Chromosome-level genome assembly of a butterflyfish, *Chelmon rostratus*

Xiaoyun Huang[1,2,3,*], Yue Song[1,2,3,*], Suyu Zhang[1,2,3,*], Yunga A[1,2,3,*], Mengqi Zhang[1,2,3], Yue Chang[1,2,3], He Zhang[1,2,3], Chang Li[1,2,3], Yong Zhao[1,2,3], Meiru Liu[1,2,3], Inge Seim[4,5], Guangyi Fan[1,2,3], Xin Liu[1,2,3,6,#], Shanshan Liu[1,2,3,#]

[1] BGI-Qingdao, BGI-Shenzhen, Qingdao, 266555, China;
[2] BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;
[3] China National Gene Bank, BGI-Shenzhen, Jinsha Road, Shenzhen 518120, China;
[4] Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal University, Nanjing 210023, China;
[5] Comparative and Endocrine Biology Laboratory, Translational Research Institute-Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology, Brisbane 4102, Queensland, Australia;
[6] James D. Watson Institute of Genome Science, 310008 Hangzhou, China;

[*] These authors contributed equally to this work
[#] These authors jointly supervised this work
Corresponding to liushanshan@genomics.cn (S.L.) and liuxin@genomics.cn (X.L.)

Abstract

Chelmon rostratus (Teleostei, Perciformes, Chaetodontidae) is a copperband butterflyfish. As an ornamental fish, the genome information for this species might help understanding the genome evolution of Chaetodontidae and adaptation/evolution of coral reef fish.

In this study, using the stLFR linked-read data, we assembled a genome of 638.70 Mb in size with contig and scaffold N50 sizes of 294.41 kb and 2.61 Mb, respectively. 94.40% of scaffold sequences were assigned to 24 chromosomes using Hi-C data and BUSCO analysis showed that 97.3% (2,579) of core genes were found in our assembly. Up to 21.47 % of the genome was found to be repetitive sequences and

21,375 protein-coding genes were annotated. Among these annotated protein-coding genes, 20,163 (94.33%) proteins were assigned with possible functions.
As the first genome for Chaetodontidae family, the information of these data helpfully to improve the essential to the further understanding and exploration of marine ecological environment symbiosis with coral and the genomic innovations and molecular mechanisms contributing to its unique morphology and physiological features.
Keywords: Chaetodontidaes; *Chelmon rostratus*; stLFR; Genome assembly; Gene annotation; Phylogenetic tree

Background & Summary

Chelmon rostratus (Teleostei, Perciformes, Chaetodontidae) lives in the western pacific, from southern Japan and Taiwan throughout the Coral Triangle to the Solomon Islands and the northern coast of Australia [1]. In the natural environment, it lives in depths from 1- 25 meters underwater, inhabiting coastal and inner reefs and often in turbid water [2]. It usually has an appealing appearance of four yellowish or yellowish-orange vertical bands with black border on the silvery white body, and a false eyespot. Thus, it is one of the protagonists in tropical marine aquarium fish [3].
Despite as a common coral reef fish, there were limited previous researches in this species. Previous researches were limited to its behavior features such as the prey capture kinematics [4]. More recently, its complete mitochondrial genome was reported [5], revealing its evolutionary position. However, there was no particular study providing the whole genomic information.
Here, we reported the first whole genome sequencing data for this species, resulting in a chromosome level genome assembly and annotation, followed by the evolutionary analysis. The data and analysis provided here, can benefit future basic studies and conservation efforts of this species.

Methods

Sample collection, library construction and sequencing

We captured an adult *Chelmon rostratus* (Figure 1) in the sea area of Qingdao, China, and muscle tissue was stored in liquid nitrogen. Then, high quality DNA was extracted using a modified DNA extraction for vertebrate tissues protocol [6] from the muscle tissue. The exacted DNA was fragmented and subjected to paired-end sequencing library construction, following single tube Long Fragment Read (stLFR) protocol [7]. For Hi-C library sequencing, about 1g living muscle tissue was used to DNA extraction and library contraction, according to Wang's method [7]. Sequencing was conducted on a BGISEQ-500 sequencer, generating 286.82Gb raw data (including 134.18Gb raw Hi-C data) (Supplementary Table 1). Data filtering was then carried out using SOAPnuke software (version 1.5) [8] with the default parameters, thus low-quality reads (more than 40% bases with quality score lower than 8), PCR duplications, adaptors and reads with high proportion (higher than 10%) of ambiguous bases (Ns) were filtered. After data filtering, 192.71 Gb 'clean data' (including 133.17Gb 'clean Hi-C data') was obtained for the further assembly (Supplementary Table 1).

Genome features revealed by *k*-mer analysis

In order to understand better about its genome features, we applied *k*-mer frequency distribution analysis to estimate its genome size and genome complexity (heterozygosity or/and repetition) [9]. We randomly selected 59.54 Gb (~90X) clean reads and carried out a 17-mer analysis using KMERFREQ_AR (version 2.0.4) [10]. The genome size was estimated to be ~711.39 Mb (Supplementary Table 2). Observing the distribution of 17-mer frequency (Supplementary Figure 1), we anticipated the genome to be a diploid species with slight heterozygosity proportion. In order to get the percentage of heterozygosity, GCE software [9] was used to calculation with 59.54 Gb clean reads and resulted that the percentage of heterozygosity was 0.72%.

Genome assembly and annotation

We assembled the genome using Supernova (version 2.1) [11] with default parameters

for 55 Gb clean stLFR data. After that, we used Gapcloser software [10] to fill the gaps within the assembly with default parameters. Finally, we obtained an assembly of 638.70 Mb in size containing 5,490 scaffolds. The N50 values of contigs and scaffolds were 294.41 Kb and 2.61 Mb (Table 1), respectively, revealing a good contiguity of the genome assembly. The longest scaffold and contig were 9.02 Mb and 1.96 Mb, respectively. The assembled length was account for ~90 % of estimated genome size.

To generate a chromosomal-level genome assembly, 133.17 Gb high-quality Hi-C data was used to further assembly. We first used HiC-Pro software (version2.8.0_devel) with default parameters to get ~26 Gb valid sequencing data, accounting for 19.31% of total Hi-C clean reads. Then, Juicer (version 1.5, an opensource tool for analyzing Hi-C datasets) [12], and the 3D de novo assembly pipeline [13] were used to connect the scaffolds to chromosomes with length of 603Mb (account for 94.40% of total genome) and scaffold sequences were assigned to 24 chromosomes, with the length from 14.13 Mb to 33.19 Mb (Table 2, Figure 2A and Figure 3).

We randomly selected 4 Gb clean data to assemble mitochondrial genome sequence using MitoZ software [14], resulting 16.52 kb of assembly with a cyclic structure (Figure 2B).

Repetitive sequence and gene annotation

We annotated the two major types of repetitive sequences (tandem repeats, TRFs and transposable elements, TEs) in the assembled genome. According to the method of previously research [15-17], TRFs were identified using Tandem Repeats Finder (version 4.04) [18]. Transposable elements (TEs) were identified by a combination of homology-based and *de novo* approaches. Briefly, for homology-based annotation, known repeats in the database (Repbase16.02) [19] were aligned against the genome assembly using RepeatMasker and RepeatProteinMask (version 3.2.9) [20] at both the DNA and protein levels. For *de novo* annotation, RepeatModeler (version 1.1.0.4) [21] was employed to build a *de novo* non-redundant repeat library and then this repeat library was searched against the genome using RepeatMasker [20]. In this way, up to

21.47 % of the assembled sequences were found to be repeat sequences (Figure 3, Supplementary Table 3).

Protein-coding gene were then predicted by a combination of two ways: (1) the *ab initio* gene prediction and (2) the homology-based annotation [22-24]. For *ab initio* gene prediction approaches, Augustus [25] and GlimmerHMM [26] were used with *Danio rerio* as the species of HMM model to predict gene models; For homology-based annotation, four homolog species including *Pundamilia nyererei*, *Maylandia zebra*, *Astatotilapia calliptera* and *Perca flavescens* were aligned against the genome assembly using BLAT software (version 0.36) [27] and GeneWise software (version 2.4.1) [28]. 21,375 protein-coding genes were obtained by combining the different evidences using Glean software (version 1.0) [29]. In the final gene models, the average length was 16,183.81 bp, with an average of 10 exons. The average length of coding sequences, exons and introns were 1,789.82bp, 179.10 bp and 1599.48 bp, respectively, similar to that of the other released fish genomes, such as *Astatotilapia calliptera*, *Maylandia zebra*, *Perca flavescens* and *Pundamilia nyererei* [30-32] (Supplementary Table 4, Supplementary Figure 2). Gene annotation of mitochondria was performed using MitoZ software [14], and 13 protein-coding genes as well as 22 tRNA genes were annotated.

Functions of the annotated protein-coding genes were inferred by searching homologs in the databases (KEGG, COG, NR, Swissprot and Interpro) [33-37]. In this way, 18,005 (84.23%), 7,343 (34.35%), 20,141 (94.23%), 19,114 (89.42%) and 19,313 (90.35%) of protein-coding genes had their homologous alignment in the above databases, respectively. The remaining 1,212 (5.67%) protein-coding genes with unknown function might be the specific feature of the *Chelmon rostratus* genome (Supplementary Table 5).

Gene family and phylogenetic analysis

To identify and analyze the gene families, we selected other eleven species with whole genome sequences available (*Lepisosteus oculatus*, *Hippocampus comes*, *Larimichthys crocea*, *Gasterosteus aculeatus*, *Takifugu rubripes*, *Oreochromis*

niloticus, *Astyanax mexicanus*, *Danio rerio*, *Cynoglossus semilaevis*, *Oryzias latipes* and *Homo sapiens* as outgroup) [24,38-47]. The protein coding genes of the total twelve species were clustered into 18,502 gene families using TreeFam [48-50]. Among these gene families, 2,301 were single-copy gene families (one copy in each of these species) (Figure 4B). The 21,375 protein-coding genes of *Chelmon rostratus* were classified into 13,797 gene families, given an average of 1.55 genes per gene family. Compared to the other species, it was similar to the gene family numbers of *Lepisosteus oculatus* (13,967) and *Gasterosteus aculeatus* (13,331), but was quite different from those of *Larimichthys crocea* (14,724), *Homo sapiens* (14,578) and *Takifugu rubripes* (12,554) (Supplementary Table 6). Among the clustered gene families from *Chelmon rostratus*, 8,190 gene families were common to at least one of the other species, the remaining 52 gene families were unique. Between the four species (*Chelmon rostratus*, *Danio rerio*, *Takifugu rubripes* and *Larimichthys crocea*), the number of common shared and unique gene was shown in Figure 4C. To understand the function of these gene families, we further performed GO enrichment with these gene families from *Chelmon rostratus*, compared with the other 11 species. The result reflected that unique gene families from *Chelmon rostratus* were enriched in muscle contraction functions (Supplementary Table 7). Phylogenetic analysis using the concatenated sequence alignment of the 2,301 single-copy genes shared by the twelve species was performed. The PhyML software (version 3.0) [51], based on the method of maximum likelihood, was used to construct the phylogenetic tree. The split time between *Chelmon rostratus* and *Larimichthys crocea* was estimated to be ~92 million years ago (Figure 4A). Based on the similarity of the protein sequences, 483 syntenic blocks were identified by using McScanX software (version 0.8) [52] (Supplementary Table 8). The time of the duplication and divergence event in these species was calculated based on the distribution of synonymous mutation rate for the gene pairs in the paralogous syntenic blocks, indicating that whole-genome duplication (WGD) event was not detected in *Chelmon rostratus* genome (Figure 3).

The expansion and contraction of the gene family analysis may reveal the evolutionary dynamics of gene families thus provide the clues for understanding the

diversity of different species. It is often inferred from the number of genes in the gene family and the phylogenetic tree. In our study, we used the CAFÉ (version 2.1) [53] software to analyze the expansion and contraction of clustered gene families (Figure 4D). As a result, a total of 18,498 gene families from the most recent common ancestor (MRCA) have been identified. Compared to the recent common ancestor between *Chelmon rostratus* and *Larimichthys crocea*, 793 gene families were expanded and the majority of the expanded gene families were found to be involved in synapse organization. (Supplementary Table 9). On the other hand, there was 2,962 gene family contracted involved in immune system process (Supplementary Table 10).

Data Records

Raw reads from BGISEQ-500 sequencing are deposited in the CNGB Nucleotide Sequence Archive (CNSA) with accession number CNP0000597 (https://db.cngb.org/cnsa). Data Citation 1: CNGB Nucleotide Sequence Archive CNP0000597).

Technical Validation

To evaluate the genome assembly, we aligned sequencing data which we filtered previously using SOAPaligner (version 2.2) [10] and found that 90.76% could be mapped back to the assembled genome. we also calculate its GC depth to rule out possible biases during sequencing or possible contaminations. We identified the average GC contents of this genome to be ~42.52% and we found a continuous GC depth distribution (Supplementary Figure 3), indicating no obvious assembly errors resulted from GC content or contamination. The genome completeness assessment was estimated with Benchmarking Universal Single-copy Orthologs (BUSCO, version 3.0.1) [54]. BUSCO analysis showed that 97.3% (2,518) of core genes were found in our assembly with 2,491 (96.3%) were single copy gene and 27 (1.0%) were duplicated (Supplementary Table 11), indicating a good coverage of the genome.

To validate the quality of predicted gene sets, we also assessed the completeness using

BUSCO (version 3) with the fish core gene database (actinopterygii_odb9) [55]. We found that about 90.20% of core gene were annotated in our gene set with 4,000 were single copy gene and 132 were duplicated (Supplementary Table 12).

Acknowledgements

This work is supported by the special funding of "Blue granary" scientific and technological innovation of China (2018YFD0900301-05). The work also received the technical support from China National Gene Bank.

Author contributions

Y.S., G.F., S.L. and X.L. conceived the work. M.Z. collected sample and X.H. sequenced the libraries. Y.S., X.H., S.Z., Y.A. collected the public data and performed the analyses. Y.C., M.L., C.L. and Y.Z. helped in the analysis. Y.S., X.H., G.F., I.S., S.L. and X.L. wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

References

1 Kuiter, R. H. & Tonozuka, T. *Pictorial guide to Indonesian reef fishes*. (Zoonetics, 2001).

2 Lieske, E. & Myers, R. Collins pocket guide: coral reef fishes. *Collins, London* (2001).

3 EA, L. & CK, T. Fine structure of the gastric epithelium of the coral fish, Chelmon rostratus Cuvier. *Okajimas folia anatomica Japonica* **51**, 285-309 (1975).

4 Ferry-Graham, L. A., Wainwright, P. C. & Bellwood, D. R. Prey capture in long-jawed butterflyfishes (Chaetodontidae): the functional basis of novel feeding habits. *Journal of Experimental Marine Biology and Ecology* **256**, 167-184 (2001).

5 Wang, L.-J., You, F. & Wu, Z.-H. Complete mitochondrial genome of copperband butterflyfish Chelmon rostratus (Teleostei, Perciformes, Chaetodontidae). *Mitochondrial DNA Part A* **27**, 2141-2142 (2016).

6 Panova, M. *et al*. DNA extraction protocols for whole-genome sequencing in marine organisms in *Marine Genomics*. 13-44 (Springer, 2016).

7 Wang, O. *et al.* Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome research* **29**, 798-808 (2019).

8 Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, gix120

(2017).

9 Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint arXiv:1308.2012* (2013).

10 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).

11 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome research* **27**, 757-767 (2017).

12 Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95-98 (2016).

13 Bonev, B. *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**, 557-572. e524 (2017).

14 Meng, G., Li, Y., Yang, C. & Liu, S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic acids research* **47**, e63-e63 (2019).

15 Li, C. *et al.* Draft genome of the Peruvian scallop Argopecten purpuratus. *GigaScience* **7**, giy031 (2018).

16 Liu, H.-P. *et al.* The sequence and de novo assembly of Oxygymnocypris stewartii genome. *Scientific data* **6**, 190009 (2019).

17 Song, L. *et al.* Draft genome of the Chinese mitten crab, Eriocheir sinensis. *GigaScience* **5**, 5 (2016).

18 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580 (1999).

19 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6, 11 (2015).

20 Smit, A., Hubley, R. & Green, P. RepeatMasker, version 4.0.9. *Available fom http://www.repeatmasker.org/* (2015).

21 Smit, A. & Hubley, R. RepeatModeler Open-1.0. *Available fom http://www. repeatmasker. org* (2008).

22 Shao, C. *et al.* Chromosome-level genome assembly of the spotted sea bass, Lateolabrax maculatus. *GigaScience* **7**, doi:10.1093/gigascience/giy114 (2018).

23 Valenzano, D. R. *et al.* The African turquoise killifish genome provides insights into evolution and genetic architecture of lifespan. *Cell* **163**, 1539-1554 (2015).

24 Ao, J. *et al.* Genome sequencing of the perciform fish Larimichthys crocea provides insights into molecular and genetic mechanisms of stress adaptation. *PLoS genetics* **11**, e1005118 (2015).

25 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435-W439 (2006).

26 Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879 (2004).

27 Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome research* **12**, 656-664 (2002).

28 Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome research* **14**, 988-995 (2004).

29 Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome biology* **8**, R13 (2007).

30 Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**, 375 (2014).

31 Ozerov, M. Y. *et al.* Highly Continuous Genome Assembly of Eurasian Perch (Perca fluviatilis) Using Linked-Read Sequencing. *G3: Genes, Genomes, Genetics* **8**, 3737-3743 (2018).

32 Malinsky, M. *et al.* Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature ecology & evolution* **2**, 1940 (2018).

33 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).

34 Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 41 (2003).

35 Kanz, C. *et al.* The EMBL nucleotide sequence database. *Nucleic acids research* **33**, D29-D33 (2005).

36 Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**, 45-48 (2000).

37 Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* (2018).

38 Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* **48**, 427-437 (2016).

39 Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J. & Kocher, T. D. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. *Bmc Genomics* **18**, 341 (2017).

40 Hinaux, H. *et al.* De novo sequencing of Astyanax mexicanus surface fish and Pachón cavefish transcriptomes reveals enrichment of mutations in cavefish putative eye genes. *PloS one* **8**, e53553 (2013).

41 Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498 (2013).

42 Chen, S. *et al.* Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature genetics* **46**, 253 (2014).

43 Kasahara, M. *et al.* Kasahara, M. et al. The medaka draft genome and insights into vertebrate genome evolution. Nature 447, 714-719. *Nature* **447**, 714-719 (2007).

44 Wong, K. H. Y., Michal, L.-S. & Pui-Yan, K. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature Communications* **9**, 3040- (2018).

45 Lin, Q. *et al.* The seahorse genome and the evolution of its specialized morphology. *Nature* **540**, 395-399 (2016).

46 Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61 (2012).

47 Aparicio & S. Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. *Science* **297**, 1301-1310 (2002).

48 Ruan, J. et al. TreeFam: 2008 update. *Nucleic acids research* 36, D735-D740 (2007).

49 Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic acids research* **34**, D572-D580 (2006).

50 Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M. & Bateman, A. TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic acids research* 42, D922-D925 (2013).

51 Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies:

assessing the performance of PhyML 3.0. *Systematic biology* 59, 307-321 (2010)

52 Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research* **40**, e49-e49 (2012).

53 De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271 (2006).

54 Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology & Evolution* **35** (2017).

55 Yang, X. *et al.* Chromosome□level genome assembly of Triplophysa tibetana, a fish adapted to the harsh high□altitude environment of the Tibetan Plateau. *Molecular ecology resources* (2019).

Data Citations

1. Ao, J. *et al.* Genome sequencing of the perciform fish Larimichthys crocea provides insights into molecular and genetic mechanisms of stress adaptation. *PLoS genetics* **11**, e1005118 (2015).
2. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**, 375 (2014).
3. Ozerov, M. Y. *et al.* Highly Continuous Genome Assembly of Eurasian Perch (Perca fluviatilis) Using Linked-Read Sequencing. *G3: Genes, Genomes, Genetics* **8**, 3737-3743 (2018).
4. Malinsky, M. *et al*. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature ecology & evolution* **2**, 1940 (2018).
5. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* **48**, 427-437 (2016).
6. Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J. & Kocher, T. D. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. *Bmc Genomics* **18**, 341 (2017).
7. Hinaux, H. *et al*. De novo sequencing of Astyanax mexicanus surface fish and Pachón cavefish transcriptomes reveals enrichment of mutations in cavefish putative eye genes. *PloS one* **8**, e53553 (2013).
8. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498 (2013).
9. Chen, S. *et al*. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature genetics* **46**, 253 (2014).
10. Kasahara, M. *et al*. Kasahara, M. et al. The medaka draft genome and insights into vertebrate genome evolution. Nature 447, 714-719. *Nature* **447**, 714-719 (2007).
11. Wong, K. H. Y., Michal, L.-S. & Pui-Yan, K. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature Communications* **9**, 3040- (2018).
12. Lin, Q. *et al*. The seahorse genome and the evolution of its specialized morphology. *Nature* **540**, 395-399 (2016).
13. Jones, F. C. *et al*. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61 (2012).
14. Aparicio & S. Whole-Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. *Science* **297**, 1301-1310 (2002).

1

Tables

Table 1 Summary of the scaffold-level assembly. The N50 values of contigs and scaffolds were 294.41 Kb and 2.61 Mb, respectively, indicating good contiguity of the scaffold-level assembly for further Hi-C data assembly.

Species name	***Chelmon rostratus***
Estimated genome size (bp)	711,393,276
Assembly size (bp)	638,700,992
Scaffold N50 (bp)	2,610,810
Longest scaffold (bp)	9,024,273
Contig N50 (bp)	294,414
Longest contig (bp)	1,961,159
Number of genes	21,375
GC content (%)	42.52
BUSCO (%)	98.40
Mapping ratio (%)	90.76
Coverage (%)	99.28

Table 2 Summary of the chromosome-level assembly. About ~26 Gb valid Hi-C sequencing data was used to connect scaffold-level assembly into chromosomes with length of 603Mb (account for 94.40% of scaffold-level assembly size) and constructed 24 chromosomes, with lengths ranging from 14.13 (chr24) Mb to 33.19 Mb (chr1).

Chromosome	Length (bp)	GC (%)
chr1	33,188,845	41.94
chr2	29,501,685	42.26
chr3	29,388,801	42.61
chr4	29,119,845	42.33
chr5	29,062,128	42.38
chr6	28,079,113	42.55
chr7	28,043,518	42.42
chr8	26,564,978	42.56
chr9	26,447,630	42.51
chr10	26,405,724	42.19
chr11	26,164,899	42.37
chr12	25,929,763	42.31
chr13	25,766,544	42.17
chr14	25,529,937	42.41
chr15	24,867,447	42.81
chr16	24,414,281	43.00
chr17	24,306,480	42.65
chr18	23,537,532	42.75
chr19	23,051,343	42.21
chr20	22,222,359	42.56
chr21	21,649,575	42.91
chr22	20,076,678	43.05
chr23	15,560,595	43.28
chr24	14,125,005	43.64
Average	25,125,196	42.58

Figure Legends

Figure 1 Photograph of *Chelmon rostratus*.

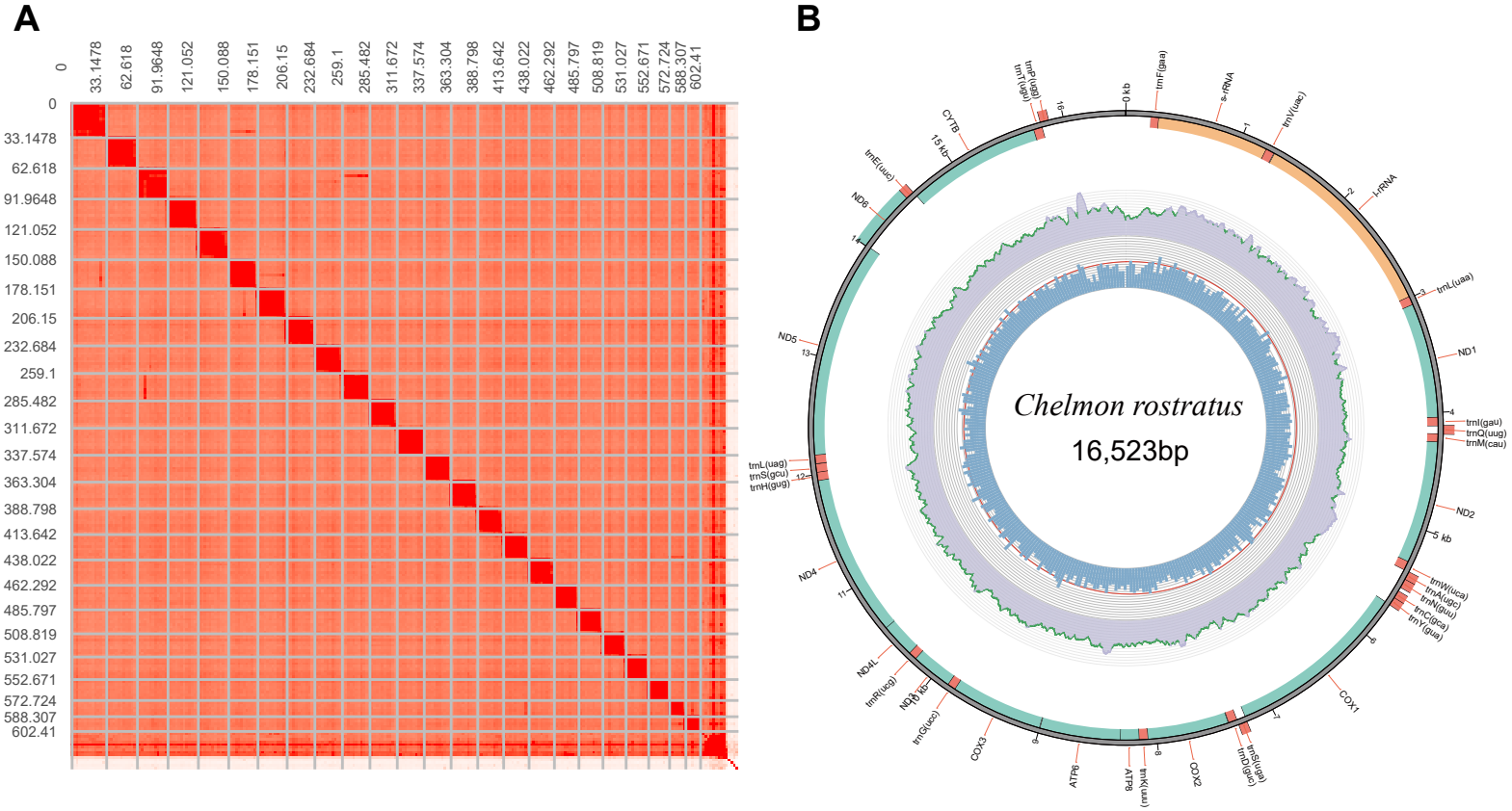
Figure 2 Heat map of interactive intensity between chromosome sequences (A) and Physical map of mitochondrial assembly (B).

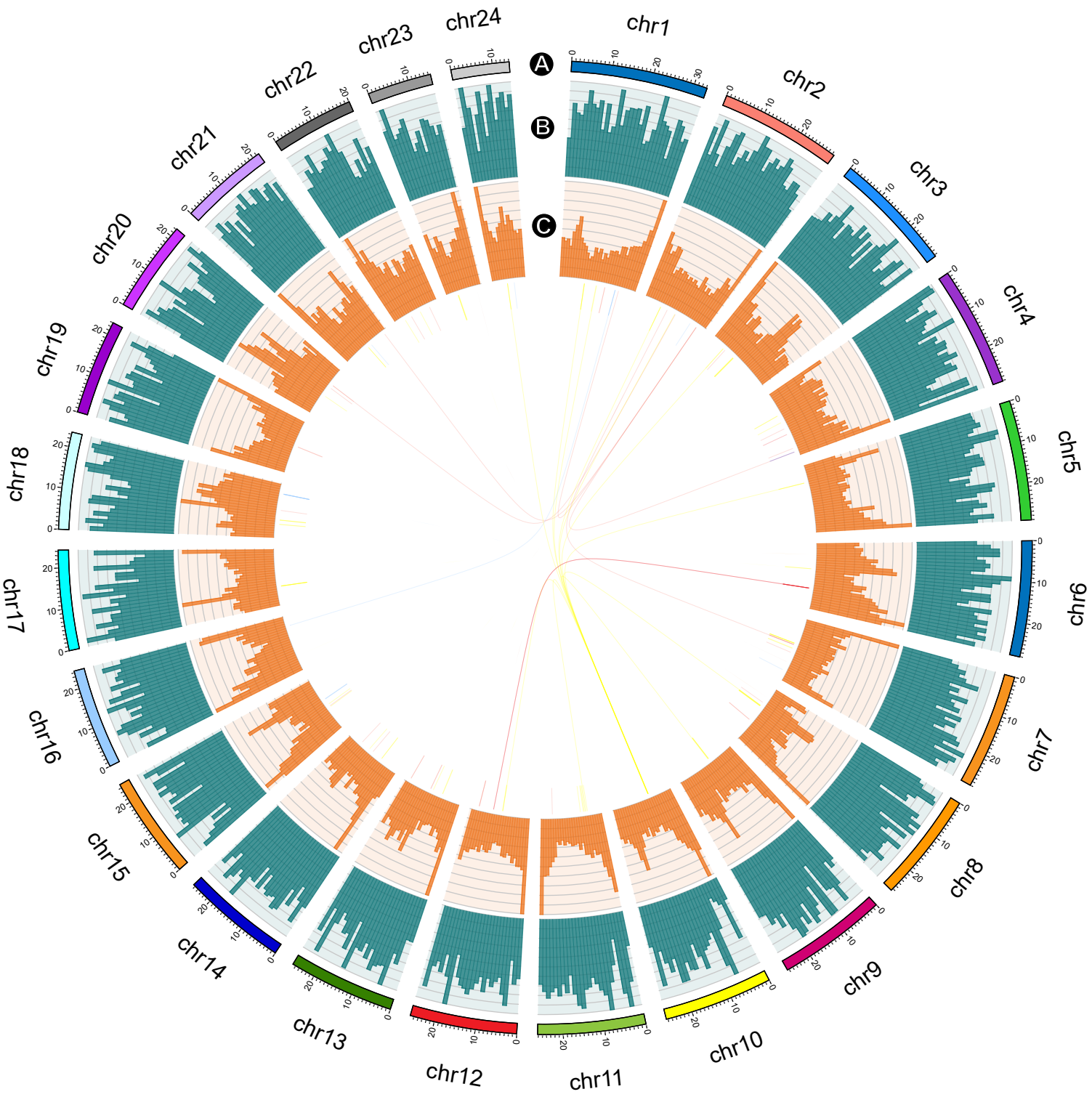
Figure 3 Distribution of basic genomic elements of *Chelmon rostratus* genome. (A) Chromosome karyotype. Different colored represented different chromosome we assembled. (B) Gene density. The histogram indicated number of genes per 1 Mb ranges from a minimum of 0.17 to a maximum of 1, illustrated by blue bar. (C) Repeat sequence density. The histogram indicated average DNA TE ratio per 1 Mb ranges from 0.16 to 1, illustrated by orange bar. Synteny blocks of each chromosome was illustrated by color lines, indicating that whole-genome duplication (WGD) event was not detected in *Chelmon rostratus* genome. Circos (Krzywinskiet al. 2009) (http://circos.ca) was used for constructing this diagram.

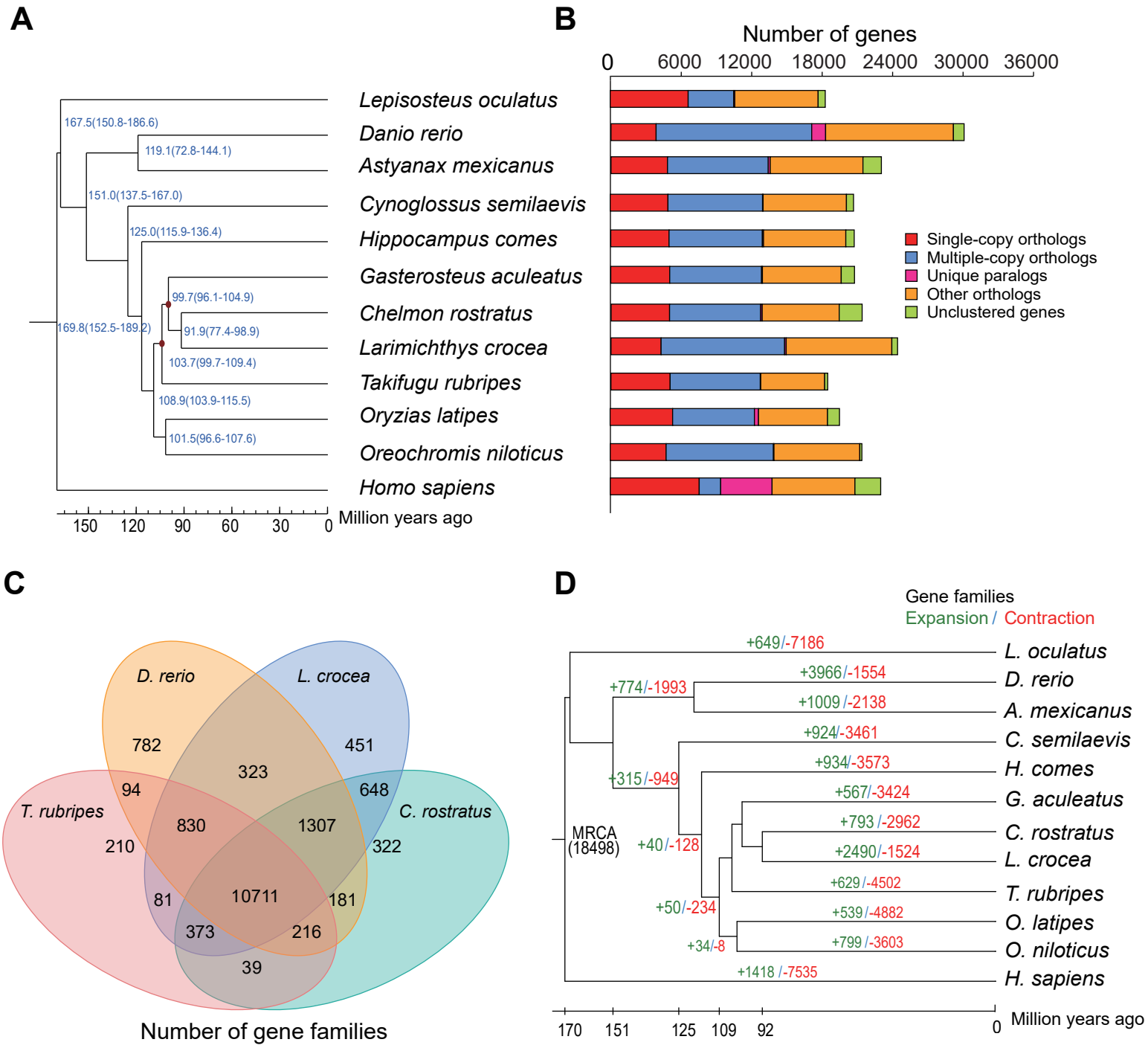
Figure 4 Comparative analysis of the *Chelmon rostratus* genome. (A) Phylogenetic analysis among *Lepisosteus oculatus*, *Hippocampus comes*, *Larimichthys crocea*, *Gasterosteus aculeatus*, *Takifugu rubripes*, *Oreochromis niloticus*, *Astyanax mexicanus*, *Danio rerio*, *Cynoglossus semilaevis*; *Oryzias latipes* and *Homo sapiens* as outgroup, by using the single-copy gene families. The species differentiation time between *Chelmon rostratus* and *Larimichthys crocea* was ~92 million years ago. (B) The protein coding genes of the total twelve species were clustered into 18,502 gene families. Among these gene families, 2,301 were single-copy gene families (one copy in each of these species). (C) Venn diagram showing overlaps of gene families between *Chelmon rostratus*, *Danio rerio*, *Takifugu rubripes* and *Larimichthys crocea*. A total of 322 gene families were unique to *Chelmon rostratus* and 10,711 were commonly shared by the other species genome. (D) Compared to the recent common ancestor between *Chelmon rostratus* and *Larimichthys crocea*, 793 gene families were expanded and 2,962 gene family contracted in *Chelmon rostratus* genome.

1

20180609 XM 006


A
0
33.1478
62.618
91.9648
121.052
150.088
178.151
206.15
232.684
259.1
285.482
311.672
337.574
363.304
388.798
413.642
438.022
462.292
485.797
508.819
531.027
552.671
572.724
588.307
602.41
B
Chelmon rostratus
16,523bp
0 kb
5 kb
10 kb
15 kb
s-rRNA
l-rRNA
trnV(uac)
trnL(uaa)
ND1
trnI(gau)
trnQ(uug)
trnM(cau)
ND2
trnW(uca)
trnA(ugc)
trnN(guu)
trnC(gca)
trnY(gua)
COX1
trnS(uga)
trnD(guc)
COX2
trnK(uuu)
ATP8
ATP6
COX3
trnG(ucc)
ND3
trnR(ucg)
ND4L
ND4
trnH(gug)
trnS(gcu)
trnL(uag)
ND5
ND6
trnE(uuc)
CYTB
trnT(ugu)
trnP(ugg)
trnF(gaa)


A
B
C
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
chr24


A
167.5(150.8-186.6)
119.1(72.8-144.1)
151.0(137.5-167.0)
125.0(115.9-136.4)
99.7(96.1-104.9)
169.8(152.5-189.2)
91.9(77.4-98.9)
103.7(99.7-109.4)
108.9(103.9-115.5)
101.5(96.6-107.6)
Lepisosteus oculatus
Danio rerio
Astyanax mexicanus
Cynoglossus semilaevis
Hippocampus comes
Gasterosteus aculeatus
Chelmon rostratus
Larimichthys crocea
Takifugu rubripes
Oryzias latipes
Oreochromis niloticus
Homo sapiens
150 120 90 60 30 0
Million years ago
B
Number of genes
0 6000 12000 18000 24000 30000 36000
Single-copy orthologs
Multiple-copy orthologs
Unique paralogs
Other orthologs
Unclustered genes
C
D. rerio
L. crocea
T. rubripes
C. rostratus
782
451
323
94
648
830
1307
210
322
81
10711
181
373
216
39
Number of gene families
D
Gene families
Expansion / Contraction
+649/-7186
+774/-1993
+3966/-1554
+1009/-2138
+924/-3461
+315/-949
+934/-3573
+567/-3424
+793/-2962
MRCA
(18498)
+40/-128
+2490/-1524
+629/-4502
+50/-234
+539/-4882
+34/-8
+799/-3603
+1418/-7535
L. oculatus
D. rerio
A. mexicanus
C. semilaevis
H. comes
G. aculeatus
C. rostratus
L. crocea
T. rubripes
O. latipes
O. niloticus
H. sapiens
170 151 125 109 92 0
Million years ago