


Article

Chromosome Genome Assembly of *Cromileptes altivelis* Reveals Loss of Genome Fragment in *Cromileptes* Compared with *Epinephelus* Species

Yang Yang¹, Lina Wu¹, Zhuoying Weng¹, Xi Wu¹, Xi Wang¹, Junhong Xia^{1,2}, Zining Meng^{1,2} 
and Xiaochun Liu^{1,2,*}

¹ State Key Laboratory of Biocontrol, Life Sciences School, Sun Yat-sen University, Guangzhou 510275, China; yangy595@mail2.sysu.edu.cn (Y.Y.); wuln5@mail2.sysu.edu.cn (L.W.); wengzhy5@mail2.sysu.edu.cn (Z.W.); wuxi577@126.com (X.W.); wangx265@mail2.sysu.edu.cn (X.W.); xiajunh3@mail.sysu.edu.cn (J.X.); mengzn@mail.sysu.edu.cn (Z.M.)

² Southern Laboratory of Ocean Science and Engineering, Zhuhai 519000, China

* Correspondence: lsslx@mail.sysu.edu.cn

Abstract: The humpback grouper (*Cromileptes altivelis*), an Epinephelidae species, is patchily distributed in the reef habitats of Western Pacific water. This grouper possesses a remarkably different body shape and notably low growth rate compared with closely related grouper species. For promoting further research of the grouper, in the present study, a high-quality chromosome-level genome of humpback grouper was assembled using PacBio sequencing and high-throughput chromatin conformation capture (Hi-C) technology. The assembled genome was 1.013 Gb in size with 283 contigs, of which, a total of 143 contigs with 1.011 Gb in size were correctly anchored into 24 chromosomes. Moreover, a total of 26,037 protein-coding genes were predicted, of them, 25,243 (96.95%) genes could be functionally annotated. The high-quality chromosome-level genome assembly will provide pivotal genomic information for future research of the speciation, evolution and molecular-assisted breeding in humpback groupers. In addition, phylogenetic analysis based on shared single-copy orthologues of the grouper species showed that the humpback grouper is included in the *Epinephelus* genus and clustered with the giant grouper in one clade with a divergence time of 9.86 Myr. In addition, based on the results of collinearity analysis, a gap in chromosome 6 of the humpback grouper was detected; the missed genes were mainly associated with immunity, substance metabolism and the MAPK signal pathway. The loss of the parts of genes involved in these biological processes might affect the disease resistance, stress tolerance and growth traits in humpback groupers. The present research will provide new insight into the evolution and origin of the humpback grouper.

Keywords: humpback grouper; *Cromileptes altivelis*; genome; evolution; collinearity



Citation: Yang, Y.; Wu, L.; Weng, Z.; Wu, X.; Wang, X.; Xia, J.; Meng, Z.; Liu, X. Chromosome Genome Assembly of *Cromileptes altivelis* Reveals Loss of Genome Fragment in *Cromileptes* Compared with *Epinephelus* Species. *Genes* **2021**, *12*, 1873. <https://doi.org/10.3390/genes12121873>

Academic Editor: Anna Rita Rossi

Received: 29 October 2021

Accepted: 22 November 2021

Published: 24 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The humpback grouper (*Cromileptes altivelis*), also called the mouse grouper, belongs to the *Cromileptes* genus of Epinephelidae. The genus only consists of the one species due to its special morphology compared to other groupers. Humpback groupers are mainly distributed in Western Pacific water [1]. The natural population of the humpback grouper is decreasing all over the world due to overfishing, climate change and habitat destruction (IUCN, <https://www.iucnredlist.org/species/39774/100458943>, accessed on 20 May 2021). Underwater surveys carried out in many areas, including Indonesia, Australia and New Caledonia, suggest that the humpback grouper is extremely rare and patchily distributed in reef habitats. The humpback grouper is one of the most expensive species in all of the grouper species in the live marine fish trade; the price has reached 130 USD per kilogram owing to its tender meat, beautiful color, special shape and rarity (IUCN). Though the artificial culture of the humpback grouper has been successful in recent years [2], the price

is still high due to its low growth rate and low output of larva fish. Molecular-assisted selection based on the genome information is essential to increase the growth rate of humpback groupers in future breeding.

The humpback grouper possesses special morphology compared with *Epinephelus* species. It has an extremely small anterior part of the head and a moderately deep body which forms a concave dorsal profile on the postorbital part, hence why it is also called humpback grouper. Moreover, it possesses a subuliform head and basiconic jaw without canine teeth so it is also widely called a mouse grouper. Generally, grouper species possess 7–11 dorsal spines. Just four groupers possess 10 dorsal spines, aside from the humpback grouper, including *Epinephelus snalogus*, *Hyporthodus exsul* and *H. nigrinus*. Based on molecular phylogenetic analysis of mitochondrial and nuclear genes of groupers, the humpback grouper is separated from *Epinephelus* species [3–6]. However, the growth ratio of humpback groupers is far less than the most closely related *Epinephelus* species, such as *E. lanceolatus* and *E. fuscoguttatus*. The speciation and evolution of humpback groupers is a controversial issue. At present, no valid genome and transcriptome have been reported for the species. Lack of genetic resources seriously hinders the research about humpback groupers. Recently, several grouper species genomes, including red-spotted grouper (*E. akaara*) [7], giant grouper (*E. lanceolatus*) [8], leopard coral grouper (*Plectropomus leopardus*) [9], kelp grouper (*Epinephelus moara*) [10] and brown-marbled grouper (*Epinephelus fuscoguttatus*) [6], have been published. These high-quality genomes provide essential genetic resources for better understanding of the evolution and phylogeny of grouper species. In the present study, the assembly of the humpback grouper genome is important for further research on the speciation, evolution, molecular-assisted selection and the developmental mechanism of special shape.

2. Materials and Methods

2.1. Sample Collection, Library Construction and Sequencing

A humpback grouper with body weight of 183.0 g and total length of 25.4 cm (Figure 1) was collected from Chenhai Aquatic Co., Ltd. (Hainan, China). The fish was immediately dissected after anesthesia with MS-222. White muscle tissue in the dorsal was sampled and immediately stored in liquid nitrogen, which was used for genomic DNA sequencing and Hi-C library construction. Moreover, ten tissues, including skin, muscle, liver, kidney, brain, intestine, fat, spleen, heart and gill, were collected and stored in RNAlater for transcriptome sequencing.



Figure 1. The characteristic of humpback grouper (*Cromileptes altivelis*).

Total DNA was extracted from white muscle tissue with a TIANamp Marine Animals DNA Kit (Tiangen Biotech Co., Ltd., Beijing, China). Quality and quantity of total DNA were determined by NanoDrop 2000 (Thermo Fisher Scientific Inc., Waltham, MA, USA). A paired-end sequencing library with insert length of 350 bp was constructed using a TruSeq Nano DNA LT Library Preparation Kit (Illumina, San Diego, CA, USA). The obtained library was then sequenced on Illumina HiSeq X Ten platform.

Genome DNA was broken into fragments by Covaris and was recycled by AMPure PB beads (Pacific Biosciences, Menlo Park, CA, USA). A SMRTbell library was constructed using SMRTbell Template Prep Kit (Pacific Biosciences, USA), according to the manufacturer's instruction and was sequenced on PacBio Bioscience Sequel platform (Pacific Biosciences, USA).

Muscle samples were fixed with fresh paraformaldehyde and then DNA–protein bonds were created. The Mbo I restriction enzyme was used to digest the DNA and the overhanging 5' ends of the DNA fragments were repaired with a biotinylated residue. The fragments, closed to each other in the nucleus during fixation, were ligated and the denatured proteins were removed. The Hi-C fragments were further sheared by sonication and then pulled down with streptavidin beads. The library was sequenced on an Illumina HiSeq X Ten platform with PE150 strategy.

Total RNA of the 10 tissues (approximately 80 mg of each) was extracted using RNAiso reagents (Takara, Dalian, China), following the manufacturer's instructions. The quantity and quality of RNA samples were determined using a microplate spectrophotometer (BioTek Company, Winooski, VT, USA) and electrophoresis which was conducted using 1% agarose gel. Total RNA of the 10 tissues was mixed with equal amounts to generate a mixed RNA pool. An RNA-seq library was prepared using NEBNext Ultra™ RNA Library Prep Kit (NEB, USA), following the manufacturer's protocols. The library's quality and quantity were measured using Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Finally, the libraries were sequenced on Illumina-Hiseq 2000 platform with PE150 paired-end approach. A total of 119.33 Gb clean data was obtained with depth of 114.26× (Table 1).

Table 1. The statistics of sequencing and assembled information of humpback grouper genome.

Raw Data	Reads Number	Reads Base (bp)	N50 (bp)	Max Length (bp)	GC Content (%)
Illumina data for annotation	39,031,897	11,658,953,034	150	150	49.7
Illumina data for survey	1,307,931,088	196,189,663,200	150	150	41.2
PacBio data	6,672,321	119,331,383,944	27,957	224,636	41.0
HiC data	336,775,366	100,882,400,830	150	150	42.6
Assembled data	Contig or Scaffold number	Genome size (bp)			
Survey	-	~1,070,000,000	-	-	41.2
Contig assembled using PacBio	470	1,044,397,337	18,092,086	49,150,803	41.3
Contig assembled using PacBio+Hi-C	283	1,013,358,489	18,269,829	49,150,803	41.2
Scaffold	164	1,013,370,389		52,436,080	41.2
Chromosome	24	1,010,598,072	43,466,351	52,436,080	41.2

2.2. Genome Assembly

We estimated the main genome characteristics of the humpback grouper through k-mer frequency distribution analysis [11]. After filtering, the clean data were used to estimate genome size and heterozygosity using 19-mer.

For the PacBio sequencing data, after removing low-quality reads, clean data were corrected and assembled using Canu version 1.8 with parameters of genomeSize = 107,000,000, corOutCoverage = 80, corMhapSensitivity = low and correctedErrorRate = 0.025 [12]. Wtdbg was used to assemble genome by constructing fuzzy Bruijn graph (<https://github.com/ruanjue/wtdbg> accessed on 20 May 2021). Quickmerge [13] was used to merge assemblies produced by Canu and wtdbg to produce a more contiguous assembly with a parameter of -hco 5.0 -c 1.5 -l 100,000 -ml 5000. In simple terms, contigs from Canu as query input and contigs from wtdbg as reference input are aligned through mummer version 4.0 [14]. The assembled genome was polished using Pilon version 1.22 [15]. For estimating genome completeness, Illumina data were mapped back to the humpback grouper genome to calculate the mapping rate using BWA v0.7.17 [16]. The genome integrality was verified based on the Core Eukaryotic Genes Mapping Approach (CEGMA) database [17] and Benchmarking Universal Single-Copy Orthologs (BUSCO) database [18].

2.3. Pseudochromosome Construction

The sequencing data of Hi-C were filtered to remove low-quality reads using Fastp version 0.12.6 [19]. The clean reads were aligned to draft genome of humpback grouper using bowtie2 version 2.3.2 [20] with end-to-end model and a parameter of very-sensitive. The draft genome was broken into 50 Kb fragments and reassembled with correct cluster, order and orient of the contigs using Lachesis with parameters of CLUSTER_MIN_RE_SITES = 100; CLUSTER_MAX_LINK_DENSITY = 2; CLUSTER_NONINFORMATIVE_RATIO = 2; ORDER_MIN_N_RES_IN_TRUN = 15; and ORDER_MIN_N_RES_IN_SHREDS = 15 [21].

2.4. Repeat Annotation

Repetitive regions of the wild humpback grouper genome were identified by de novo and homology predictions. Transposable elements (TEs) were identified using LTR Finder version 1.05 [22], RepeatScout v1.0.5 (<http://www.repeatmasker.org> accessed on 20 March 2020) and PILER v1.0 [23]. The TEs were classified and annotated using PASTEC classifier version 1.0 using TEdenovo pipeline [24]. Coding sequences were removed from the predicted repeat sequences through alignment to the SwissProt database using blastx with e-value $< 1 \times 10^{-4}$, identity > 30 , coverage $> 30\%$, and length > 90 bp. RepeatMasker v4.0.5 [25] was used to identify repeats based on the RepBase library version 19.06 [26] of known transposable elements (TEs).

2.5. Genome Prediction and Annotation

Predicted non-coding RNAs includes micro RNA (miRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA). tRNA was predicted using tRNAscan-SE v1.3.1 [27]. microRNA and rRNA were predicted using Blast+ version 2.2.31 (https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download, accessed on 20 March 2021) with e-value $< 1 \times 10^{-5}$ based on the Rfam database (<ftp://ftp.ebi.ac.uk/pub/databases/Rfam> accessed on 20 May 2021) [28]. miRNAs were predicted using Infernal version 1.1.1 [29] based on miRBase version 21 [30].

Protein-coding genes were predicted using three methods, i.e., de novo prediction, homologous sequences prediction and RNA-seq-assisted methods. For de novo prediction, the genome without repeat regions was applied to generate gene structures using Genscan [31], Augustus version 2.4 [32], GlimmerHMM version 3.0.4 [33], GeneID version 1.4 [34] and SNAP version 2006-07-28 [35]. Four fish species genomes and their annotation files, including tilapia (*Oreochromis niloticus*, GCF_001858045.2), zebra fish (*Danio rerio*, GCF_000002035.6), large yellow croaker (*Larimichthys crocea*, GCF_000972845.2) and Atlantic salmon (*Salmo salar*, GCF_000233375.1) were downloaded from Genbank database. The genome of humpback grouper was aligned to these genomes downloaded from NCBI using GeMoMa [36] to obtain the exact exons, introns and splice sites [36]. RNA-seq data from ten tissues were aligned to the genome using HISAT2 version 2.0.4 [37] and were assembled using Stringtie version 1.2.3 [38]. Open reading frames (ORFs) were predicted using PASA version 2.0.2 [39], TransDecoder version 2.0 (<https://github.com/TransDecoder/TransDecoder>, accessed on 20 March 2021) and GeneMarkS-T version 5.1 [40]. The genes predicted by the three methods were merged using EVM (<http://evidencemodeler.sourceforge.net/>, accessed on 20 March 2021) [41].

The putative functions of predicted genes were annotated by aligning them to NR [42], KOG [43], GO [44], KEGG [45] and TrEMBL [46] databases using Blast+ version 2.2.31.

2.6. Evolution Analyses

For identifying gene family, eight fish species including seven grouper species (giant grouper (*E. lanceolatus*), orange-spotted grouper (*E. coioides*), brown-marbled grouper (*E. fuscoguttatus*), humpback grouper (*Cromileptes altivelis*), kelp grouper (*E. moara*), red-spotted grouper (*E. akaara*) and leopard coral grouper (*Plectropomus leopardus*)) and an outgroup large yellow croaker (*Larimichthys crocea*) were collected for comparison

using all-to-all BLAST with 1×10^{-5} of e-value. The orthogroups of eight species were predicted using Orthofinder version 2.3.7 [47].

The phylogenetic tree was constructed using shared single-copy genes among the eight fish species mentioned above. Protein sequences of these single-copy genes were aligned using MAFFT version 7.394 [48]. Gblocks version 0.91b [49] was used to remove low-quality alignments. The phylogenetic tree was constructed using RAxML software version 8.0.9 [50] and IQ-tree version 1.6.11 [51] with bootstrap value of 1000.

The genomic collinearity analyses of brown-marbled grouper and other groupers were carried out using MCScanX [52]. The divergence time at each tree node was predicted using MCMCtree in PAML package version 4.9 [53]. The two calibration times including *E. akaara* vs. *E. fuscoguttatus* with 15.2~28.5 Myr and *L. crocea* vs. *P. leopardus* with 99~127 Myr were obtained from the TimeTree database [54].

3. Results

3.1. Sequencing and Genome Assembly

Firstly, next-generation sequencing was applied to estimate the genome characteristics, including size, heterozygosity and repeat ratio, using the k-mer method. A total of 196.19 Gb clean data ($184\times$) was obtained, with Q20 of 96.80% and Q30 of 91.90% (Table 1). The main peak of 19-mer in frequency distribution was at a depth of $152\times$ (Figure 2A). The predicted genome size of the humpback grouper is 1.07 Gb with repeats of 29.53%. The heterozygosity and GC content were 0.09% and 41.2%, respectively.

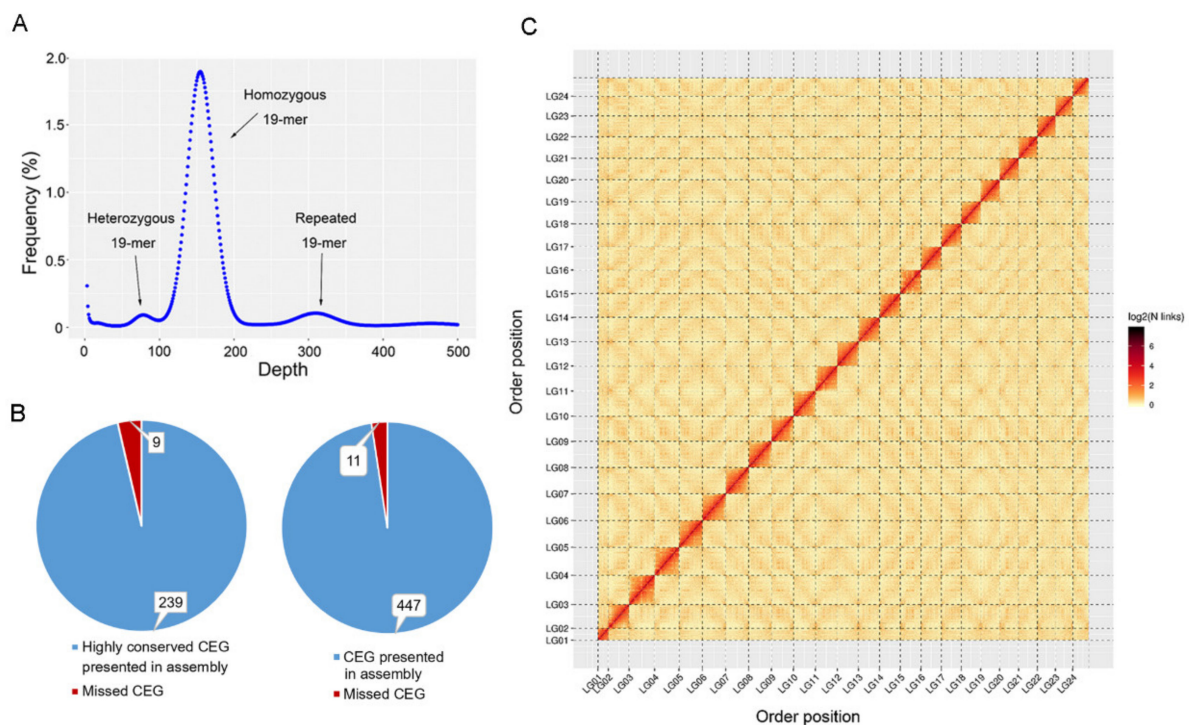


Figure 2. Genome assembly of humpback grouper. (A) Estimation of genome size, repeat content and heterozygous by survey using 19-mers in humpback grouper; (B) Verification of genome integrality based on the Core Eukaryotic Genes Mapping Approach (CEGMA) database; red parts and blue parts indicated missed and presented core eukaryotic genes (CEG) in humpback grouper genome, respectively; (C) The genome contig contact matrix; the blocks indicated the contacts between linkage groups; color depth indicated the degree of contacts.

For genome assembly, a total of 119.33 Gb PacBio data ($114.26\times$) and 100.88 Gb Hi-C data were obtained. The clean data included 6,672,321 reads with N50 of 27.96 Kb and an average length of 17.89 Kb (Table 1). A draft genome was assembled with 470 contigs with a length of 1.04 Gb and a contig N50 of 18.09 Mb using Canu and wtdbg (Table 1).

The genome quality was assessed by illumina data using BWA. The results showed that 98.29% of reads mapped into the draft genome, while 96.25% of reads properly mapped into the genome. The genome was aligned against the CEGMA database, which was constructed by a set of 458 core eukaryotic genes (CEGs) that are present in a wide range of taxa. A total of 447 (97.60%) CEGs were mapped (Figure 2B). The genome was then aligned to a reference gene set from the BUSCO database, which was constructed from 20 fish species, consisting of 4584 genes. A total of 4351 (94.92%) genes were mapped completely.

The Hi-C data was applied to correct misjoins, order and orientation in the draft genome. At last, a total of 283 contigs with 1,013,358,489 bp were obtained. Of them, 163 (57.60%) contigs were anchored into 24 pseudochromosomes with 1.012 Gb (99.82%) (Figure 2C). In the anchored contigs, 143 (87.73%) contigs with 1.011 Gb (99.91%) were correctly assembled into chromosomes (Table 1).

3.2. Annotation

A total of 1,618,118 repeat sequences were predicted with 380.56 Mb, which account for 37.55% of the humpback grouper genome (Table S1). Of them, 2184 microsatellite sequences with 1,738,061 bp were detected, which account for 0.17% of the humpback grouper genome.

For non-coding RNAs, a total of 490 miRNA, 530 rRNA and 1239 tRNA were predicted.

For protein-coding genes, a total of 26,037 genes were predicted (Table S2); the total length and average length of these genes were 443.64 Mb and 17,038.97 bp. There were 252,243 exons, 226,206 introns and 245,667 coding sequences in the whole genome (Table S3). A total of 25,243 (96.95%) genes were aligned into at least one database. A total of 12,734 (48.91%), 15,086 (57.94%), 16,598 (63.75%), 24,908 (95.66%) and 25,218 (96.85%) genes were mapped into the GO, KEGG, KOG, TrEMBL and NR databases, respectively.

3.3. Evolution analyses

The orthogroups of eight fish species were predicted using Orthofinder software. A total of 23,448 orthogroups were detected in all species. A total of 20,971 orthogroups were predicted in the humpback grouper. Of them, 25 species-specific orthogroups and 65 species-specific genes were detected (Figure 3A). Parts of these genes could be annotated and associated with myofiber structure and immunization. The number of shared single-copy orthogroups is 5066, which was applied in the establishment of the phylogenetic tree.

The phylogenetic tree from RAxML software was consistent with that from the IQ-tree software (Figures 3B and S1): the humpback grouper and other *Epinephelus* species clustered into one branch with high bootstrap value (≥ 85). The giant grouper is the most closely related species to the humpback grouper with the shortest divergence time (3.22~16.30 Mya) compared to other grouper species. Then, the two groupers were clustered with the orange-spotted grouper, brown-marbled grouper, kelp grouper, red-spotted grouper and leopard coral grouper in turn (Figure 2B). The divergence time between the humpback grouper and all the *Epinephelus* species was less than 16.55 Mya, while the divergence time between *Epinephelus* species and the leopard coral grouper reached 16.20~92.98 Mya (Figure 3C). The results suggested that the humpback grouper may diverge from *Epinephelus* species.

The collinearity analyses were carried out using MCScanX. There was high collinearity between the humpback grouper and giant grouper (Figure 4A). However, there was a gap in chromosome 6 of the humpback grouper compared to the most closely related *Epinephelus* species (Figures 4A and S2). Most of the genes located in 29,818,799~43,791,507 of the giant grouper genome were loosed in the humpback grouper (Figure 4B). Based on KEGG enrichment analysis (Figure 4C), the missed genes were mainly involved in immunity (IL1s, KRABs and JAK), substance metabolism (UGTs and GSTs) and MAPK signal transduction (CACNAs, FGF, EREG, MAPK1/3, MAPKAPK5 and MKP).

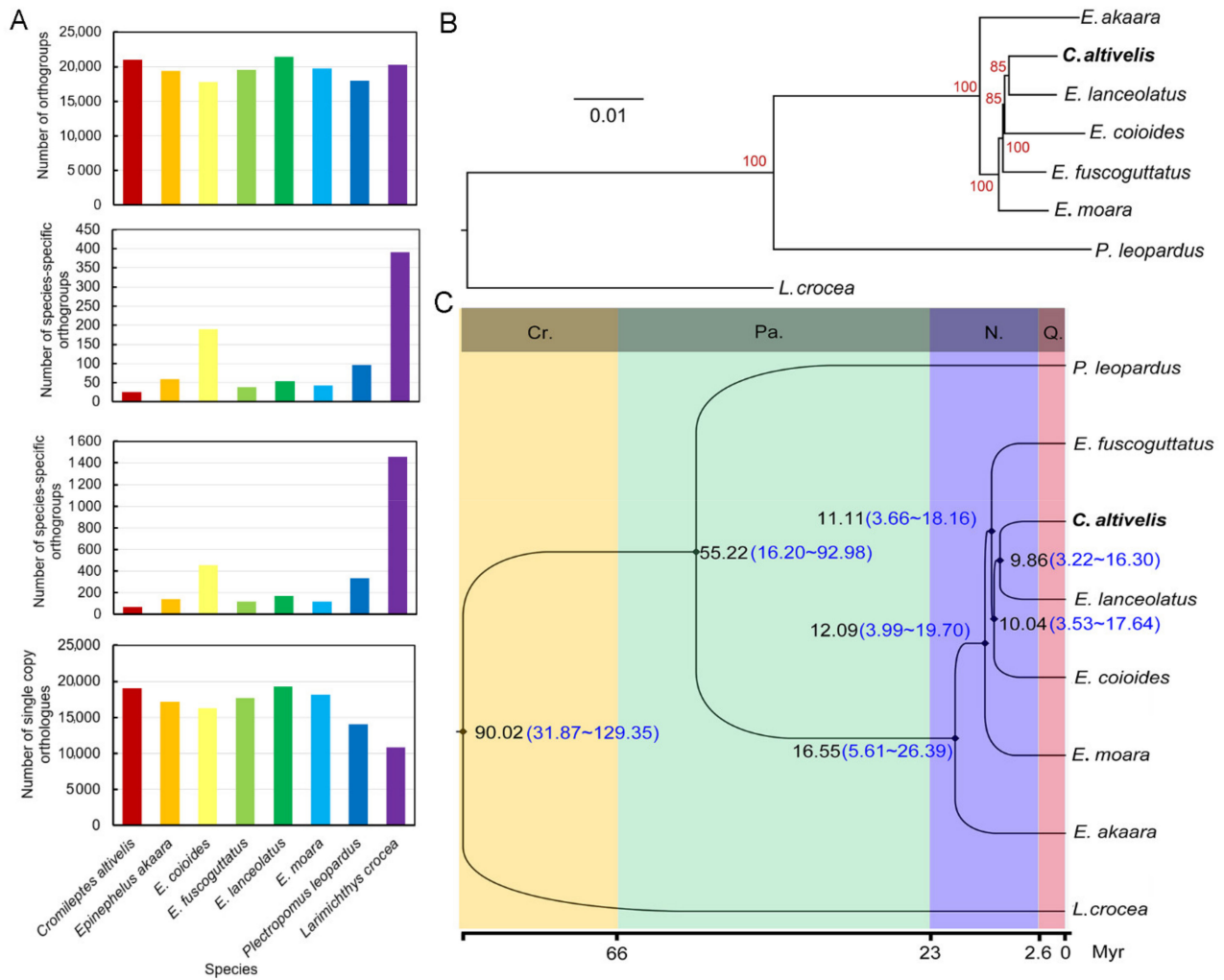


Figure 3. Analysis of divergence time of humpback grouper and other grouper species. Large yellow croaker (*L. crocea*) was set as outgroup. (A) The orthogroups statistics of the eight species; (B) phylogenetic tree of humpback grouper and other grouper species, red number indicated the bootstrap values; (C) Analysis of divergence time between humpback grouper and other grouper species using MCMCtree in PAML, the bold word indicated fish species used in the present study; two calibration times were obtained from TimeTree database; black number indicated the mean value of divergence time, blue number indicated the 95% confidence intervals of divergence times; Cr.—Cretaceous, Pa.—Paleogene, N.—Neogene.

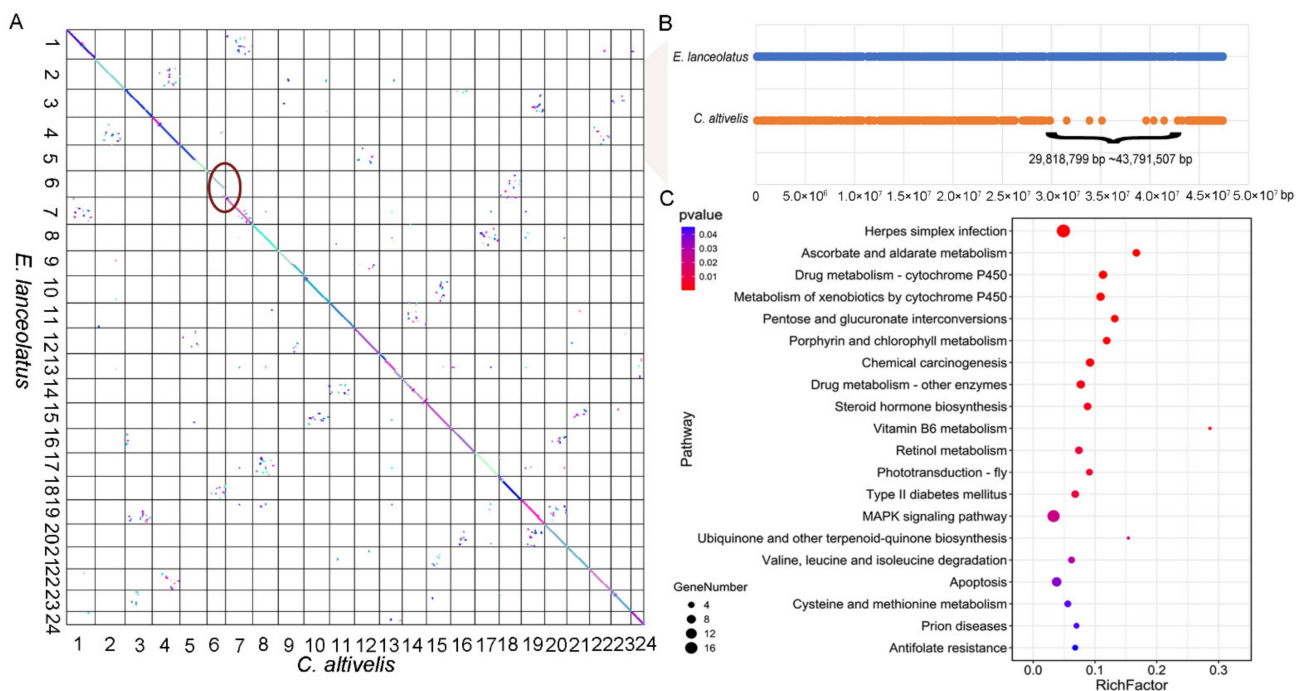


Figure 4. Collinearity analysis between humpback grouper and giant grouper. (A) Scatterplot of collinearity between humpback grouper and giant grouper, red circle showed the missing fragment in chromosome 6 of humpback grouper genome; (B) The distribution of genes in chromosome 6 of giant grouper and humpback grouper, multiple genes located in located in 29,818,799~43,791,507 bp of chromosome 6 were missed. (C) The KEGG enrichment analysis of missed genes.

4. Discussion

The humpback grouper, the only species in the *Cromileptes* genus of Epinephelidae, possesses special morphologies compared with other groupers. Thus, its speciation, evolution and phylogenetic position has attracted much attention. The high-quality chromosome-level genome of the humpback grouper provides an important foundation for the analyses of its origin and evolution.

In the research, a chromosome-level genome of the humpback grouper was assembled using PacBio sequencing and high-throughput chromatin conformation capture (Hi-C) technology. The heterozygosity of the genome was 0.09%. Its heterozygosity is lower than other grouper species, such as 0.375% in red-spotted grouper [7], 0.42% in leopard coral grouper [9] and 0.35% in brown-marbled grouper [6], which implies that the genetic diversity of the humpback is lower than other grouper species. The assembled genome was 1.013 Gb with 283 contigs, of which, a total of 143 contigs with 1.011 Gb in size were correctly anchored into 24 chromosomes. The genome size is similar with groupers in *Epinephelus*, such as 1.054 Gb in giant grouper [8], 1.047 Gb in brown-marbled grouper [6], 1.135 Gb in red-spotted grouper [7] and 1.08 Gb in kelp grouper [10].

The percentage of repeat contents in the humpback grouper was 37.55%, which was slightly higher than 34.6% of leopard coral grouper [9] and lower than 43.02% of red-spotted grouper [7], 41.1% of giant grouper [8] and 43.16% of brown-marbled grouper [6]. For protein-coding genes, a total of 26,037 protein-coding genes were predicted, of them, 25,243 (96.95%) genes could be functionally annotated. The results showed a high annotation rate.

The specific genes were predicted using Orthofinder software. A total of 65 specific genes in the humpback grouper were obtained. These genes were mainly associated with myofiber structure and immunity, which may be involved in the growth and disease resistance of the humpback grouper. A total of 5066 shared single-copy orthogroups were obtained in the seven grouper species and the outgroup, which was higher than other research that used more outgroups [6,7,9]. The application of more single-copy orthogroups made the phylogenetic analysis more accurate. Based on the phylogenetic

analyses of the humpback grouper and other *Epinephelus* species clustered into one branch with high bootstrap value, the giant grouper is the most closely related species of humpback grouper with a divergence time of 3.22~16.30 Mya. The divergence time between the humpback grouper and all the *Epinephelus* species was less than 16.55 Mya. Similar conclusions have been reported in several studies. For example, in the phylogenetic tree constructed by mitochondrial genome, the humpback grouper was clustered with the giant grouper as one clade, then clustered with all *Epinephelus* species as one clade [3]; in the phylogenetic tree based on cytochrome b gene, the humpback grouper was also clustered with the *Epinephelus* species as one clade [4]; and in the phylogenetic tree constructed by single-copy orthogroups, the humpback grouper was firstly clustered with the brown-marbled grouper, then was clustered with *Epinephelus* species [6]. Though there were slight differences in topological structure, these results confirmed that *Cromileptes* originated from the *Epinephelus* genus.

There was high collinearity between the genomes of the humpback grouper and *Epinephelus* species. However, we found a gap in chromosome 6 which spanned 29,818,799~43,791,507 bp in the humpback grouper genome compared to the giant grouper. The result was also observed in the collinearity analysis between the brown-marbled grouper and humpback grouper [6]. Missed genes in the gap were mainly involved in immunity, substance metabolism and the MAPK signal pathway. Recently published research also reported that there were expansions of gene families involved in immunity in the brown-marbled grouper compared with the humpback grouper [6]. The change of genes involving immunity might induce a difference in disease resistance between the two species. For missed genes involved in immunity, interleukin 1s (IL1s) play important roles in innate inflammation through stimulating thymocyte proliferation and B-cell maturation and proliferation [55]. Tyrosine protein kinase (JAK) is involved in various processes such as metabolism, immunity and cell cycle control [56–59]. For substance metabolism, glucuronosyltransferases (UGTs) are essential factors in the elimination and detoxification of drugs and metabolism of endogenous and xenobiotics substances [60–62]; glutathione S transferases (GST) as detoxification enzymes could catalyze a combination of glutathione with electrophilic groups of substances, which play important roles in resistance to drugs, environmental pollutants and reactive oxygen species [63]. The MAPK signal pathway plays an important role in the regulation of cell proliferation [64]. The loss of the parts of genes involved in these biological processes might affect the disease resistance, stress tolerance and growth traits in the humpback grouper. In the missed fragment, the gene directly associated with morphology was not detected. The gene involved in the form of the humpback in the humpback grouper still need to be explored.

5. Conclusions

In the research, a high-quality chromosome-level genome of humpback grouper was assembled using PacBio sequencing and high-throughput chromatin conformation capture (Hi-C) technology, which will provide pivotal genomic information for future research of speciation, evolution and molecular-assisted breeding in humpback groupers. In addition, phylogenetic analysis, based on single-copy orthologues of grouper species, showed that the humpback grouper is included in the *Epinephelus* genus and clustered with the giant grouper to one clade with a divergence time of 9.86 Myr. Moreover, a gap in chromosome 6 of the humpback grouper was detected based on collinearity analysis; the missing genes were mainly associated with immunity, substance metabolism and the MAPK signal pathway. The loss of the parts of genes involved in these biological processes might affect the disease resistance, stress tolerance and growth traits in the humpback grouper.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/genes12121873/s1>, Figure S1: The phylogenetic tree of humpback grouper and other groupers using IQ-tree software, Figure S2: Collinearity analysis of *Cromileptes altivelis* and *Epinephelus coioides*, Table S1: The statistics of repeat information in humpback grouper genome, Table S2: Gene prediction of humpback grouper, Table S3: Construction statistics of predicted genes in humpback grouper genome.

Author Contributions: Y.Y., Z.M. and X.L. designed the study. Y.Y., X.W. (Xi Wang) and L.W. collected the samples. Y.Y., X.W. (Xi Wu) and Z.W. performed the laboratory work. Y.Y. and L.W. performed the analyses. Y.Y., Z.M., J.X. and X.L. drafted and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Agriculture Research System of MOF and MARA (CARS-47), National Natural Science Foundation of China (31872572), Yang Fan Innovative and Entrepreneurial Research Team Project (No.201312H10), Huizhou Swan Project (20170214023102296) and Science and Technology Planning of Guangzhou (201804020013).

Data Availability Statement: The chromosome-level genome assembly of the humpback grouper was deposited in the GenBank database with Whole Genome Shotgun projects: JAAWWW000000000. The raw data of NGS sequencing, Pacbio, Hi-C and RNA-seq for genome assembly of the humpback grouper were reserved in the GSA (Genome Sequence Archive) database and the accession numbers were CRA002771, CRA002772, CRA002773 and CRA002774.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Froese, R.; Pauly, D. FishBase. World Wide Web Electronic Publication; Version (12/2019). 2019. Available online: www.fishbase.org (accessed on 20 March 2021).
2. Shapawi, R.; Mustafa, S.; Ng, W.K. Effects of dietary carbohydrate source and level on growth, feed utilization, and body composition of the humpback grouper, *Cromileptes altivelis* (Valenciennes). *J. Appl. Aquac.* **2011**, *23*, 112–121. [[CrossRef](#)]
3. Zhuang, X.; Qu, M.; Zhang, X.; Ding, S. A Comprehensive Description and Evolutionary Analysis of 22 Grouper (Perciformes, *Epinephelidae*) Mitochondrial Genomes with Emphasis on Two Novel Genome Organizations. *PLoS ONE* **2013**, *8*, e73561. [[CrossRef](#)]
4. Ding, S.; Zhuang, X.; Guo, F.; Wang, J.; Su, Y.; Zhang, Q.; Li, Q. Molecular phylogenetic relationships of China Seas groupers based on cytochrome b gene fragment sequences. *Sci. China Ser. C Life Sci.* **2006**, *49*, 235–242. [[CrossRef](#)]
5. Ma, K.Y.; Craig, M.T.; Choat, J.H.; van Herwerden, L. The historical biogeography of groupers: Clade diversification patterns and processes. *Mol. Phylogenet. Evol.* **2016**, *100*, 21–30. [[CrossRef](#)]
6. Yang, Y.; Wang, T.; Chen, J.; Wu, L.; Wu, X.; Zhang, W.; Luo, J.; Xia, J.; Meng, Z.; Liu, X. Whole-genome sequencing of brown-marbled grouper (*Epinephelus fuscoguttatus*) provides insights into adaptive evolution and growth differences. *Mol. Ecol. Resour.* **2021**. [[CrossRef](#)] [[PubMed](#)]
7. Ge, H.; Lin, K.; Shen, M.; Wu, S.; Wang, Y.; Zhang, Z.; Wang, Z.; Zhang, Y.; Huang, Z.; Zhou, C.; et al. De novo assembly of a chromosome-level reference genome of red-spotted grouper (*Epinephelus akaara*) using nanopore sequencing and Hi-C. *Mol. Ecol. Resour.* **2019**, *19*, 1461–1469. [[CrossRef](#)] [[PubMed](#)]
8. Zhou, Q.; Gao, H.; Zhang, Y.; Fan, G.; Xu, H.; Zhai, J.; Xu, W.; Chen, Z.; Zhang, H.; Liu, S.; et al. A chromosome-level genome assembly of the giant grouper (*Epinephelus lanceolatus*) provides insights into its innate immunity and rapid growth. *Mol. Ecol. Resour.* **2019**, *19*, 1322–1332. [[CrossRef](#)] [[PubMed](#)]
9. Yang, Y.; Wu, L.N.; Chen, J.F.; Wu, X.; Xia, J.H.; Meng, Z.N.; Liu, X.C.; Lin, H.R. Whole-genome sequencing of leopard coral grouper (*Plectropomus leopardus*) and exploration of regulation mechanism of skin color and adaptive evolution. *Zool. Res.* **2020**, *41*, 328–340. [[CrossRef](#)]
10. Zhou, Q.; Gao, H.; Xu, H.; Lin, H.; Chen, S. A Chromosomal-scale Reference Genome of the Kelp Grouper *Epinephelus moara*. *Mar. Biotechnol.* **2021**, *23*, 12–16. [[CrossRef](#)] [[PubMed](#)]
11. Liu, B.; Shi, Y.; Yuan, J.; Galaxy, Y.; Zhang, H.; Li, N.; Li, Z.; Chen, Y.; Mu, D.; Parkin, I. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* **2013**, arXiv:1308.2012.
12. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive κ -mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736. [[CrossRef](#)]
13. Chakraborty, M.; Baldwin-Brown, J.G.; Long, A.D.; Emerson, J.J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **2016**, *44*, e147. [[CrossRef](#)] [[PubMed](#)]
14. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. *Genome Biol.* **2004**, *5*, R12. [[CrossRef](#)]

15. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K.; et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **2014**, *9*, e112963. [[CrossRef](#)]
16. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595. [[CrossRef](#)]
17. Parra, G.; Bradnam, K.; Korf, I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **2007**, *23*, 1061–1067. [[CrossRef](#)] [[PubMed](#)]
18. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [[CrossRef](#)] [[PubMed](#)]
19. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [[CrossRef](#)]
20. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [[CrossRef](#)]
21. Burton, J.N.; Adey, A.; Patwardhan, R.P.; Qiu, R.; Kitzman, J.O.; Shendure, J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **2013**, *31*, 1119–1125. [[CrossRef](#)]
22. Xu, L.; Zhang, Y.; Su, Y.; Liu, L.; Yang, J.; Zhu, Y.; Li, C. Structure and evolution of full-length LTR retrotransposons in rice genome. *Plant Syst. Evol.* **2010**, *287*, 19–28. [[CrossRef](#)]
23. Edgar, R.C.; Myers, E.W. PILER: Identification and classification of genomic repeats. *Bioinformatics* **2005**, *21*, i152–i158. [[CrossRef](#)]
24. Hoede, C.; Arnoux, S.; Moisset, M.; Chaumier, T.; Inizan, O.; Jamilloux, V.; Quesneville, H. PASTEC: An automatic transposable element classification tool. *PLoS ONE* **2014**, *9*, e91929. [[CrossRef](#)]
25. Tarailo-Graovac, M.; Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **2009**, 11–14. [[CrossRef](#)]
26. Bao, W.; Kojima, K.K.; Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **2015**, *6*, 11. [[CrossRef](#)]
27. Lowe, T.M.; Eddy, S.R. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **1996**, *25*, 955–964. [[CrossRef](#)] [[PubMed](#)]
28. Lagesen, K.; Hallin, P.; Rødland, E.A.; Stærfeldt, H.H.; Rognes, T.; Ussery, D.W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **2007**, *35*, 3100–3108. [[CrossRef](#)] [[PubMed](#)]
29. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. [[CrossRef](#)] [[PubMed](#)]
30. Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. MiRBase: From microRNA sequences to function. *Nucleic Acids Res.* **2019**, *47*, D155–D162. [[CrossRef](#)] [[PubMed](#)]
31. Burge, C.; Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **1997**, *268*, 78–94. [[CrossRef](#)]
32. Stanke, M.; Keller, O.; Gunduz, I.; Hayes, A.; Waack, S.; Morgenstern, B. AUGUSTUS: A b initio prediction of alternative transcripts. *Nucleic Acids Res.* **2006**, *34*, W435–W439. [[CrossRef](#)]
33. Majoros, W.H.; Pertea, M.; Salzberg, S.L. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **2004**, *20*, 2878–2879. [[CrossRef](#)]
34. Alioto, T.; Blanco, E.; Parra, G.; Guigó, R. Using geneid to Identify Genes. *Curr. Protoc. Bioinform.* **2018**, *64*. [[CrossRef](#)]
35. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **2004**, *5*, 59. [[CrossRef](#)]
36. Slater, G.S.C.; Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **2005**, *6*, 31. [[CrossRef](#)] [[PubMed](#)]
37. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [[CrossRef](#)] [[PubMed](#)]
38. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562–578. [[CrossRef](#)]
39. Campbell, M.A.; Haas, B.J.; Hamilton, J.P.; Mount, S.M.; Robin, C.R. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genom.* **2006**, *7*, 327. [[CrossRef](#)] [[PubMed](#)]
40. Tang, S.; Lomsadze, A.; Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **2015**, *43*, e78. [[CrossRef](#)]
41. Haas, B.J.; Salzberg, S.L.; Zhu, W.; Pertea, M.; Allen, J.E.; Orvis, J.; White, O.; Robin, C.R.; Wortman, J.R. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **2008**, *9*, R7. [[CrossRef](#)]
42. Marchler-Bauer, A.; Lu, S.; Anderson, J.B.; Chitsaz, F.; Derbyshire, M.K.; DeWeese-Scott, C.; Fong, J.H.; Geer, L.Y.; Geer, R.C.; Gonzales, N.R.; et al. CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **2011**, *39*, D225–D229. [[CrossRef](#)]
43. Koonin, E.V.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Krylov, D.M.; Makarova, K.S.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; Rao, B.S.; et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **2004**, *5*, R7. [[CrossRef](#)] [[PubMed](#)]
44. Dimmer, E.C.; Huntley, R.P.; Alam-Faruque, Y.; Sawford, T.; O'Donovan, C.; Martin, M.J.; Bely, B.; Browne, P.; Chan, W.M.; Eberhardt, R.; et al. The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.* **2012**, *40*, D565–D570. [[CrossRef](#)] [[PubMed](#)]

45. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462. [[CrossRef](#)]
46. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.C.; Estreicher, A.; Gasteiger, E.; Martin, M.J.; Michoud, K.; O'Donovan, C.; Phan, I.; et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365–370. [[CrossRef](#)] [[PubMed](#)]
47. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **2019**, *20*, 1–14. [[CrossRef](#)]
48. Rozewicki, J.; Li, S.; Amada, K.M.; Standley, D.M.; Katoh, K. MAFFT-DASH: Integrated protein sequence and structural alignment. *Nucleic Acids Res.* **2019**, *47*, W5–W10. [[CrossRef](#)] [[PubMed](#)]
49. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **2000**, *17*, 540–552. [[CrossRef](#)] [[PubMed](#)]
50. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [[CrossRef](#)] [[PubMed](#)]
51. Nguyen, L.T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [[CrossRef](#)] [[PubMed](#)]
52. Wang, Y.; Tang, H.; Debarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.H.; Jin, H.; Marler, B.; Guo, H.; et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **2012**, *40*, e49. [[CrossRef](#)]
53. Yang, Z. Paml: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **1997**, *13*, 555–556. [[CrossRef](#)]
54. Kumar, S.; Stecher, G.; Suleski, M.; Hedges, S.B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **2017**, *34*, 1812–1819. [[CrossRef](#)] [[PubMed](#)]
55. Dinarello, C.A. Overview of the IL-1 family in innate inflammation and acquired immunity. *Immunol. Rev.* **2018**, *281*, 8–27. [[CrossRef](#)] [[PubMed](#)]
56. Saltzman, A.; Stone, M.; Franks, C.; Searfoss, G.; Munro, R.; Jaye, M.; Ivashchenko, Y. Cloning and characterization of human Jak-2 kinase: High mRNA expression in immune cells and muscle tissue. *Biochem. Biophys. Res. Commun.* **1998**, *246*, 627–633. [[CrossRef](#)]
57. Jäkel, H.; Weinl, C.; Hengst, L. Phosphorylation of p27Kip1 by JAK2 directly links cytokine receptor signaling to cell cycle control. *Oncogene* **2011**, *30*, 3502–3512. [[CrossRef](#)] [[PubMed](#)]
58. Berry, D.C.; Jin, H.; Majumdar, A.; Noy, N. Signaling by vitamin A and retinol-binding protein regulates gene expression to inhibit insulin responses. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4340–4345. [[CrossRef](#)] [[PubMed](#)]
59. Villarino, A.V.; Kanno, Y.; Ferdinand, J.R.; O'Shea, J.J. Mechanisms of Jak/STAT signaling in immunity and disease. *J. Immunol.* **2015**, *194*, 21–27. [[CrossRef](#)]
60. Alonen, A.; Finel, M.; Kostianen, R. The human UDP-glucuronosyltransferase UGT1A3 is highly selective towards N2 in the tetrazole ring of losartan, candesartan, and zolarsartan. *Biochem. Pharmacol.* **2008**, *76*, 763–772. [[CrossRef](#)]
61. Miley, M.J.; Zielinska, A.K.; Keenan, J.E.; Bratton, S.M.; Radominska-Pandya, A.; Redinbo, M.R. Crystal structure of the cofactor-binding domain of the human phase II drug-metabolism enzyme UDP-glucuronosyltransferase 2B7. *J. Mol. Biol.* **2007**, *369*, 498–511. [[CrossRef](#)]
62. Perreault, M.; Gauthier-Landry, L.; Trottier, J.; Verreault, M.; Caron, P.; Finel, M.; Barbier, O. The Human UDP-glucuronosyltransferase UGT2A1 and UGT2A2 enzymes are highly active in bile acid glucuronidation. *Drug Metab. Dispos.* **2013**, *41*, 1616–1620. [[CrossRef](#)] [[PubMed](#)]
63. Hayes, J.D.; Pulford, D.J. The glutathione S-transferase supergene family: Regulation of GST and the contribution of the isoenzymes to cancer chemoprotection and drug resistance part I. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 445–520. [[CrossRef](#)] [[PubMed](#)]
64. Wei, Z.; Liu, H.T. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* **2002**, *12*, 9–18. [[CrossRef](#)]