Oct 06, 2021

# High-quality reference genome of Fasciola gigantica: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions

Xier Luo[1]

[1]Agricultural Genomics Institute at Shenzhen

1

dx.doi.org/10.17504/protocols.io.bxatpien

Anonymous

*Fasciola gigantica* and *Fasciola hepatica* are causative pathogens of *fascioliasis*, with the widest latitudinal, longitudinal, and altitudinal distribution; however, among parasites, they have the largest sequenced genomes, hindering genomic research. In the present study, we used various sequencing and assembly technologies to generate a new high-quality *Fasciola gigantica* reference genome. We improved the integration of gene structure prediction, and identified two independent transposable element expansion events contributing to (1) the speciation between Fasciola and Fasciolopsis during the Cretaceous-Paleogene boundary mass extinction, and (2) the habitat switch to the liver during *the Paleocene-Eocene Thermal Maximum*, accompanied by gene length increment. Long interspersed element (LINE) duplication contributed to the second transposon-mediated alteration, showing an obvious trend of insertion into gene regions, regardless of strong purifying effect. Gene ontology analysis of genes with long LINE insertions identified membrane-associated and vesicle secretion process proteins, further implicating the functional alteration of the gene network. We identified 852 predicted excretory/secretory proteins and 3300 protein-protein interactions between *Fasciola gigantica* and its host. Among them, copper/zinc superoxide dismutase genes, with specific gene copy number variations, might play a central role in the phase I detoxification process. Analysis of 559 single-copy orthologs suggested that *Fasciola gigantica* and *Fasciola hepatica* diverged at 11.8 Ma near the Middle and Late Miocene Epoch boundary. We identified 98 rapidly evolving gene families, including actin and aquaporin, which might explain the large body size and the parasitic adaptive character resulting in these liver flukes becoming epidemic in tropical and subtropical regions.

Manuscript for Fasiola gigantica.doc

DOI

dx.doi.org/10.17504/protocols.io.bxatpien

https://doi.org/10.1371/journal.pntd.0009750

Xier Luo 2021. High-quality reference genome of Fasciola gigantica: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.bxatpien

protocol

Luo X, Cui K, Wang Z, Li Z, Wu Z, Huang W, Zhu X, Ruan J, Zhang W, Liu Q (2021) High-quality reference genome of *Fasciola gigantica*: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions. PLoS Negl Trop Dis 15(10): e0009750. doi: 10.1371/journal.pntd.0009750

protocol ,

Aug 11, 2021

Oct 06, 2021

52275

## 1 Sample collection

All animal work was approved by the Guangxi University Institutional Animal Care and Use Committee. For the reference genome sequencing, one *F. gigantica* at adult stage was derived from infected buffalo in the Guangxi Zhuang Autonomous Region.

## 2 De novo sequencing and assembly

### 2.1 Sequencing

Three *de novo* genome sequencing methods were performed on the liver fluke: We generated (1) 122.4 Gb (~88× depth) PacBio Sequel II single-molecule long reads, with an average read length of 15.8 kb (PacBio,Menlo Park, CA, USA); (2) 89.5 Gb (~66× depth) Illumina HiSeq PE150 pair-end sequencing to correct errors (Illumina,San Diego, CA, USA); and (3) 134 Gb (~100× depth) chromosome conformation capture sequencing (Hi-C) data (sequenced by Illumina platform).

### 2.2 Assembly

A PacBio-only assembly was performed using Canu v2.0 (59, 60) using new overlapping and assembly algorithms, including an adaptive overlapping strategy based on *tf-idf* weighted MinHash and a sparse assembly graph construction that avoids collapsing diverged repeats and haplotypes. To remove haplotigs and contig overlaps in the assembly, we used Purge_Dups based on the read depth (61). Arrow (https://github.com/PacificBiosciences/GenomicConsensus) was initially used to reduce the assembly error in the draft assembly, with an improved consensus model based on a more straightforward hidden Markov model approach. Pilon (62) was used to improve the local base accuracy of the contigs via analysis of the read alignment information based on paired-end bam files (thrice). As a result, the initial assembly resulted had an N50 size of 4.89 Mb for the F. gigantica reference genome. ALLHiC was capable of building chromosomal-scale scaffolds for the initial genome using Hi-C paired-end reads containing putative restriction enzyme site information (63). The whole genome assembly (contig version) have been deposited in the Genome Warehouse in BIG Data Center under accession number GWHAZTT00000000 and NCBI under Bioproject PRJNA691688.

### 2.3 Evaluation

Three methods were used to evaluate the quality of the genomes. First, we used QUality ASsessment Tool (QUAST) (64) to align the Illumina and PacBio raw reads to the *F. gigantica* reference genome to estimate the coverage and mapping rate. Second, all the Illumina paired-end reads were mapped to the final genome using BWA (65), and single nucleotide polymorphisms (SNPs) were called using Samtools and Bcftools (66). The predicted error rate was calculated by the homozygous substitutions divided by length of the whole genome, which included the discrepancy between assembly and sequencing data. Thirdly, we assessed the completeness of the genome assemblies and annotated the genes using BUSCO (18).

## 3 Genome annotation

Three gene prediction methods, based on *de novo* prediction, homologous genes, and transcriptomes, were integrated to annotate protein-coding genes. RNA-seq data of *F. gigantica* were obtained from the NCBI Sequence Read Archive, SRR4449208 (67). RNA-seq reads were aligned to the genome assembly using HISAT2 (v2.2.0) (68) and subsequently assembled using StringTie (v2.1.3) (69). PASA (v2.4) (70) was another tool used to assemble RNA-seq reads and further generated gene models to train *de novo* programs. Two *de novo* programs, including Augustus (v3.0.2) (71) and SNAP (v2006-07-28) (72), were used to predict genes in the repeat-masked genome sequences. For homology-based prediction, protein sequences from UniRef100 (73) (plagiorchiida-specific, n =75,612) were aligned on the genome sequence using TBLASTn (74) (e-value < $10^{-4}$), and GeneWise (version 2.4.1) (75) was used to identify accurate gene structures. All predicted genes from the three approaches were combined using MAKER (v3.1.2) (76) to generate high-confidence gene sets. To obtain gene function annotations, Interproscan (v5.45) (77) was used to identify annotated genes features, including protein families, domains, functional sites, and GO terms from the InterPro



**Citation:** Xier Luo High-quality reference genome of Fasciola gigantica: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions https://dx.doi.org/10.17504/protocols.io.bxatpien

database. SwissProt and TrEMBL protein databases were also searched using BLASTp (78) (e-value < $10^{-4}$). The best BLASTp hits were used to assign homology-based gene functions. BlastKOALA (79) was used to search the KEGG ORTHOLOGY (KO) database. The subsequent enrichment analysis was performed using clusterProfiler using total annotated genes as the background with the “enricher” function (80).

## 4 Repeat annotation and analysis

### 4.1 Annotation

We combined *de novo* and homology approaches to identify repetitive sequences in our assembly and previous published assemblies, including *F. gigantica*, *F. hepatica*, and *Fasciolopsis buski*. RepeatModeler (v2.0.1) (24) was first used to construct the *de novo* identification and accurate compilation of sequence models representing all of the unique TE families dispersed in the genome. Then, RepeatMasker (v4.1.0) (25) was run on the genome using the combination of *de novo* libraries and a library of known repeats (Repbase-20181026).

### 4.2 Analysis

The relative position between a repeat and a gene was identified using bedtools (81), and the type of repeat was further divided to intronic and intergenic origin. The repeat landscape was constructed using sequence alignments and the complete annotations output from RepeatMasker, depicting the Kimura divergence (Kimura genetic distances between identified repeat sequences and their consensus) distribution of all repeats types. The most notable peak in the repeat landscapes was considered as the most convincing time of repeat duplication in that period. We inferred the time of LINEs insertion by transferring Kimura divergence in RepeatMasker to age (t=d/2mu). The distributions of TE elements were calculated with sliding windows (n=50). In each sliding window, we calculated the relative proportion of TE between intronic and intergenic regions, and further corrected them using the whole ratio between intronic and intergenic regions. To calculate mutation rate, we used 559 single-copy orthologs multiple sequence alignment among 8 species produced in the latter gene family analysis, and estimated the mutation rate using MCMCtree with global clock. A Markov chain Monte Carlo (MCMC) process was run for 2,000,000 iterations, with sample frequency of 100 after a burn-in of 1,000 iterations. The median of simulated data was selected as mutation rate (mu = $1.73\times10^{-9}$ per base per year).

## 5 Genome-wide host-parasite protein interaction analysis

In addition to the genome data that we generated for F. gigantica, we downloaded genome annotation information for human (GCA_000001405.28), swamp buffalo



**Citation:** Xier Luo High-quality reference genome of Fasciola gigantica: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions https://dx.doi.org/10.17504/protocols.io.bxatpien

(GWHAAJZ00000000), *F. hepatica (*GCA_002763495.2), *Fasciolopsis buski (*GCA_008360955.1), *Clonorchis sinensis* (GCA_003604175.1), *Schistosoma mansoni* (GCA_000237925.2), and *Taenia multiceps* (GCA_001923025.3) from the NCBI database and BIG Sub (China National Center for Bioinformation, Beijing, China). Proteases and protease inhibitors were identified and classified into families using BLASTp (e-value < $10^{-4}$) against the MEROPS peptidase database (merops_scan.lib; (European Bioinformatics Institute (EMBL-EBI), Cambridge, UK)), with amino acids at least 80% coverage matched for database proteins. These proteases were divided into five major classes (aspartic, cysteine, metallo, serine, and threonine proteases). E/S proteins (i.e., the secretome) were predicted by the programs SignalP 5.0 (82), TargetP (83), and TMHMM (84). Proteins with a signal peptide sequence but without a transmembrane region were identified as secretome proteins, excluding the mitochondrial sequences. Genome-wide host-parasite protein interaction analysis was perform by constructing the PPIs between the *F. gigantica* secretome and human proteins expressed in the tissues related to the liver fluke life cycle. For the hosts, we selected human proteins expressed in the small intestine and liver, and located in the plasma membrane and extracellular region. The gene expression and subcellular location information were obtained from the TISSUES (85) and Uniprot (EMBL-EBI) databases, respectively. For *F. gigantica*, secretome molecules were mapped to the human proteome as the reference, using the reciprocal best-hit BLAST method. These two gene datasets were used to construct host-parasite PPI networks. We downloaded the interaction files (protein.links.v11.0) in the STRING database (86), and only highly credible PPIs were retained by excluding PPIs with confidence scores below 0.7. The final STRING network was plotted using Cytoscape (87).

## 6 Gene family analysis

We chose the longest transcript in the downloaded annotation dataset to represent each gene, and removed genes with open reading frames shorter than 150 bp. Gene family clustering was then performed using OrthoFinder (v2.3.12) (88), based on the predicted gene set for eight genomes, including *F. gigantica* (our assembly), *F. hepatica* (NCBI: GCA_002763495.2), *Fasciolopsis buski* (NCBI: GCA_008360955.1), *Clonorchis sinensis* (NCBI: GCA_003604175.1), *Schistosoma mansoni* (NCBI: GCF_000237925.1), *Taenia multiceps* (NCBI: GCA_001923025.3), swamp buffalo (BIG sub: GWHAAJZ00000000), and human (NCBI: GCF_000001405.39). This analysis yielded 17,992 gene families. To identify gene families that had undergone expansion or contraction, we applied the CAFE (v5.0.0) program (89), which inferred the rate and direction of changes in gene family size over a given phylogeny. Among the eight species, 559 single-copy orthologs were aligned using MUSCLE (v3.8.1551) (90), and we eliminated poorly aligned positions and divergent regions of the alignment using Gblock 0.91b (91). RAxML (v8.2.12) was then used with the PROTGAMMALGF model to estimate a maximum likelihood tree. Divergence times were estimated using PAML MCMCTREE (92). A Markov chain Monte Carlo (MCMC) process was run for 2,000,000 iterations, with a sample frequency of 100 after a burn-in of 1,000 iterations under an independent rates model. Two independent runs were performed to check the convergence. The fossil-calibrated eukaryote phylogeny was used to set the root height for the species tree, taken from the age of Animals (602–661 Ma) estimated in a previous fossil-calibrated eukaryotic phylogeny (93) and the divergence time between the euarchontoglires and laurasiatheria: (95.3–113 Ma) (94).



**Citation:** Xier Luo High-quality reference genome of Fasciola gigantica: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions https://dx.doi.org/10.17504/protocols.io.bxatpien