tv (Transition/Transversion) was obtained and calculated through the whole genome, the heterozygosity was calculated as heterozygous SNP rate across the whole genome (**Table 4**). Next, SnpEff45 software was used to annotation the identified variants (**Fig. 5**).

**4.3 Pseudogenization enrichment analysis**

The whole genome SNPs and indels were used for pseudogene annotation. We found that 1,541 SNPs and 438 indels were annotated to result in pseudogenization of 401 genes. For further analysis, these candidate genes were used to KEGG enrichment and gene ontology analysis using DAVID 6.846. The enrichment results are summarized in **Table 5**. Twenty-nine pseudogenes related to eyes are summarized in **Supplementary Table 12**.

**Supplementary Table 12. Pseudogenes related to retina and eye development.**

|  |  |
| --- | --- |
| **Gene name** | **Gene description** |
| *cdh2* | cadherin 2, type 1, N-cadherin |
| *polr3f* | polymerase (RNA) III (DNA directed) polypeptide F  |
| *smo* | smoothened, frizzled class receptor |
| *wdr55* | WD repeat domain 55 |
| *lgsn* | lengsin, lens protein with glutamine synthetase domain |
| *cdh6* | cadherin 6 |
| *adcyap1b* | adenylate cyclase activating polypeptide 1b |
| *tyms* | thymidylate synthetase |
| *bcor* | BCL6 corepressor |
| *paics* | phosphoribosylaminoimidazole carboxylase |
| *atp6v1e1b* | ATPase H+ transporting V1 subunit E1b |
| *mipa* | major intrinsic protein of lens fiber a |
| *aldh1a3* | aldehyde dehydrogenase 1 family, member A3 |
| *tfap2a* | transcription factor AP-2 alpha |
| *epb41l5* | erythrocyte membrane protein band 4.1 like 5 |
| *six7* | SIX homeobox 7 |
| *nsfa* | N-ethylmaleimide-sensitive factor a |
| *pbx4* | pre-B-cell leukemia transcription factor 4 |
| *rx1* | retinal homeobox gene 1 |
| *vps39* | vacuolar protein sorting 39 homolog |
| *mbnl2* | muscleblind-like splicing regulator 2 |
| *alcama* | activated leukocyte cell adhesion molecule a |
| *tmx3a* | thioredoxin related transmembrane protein 3a |
| *actn2b* | actinin, alpha 2b |
| *mfn2* | mitofusin 2 |
| *apc* | adenomatous polyposis coli |
| *six6b* | SIX homeobox 6b |
| *cad* | carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase |
| *six6a* | SIX homeobox 6a |

**5. References**

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760 (2009).
2. Xu HB, Luo X, Qian J, et al. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *Plos One* **7**, e52249 (2012).
3. 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* **27**, 722-736 (2017).
4. Walker B J, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS One* **9**, e112963 (2014).
5. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).
6. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
7. Wang Y, Tang HB, Jeremy D, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49 (2012).
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).
9. Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265-268 (2007).
10. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research* **38**, e199 (2010).
11. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358 (2005).
12. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1**, i152-158 (2005).
13. Wicker T*,* Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973-982 (2007).
14. Bao, W, Kojima, K & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6, 11 (2015).
15. Tarailograovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **Chapter 4**, Unit 4 10 (2004).
16. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78-94 (1997).
17. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-225 (2003).
18. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879 (2004).
19. Blanco E, Parra G, Guigo R. Using geneid to identify genes. *Current Protocols in Bioinformatics* **Chapter 4**, Unit 4 3 (2007).
20. Korf I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
21. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**, e89 (2016).
22. Kent, WJ. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664 (2002).
23. Kim D, Landmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357-360 (2015).
24. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 33, 290-295 (2015).
25. Haas B, Papanicolaou A. TransDecoder (find coding regions within transcripts). (2016).
26. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* **43**,588-588 (2014).
27. Haas B J, Salzberg S L, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
28. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**, D258–D261 (2004).
29. Tatusov R L, Fedorova N D, Jackson J D, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41-41 (2003).
30. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27-30 (2000).
31. Sayers EW, Agarwala R, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **5**, (2018).
32. Boeckmann B*,* Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365-370 (2003).
33. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955-964 (1997).
34. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).
35. Griffiths-Jones, S, Bateman, A, Marshall, M, Khanna, A & Eddy, S.R Rfam: an RNA family database. *Nucleic Acids Research* **31**, 439-441 (2003).
36. Altschul SF, Thomas L, Madden AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402 (1997).
37. Thomas PD, Kejariwal A, Campbell MJ, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Research* **31**, 334-341 (2003).
38. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).
39. Guindon S, Dufayard JF, Lefort V, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307-321 (2010).
40. Hedges SB, Marin J, Suleski M, Paymer M & Kumar S. Tree of life reveals clock-like speciation and diversification. Molecular Biology & Evolution 32, 835-845 (2015).
41. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271 (2006).
42. Andrews, S. FastQC, <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> *Babraham Bioinformatics* (2010).
43. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
44. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303 (2010).
45. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3., *Fly*, **6**, 80-92 (2012).
46. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols* **4**, 44–57 (2009).