BMC Genomics

RESEARCH ARTICLE Open Access

Long-read sequencing and de novo genome assembly of marine medaka (*Oryzias melastigma*)

Pingping Liang[1], Hafiz Sohaib Ahmed Saqib[2], Xiaomin Ni[1,3] and Yingjia Shen[1*]

Abstract

Background: Marine medaka (*Oryzias melastigma*) is considered as an important ecotoxicological indicator to study the biochemical, physiological and molecular responses of marine organisms towards increasing amount of pollutants in marine and estuarine waters.

Results: In this study, we reported a high-quality and accurate de novo genome assembly of marine medaka through the integration of single-molecule sequencing, Illumina paired-end sequencing, and 10X Genomics linked-reads. The 844.17 Mb assembly is estimated to cover more than 98% of the genome and is more continuous with fewer gaps and errors than the previous genome assembly. Comparison of *O. melastigma* with closely related species showed significant expansion of gene families associated with DNA repair and ATP-binding cassette (ABC) transporter pathways. We identified 274 genes that appear to be under significant positive selection and are involved in DNA repair, cellular transportation processes, conservation and stability of the genome. The positive selection of genes and the considerable expansion in gene numbers, especially related to stimulus responses provide strong supports for adaptations of *O. melastigma* under varying environmental stresses.

Conclusions: The highly contiguous marine medaka genome and comparative genomic analyses will increase our understanding of the underlying mechanisms related to its extraordinary adaptation capability, leading towards acceleration in the ongoing and future investigations in marine ecotoxicology.

Keywords: de novo genome assembly, Marine ecotoxicology, Pacific biosciences SMRT sequencing, Transposable elements

Background

With the rapid development of global industrialization, pollutants, such as oil contaminations and heavy metals, released into the rivers and coastal waters increase every year [1–3]. Those pollutions have drawn extensive attention because they are toxic, non-biodegradable, easy to accumulate and they have drastic effects on living organisms and the ecosystem. Furthermore, the ecotoxicological impacts of pollutants are different on inhabiting flora and fauna between seawater and freshwater ecosystems [4, 5]. Whereas, many characteristics of seawater are dramatically different from those of freshwater (i.e., ionic strength, buoyancy, salinity, density, dissolved oxygen and pH) [6, 7]. These differences modulate the impact of ecotoxicological features of pollutants, such as the packing fraction and size, the bioaccumulation of the pollutants, the distribution and composition of the pollutants in liquid and solid phases [8]. Thus, the rising level of anthropogenic pollutants in coastal and estuaries waters is attracting researchers to establish an appropriate seawater model organism to precisely examine the

* Correspondence: shenyj@xmu.edu.cn
[1]College of the Environment and Ecology, Xiamen University, Xiamen 361102, China
Full list of author information is available at the end of the article

ecotoxicological effects of contaminants on evolutionary adaptations of marine fauna.

Over the past decades, several fish species such as tilapia (*Oreochromis niloticus*) [9], rainbow trout (*Oncorhynchus mykiss*) [10], Japanese medaka (*Oryzias latipes*) [11] and zebrafish (*Danio rerio*) [12] have been widely used as model organisms to study the ecotoxicological impacts on freshwater ecosystems in laboratory experiments. Researchers have found that some coastal or estuaries candidate species; such as *Enteromorpha linza* [13], *Corophium acherusicum* [14] and *Ctenogobius giurinus* [15], can potentially be used for the ecotoxicological investigations in seawater ecosystems. However, research findings based on these seawater species lag far behind than their freshwater counterparts because of high species specificity to the living environment and the lack of adequate genetic information [16]. Consequently, researchers are in urgent need of marine sentinel model organisms, as many estuaries and coastal waters are highly contaminated.

The marine medaka, *Oryzias melastigma*, also designated as *O. dancena*, distributes broadly in the coastal and fresh waters of Pakistan, India, Myanmar and Thailand [17]. *O. melastigma* is considered as a pragmatic model fish due to its smaller size (4.5 to 23 mm), short life span (2–3 months), high fecundity, distinctive life stages, prominent gender dimorphism in the morphology of anal fin [18] and adaptability to survive in varying aquatic salinity, ranging 0–35 ppt [4, 5]. These physical and morphological characteristics have made the *O. melastigma* a model organism for ecotoxicological investigations [16, 19–24]. In recent years, many ecotoxicological studies have been focused on molecular responses of *O. melastigma* against several environmental stresses [4]. However, previous methodologies or sequencing technology had limitations that need to be amended for correct demonstration of genomes and the better understanding of molecular adaptations.

Fortunately, plummeting cost and numerous advancements in sequencing technologies and bioinformatic algorithms have made assembling of highly sophisticated genomes possible with relatively low cost. Currently, one draft genome of *O. melastigma* has published based on a reference genome assistant assembling approach [25]. The published genome of *O. melastigma* was generated using Illumina reads from several libraries, including three paired-end libraries (PE400, PE500 and PE800) and four MP (mate-pair) libraries (MP2kb, MP5kb, MP10kb and MP20kb). Then scaffolds and pseudochromosomes were assembled based on alignment to the chromosomes of Japanese medaka (*Oryzias latipes*) genome [25]. However, studies have shown that the usage of short Illumina sequencing reads for whole-genome sequencing is a cost-efficient way, but it can also omit the most exciting and perhaps evolutionarily important genome regions [26]. Moreover, duplicated regions of the genome are too tricky to assemble due to their high sequences identity and repetitive nature [27–29]. Therefore, the recently duplicated and high repetitive regions in the previous genome assembly of *O. melastigma* may collapse characteristically. Because using only short Illumina sequencing reads is futile to assemble the duplicated and repetitive “dark-matter” regions of the genome.

Long-read genome sequencing is a more promising approach that provides high consensus accuracy, long reads length, low level of bias and simultaneous epigenetic characterization of complex vertebrate genomes [29–33]. These advantages of long-read based genome sequencing make it a useful tool for the whole genome sequencing, analyses of hard-to-sequence regions in complex genomes, targeted sequencing, evolutionary and phylogenetic relationship analyses of complex populations and epigenetic characterizations [34, 35]. Therefore, this study was carried out; i) to emphasize the use of long-read sequencing in molecular-based ecotoxicological studies by correct assembling of recently duplicated regions, filling gaps and characterize the high repetitive regions in the genome of *O. melastigma*, Furthermore, ii) to apply the comparative genomics and phylogenetic relationship approaches to understand the origin and evolutionary adaptations of *O. melastigma*. This study will provide a high-quality genome assembly and better understandings of the underlying evolutionary mechanisms by which *O. melastigma* has adapted to diverse living environments.

Results

Genome sequencing, assembly and annotation

The genome of adult male marine medaka was sequenced using a combination of several sequencing approaches. The primary genome assembly of *O. melastigma* was generated using single-molecule real-time (SMRT) sequencing (PacBio Sequel), Illumina paired-end sequencing (Illumina HiSeqX ten) and 10X Genomics linked-reads. The whole-genome size of *O. melastigma* was estimated to be ~ 855 Mb by k-mer analysis (Additional file 2: Table S1). We assembled the *O. melastigma* genome using 80.26X long-read coverage of PacBio sequencing data (68.61Gb) (Additional file 2: Table S2). The sequenced reads were self-corrected and the resulting genome assembly consisted of 2610 contigs (with contig N50 of 700 kb), yielding a high-quality consensus sequence with a total length of ~ 835 Mb. Then, the contigs were connected to scaffolds by 10X Genomics linked-read data. Finally, Illumina paired-end sequencing data was used for error correction (Additional file 2: Table S2). The total size of the assembly was 835.41 Mb in the contig level (with contig N50 of 707.80Kb), and the total length of the final assembly was 844.17 Mb with the most extended scaffold reaching up to 8.67 Mb (39.31% GC-contents and obtained 1257 scaffolds with a scaffold N50 of 1.71 Mb) (Table 1).

To evaluate the accuracy of the genome, we mapped paired-end sequence data generated by Illumina HiSeqX ten platform to the *O. melastigma* genome with BWA (Burrows-Wheeler Aligner) [36]. The 96.19% mapping rate and 99.15% coverage rate showed a high consistency between reads and the genome assembly (Additional file 2: Table S3). Furthermore, we performed a variant calling using SAMtools [37] to evaluate the accuracy of the genome at the single-base level. We identified a total of 3,785,501 single-nucleotide polymorphisms (SNPs) (0.47% of the genome). The 28,611 SNPs (0.0036% of the genome) belonged to homozygous single-nucleotide polymorphisms (Additional file 2: Table S4), indicating high accuracy of genome assembly at the single-base level. Half of the total SNPs were located in the genic regions, with about 5% were distributed in the exon regions (Additional file 2: Table S5).

To assess the completeness of the marine medaka assembly, we compared the assembly to the established core vertebrate gene sets by two methods, Benchmarking Universal Single-Copy Orthologs (BUSCO) [38] and Core Eukaryotic Genes Mapping Approach (CEGMA) [39]. BUSCO and CEGMA analysis identified 94.90% of the eukaryotic BUSCO conserved gene set, and 96.37% of CEGMA gene sets entirely assembled in the current version of the genome (Additional file 2: Table S6).

The reference genome of *O. melastigma* was annotated with 25,699 protein-coding genes (avg.exon/coding genes: 8.89) using transcriptome sequencing data from five tissues, combined with ab initio prediction and homology-based approaches. This number is comparable to those found in other vertebrate genomes [40–43] (Table 2). Furthermore, we were able to generate functional assignments for 99.2% of the marine medaka genes from at least one of the public protein databases (Additional file 2: Table S7). The predicted noncoding RNA genes in the *O. melastigma* genome consisted of 926 miRNA, 1916 tRNA, 2474 lncRNA, 825 rRNA and 295 snRNA genes (Additional file 2: Table S8).

Table 1 Assembly statistics of new *O. melastigma* assembly

Sample ID	Length		Number	
	Contig (bp)	Scaffold (bp)	Contig	Scaffold
Total	835,406,597	844,166,318	2588	1257
Max	5,175,882	8,672,543	–	–
Number > =2000	–	–	2572	1241
N50	707,795	1,709,016	314	151
N60	537,651	1,265,745	449	208
N70	399,374	1,001,150	629	283
N80	271,770	678,245	881	385
N90	157,232	413,071	1281	543

Improvements in genome assembly over the previous version

This new genome assembly of *O. melastigma* significantly improved the contiguity in terms of gap-filling and contig sizes. The previously reported marine medaka genome assembly was generated by Illumina reads [25]. Table 3 provides summary statistics for the comparison between previous *O. melastigma* genome assembly and our new assembly. The total length of the new genome assembly is 844 Mb compared to 779 Mb of the previous genome assembly. However, the scaffold N50 (1.71 Mb) of the new genome assembly is shorter than the scaffold N50 (23.73 Mb) of the previous genome assembly. But the new genome assembly contains only 1331 gaps with the length of 8.76 Mb (1.04% of the genome), which is considerably lesser than the previous assembly (51,440 gaps with a total length of 41.24 Mb, 5.29% of the genome). Compared with the previous genome assembly, this assembly represents a considerable decrease in assembly fragmentation (59,791 versus 2588 contigs). We achieved a 25-fold improvement over the previous *O. melastigma* genome assembly (708Kb vs 29Kb) using N50 contig length as a metric. (Additional file 1: Fig. S1). Furthermore, the length distribution of gaps indicated that the previous assembly has many big gaps in addition to thousands of small gaps (Additional file 1: Fig. S2).

Previously published *O. melastigma* genome assembly was generated by reference-assisted chromosome assembly (RACA) [44] which ordered sequences generated by Illumina short reads and assemble into chromosomal fragments based on information from closely related species and out-groups. We compared previously assembled chromosomes to our de novo contigs to detect the differences in two assemblies. Both assemblies showed a high mapping rate (94%), but we identified several misassembled regions in the previous genome of *O. melastigma* (Fig. 1). For example, two regions of different chromosomes (RACA_21 and RACA_24) in the previous assembly mapped to contig439 of the current genome assembly (Fig. 1a, d). To validate the accuracy of our contigs, we mapped the PacBio long reads back to our de novo contigs. The read depth of the region around the breakpoint of contig439 showed at least 39 mapping reads which spanned the breakpoint, suggesting this region of contig439 is continuous (Fig. 1a, d; Additional file 1: Fig. S3). The similar phenomenon happened to contig1840 and contig1980 (Fig. 1b, c, e; Additional file

Table 2 General statistics for the genomes used by the homolog-based method

Species	Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
Dre	25,619	25,207.59	1642.64	9.42	174.39	2798.97
Gac	20,787	8451.06	1548.67	10.40	148.94	734.44
Gmo	20,095	15,245.21	1459.03	12.72	114.67	1175.90
Ola	19,699	12,145.58	1515.82	10.25	147.82	1148.61
Ome	25,699	14,538.33	1484.55	8.89	167.04	1655.11
Oni	21,437	14,903.11	1714.22	10.90	157.25	1332.07
Tni	19,602	6066.17	1516.59	10.52	144.20	478.02
Tru	18,523	7492.75	1693.53	11.10	152.61	574.33
Xma	20,379	13,751.42	1643.24	10.69	153.77	1250.06

Note: *Takifugu rubripes* (Tru), *Ctenopharyngodon idellus* (Cid), *Cyprinus carpio* (Cca), *Danio rerio* (Dre), *Sinocyclocheilus graham* (Sga), *Ictalurus punctatus* (Ipu), *Homo sapiens* (Hom) and *Mus musculus* (Mmu)

1: Fig. S4; Additional file 1: Fig. S5). Interestingly, the last fragment of RACA_8 and the starting fragment of RACA_11 in the previous assembly is mapped to two distinct regions of contig1840 in the new assembly. These two regions are continuous in contig1840 supported by high mapping rate of PacBio long reads (Additional file 1: Fig. S4). These results indicated that the previous genome misassembled the two pseudo-chromosome RACA_8 and RACA_11. They may belong to one chromosome in *O. melastigma*. Additionally, one fragment (with the length of 5,627,693 bp) at RACA_27 from the previous assembly mapped to contig1980 in the new assembly (with a size of 4,835,141 bp), of which only ~ 4.17 Mb were mapped. Moreover, 77 kb unsequenced gaps in RACA_27 were filled in the new assembly, suggesting that long-read sequencing filled the gaps in the previous assembly (Fig. 1c; Additional file 1: Fig. S5).

To further illustrate the genome-wide synteny, the syntenic blocks among Japanese medaka, previous and new assemblies of marine medaka were analyzed. A total of 18,724 syntenic gene pairs found between new and previous assemblies. Comparison with Japanese medaka showed that many conserved syntenic genes lacked in the previous genome assembly but located in the new assembly. For example, the synteny was observed between new assembly and Japanese medaka but not between Japanese medaka and previous assembly (Fig. 1f). Furthermore, there are also some rearrangements between Japanese medaka and marine medaka genome even they are closely related species (Fig. 2). A total of 8317 structural variations (SVs) were identified after mapping our long-reads to Japanese medaka. The 5944 SVs overlapped with genes, and 1564 were located in coding sequences (Fig. 2b, c).

Table 3 Comparison of previous and new *O. melastigma* assemblies on genomic sequences level

Assembly	Previous assembly		New assembly	
	Contig	Scaffold	Contig	Scaffold
Total size (bp)	738,232,102	779,469,774	835,406,597	844,166,318
Number	59,792	8603	2588	1257
Longest (bp)	268,000	37,948,421	5,175,882	8,672,543
Mean (bp)	12,346	90,604	322,800	671,572
N50	28,594	23,737,187	707,795	1,709,016
%N	0	5.29	0	1.04
Number of gaps	0	51,440	0	1331
Total gap size (bp)	0	41,237,672	0	8,759,721

The most striking differences between the two versions of *O. melastigma* genome assembly was present within the highly repeated regions. The total size of predicted repetitive elements in the new assembly of *O. melastigma* was 326.6 Mb, accounting for 38.69% of the total genome size, which was higher than the previous assemblies of *O. melastigma* and *O. latipes* (33.67%, 262.5 Mb and 37.84%, 277.8 Mb, respectively, Additional file 2: Table S9). Similarly, long terminal repeat retrotransposons (LTR-RT) (10.5%, 88.67 Mb) and long interspersed nuclear elements (LINE) (15.71%, 132.64 Mb) were also more abundant in the new assembly of *O. melastigma* (Additional file 2: Table S9) than the previous assemblies. Figure 3 demonstrated the repeat elements of new and previously published assemblies of *O. melastigma* and *O. latipes*. Both versions of *O. melastigma* showed similar trends for each type of TEs, except for one major intermediate burst of transposition mainly involving DNA, LINE and SINE, in addition to one significant ancient expansion of LTR and LINE (Fig. 3a, c). The number of repetitive elements increased at almost all divergence levels (Additional file 2: Table S9), with most at higher divergences, especially for the LTR (long terminal repeat) (Additional file 1: Fig. S6). Examining the length distribution of LTR in these two assemblies (Fig. 3a, b), we found more and longer LTRs in the new

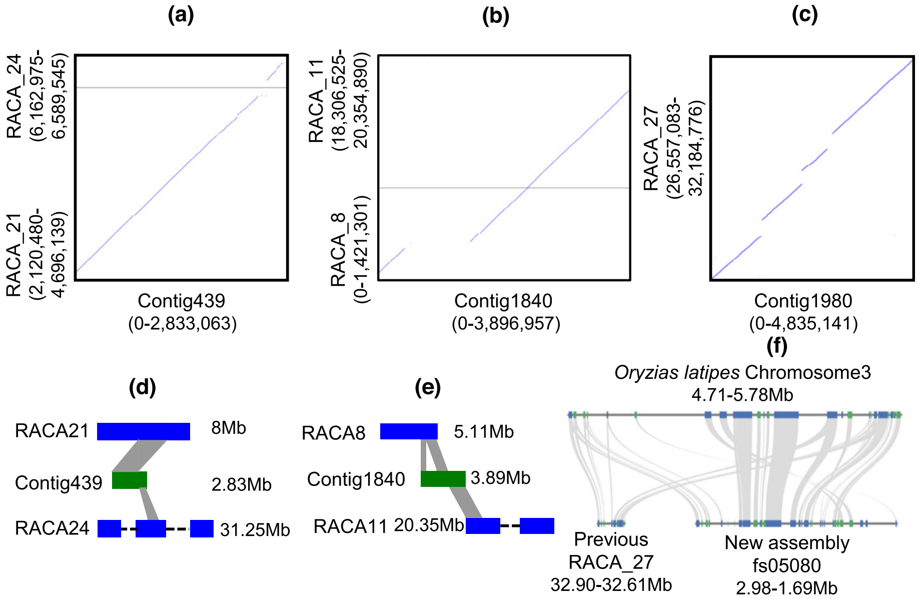

Fig. 1 Examples of comparisons between our de novo contigs and the previous assembly. **a**, **d** A dotplot and sketch map between our de novo contig439 and previous assembly, each dot indicates a region of the close similarity between them. Two chromosomes (RACA_21, RACA_24) of the previous assembly mapped to this contig. **b**, **e** A dotplot and sketch map between our de novo contig1840 and previous assembly. Two chromosomes (RACA81, RACA_11) of the previous assembly mapped to this contig. **c** A dotplot between our de novo contig1980 and previous assembly. **f** One local syntenic block among Japanese medaka, previous assembly and our assembly There are some conserved orthologous genes lacked in the previous assembly (Note: the dotted line in d and e means it is a long chromosome)

assembly, suggesting that using single-molecule sequencing reads can overcome the limitations of short-read sequencing by producing long reads which span the repetitive genomic regions. The previous assembly has almost the same size of repetitive elements with Japanese medaka, but with a lower proportion (Additional file 2: Table S9). Overall, TEs are better assembled in the new assembly (Additional file 2: Table S9). Compared with Japanese medaka, TEs in the *O. melastigma* showed less recent activities (Fig. 3).

Genomic comparison between *Oryzias melastigma* and other vertebrates

Specifying the origin of *O. melastigma* is very important to illustrate the evolution and function of a genome. Mainly, clusters of homologous gene pairs are evidence of candidate homologous regions in distantly related genomes. Demonstrating the statistical significance of such ‘gene clusters’ is an essential component of comparative genomic analyses. A total of 25,595 genes of *O. melastigma* have homologs in other vertebrates and classified into 25,227 orthogroups, 391 of those are single-copy gene families with one-to-one correspondence in different genomes. The distribution of gene family types in each species is shown in Additional file 1: Fig. S7. The trends of gene family types of each species are almost the same except the *Cyprinus carpio* and *Salmo salar* which have more genes. In total, we found 3975 orthogroups shared by all species.

Compared with *Oryzias latipes*, *Nothobranchius furzeri* and *Xiphophorus maculatus*, we found 134 gene families (including 301 genes) unique to *O. melastigma*, which include gene *Hipk1* (Homeodomain-interacting protein kinase 1), apoptosis regulator BAX, *pvrl4*, *Hdlbp*, *CNGB3*, *Fam19a5*, *pycard* and *Clcc1*, etc. Gene Ontology (GO) annotation showed that the unique genes were significantly enriched in functional categories of biological processes, such as apoptotic process (14 genes, Adj. *P*-value = 8.82e-06), programmed cell death (14 genes, Adj. *P*-value = 8.82e-06), cell death (14 genes, Adj. *P*-value = 8.82e-06), death (14 genes, Adj. *P*-value = 8.82e-06), regulation of apoptotic process (12 genes, Adj. *P*-value = 1.36e-05), regulation of cell death (12 genes, Adj. *P*-value = 1.36e-05) and regulation of programmed cell death (12 genes, Adj. *P*-value = 1.36e-05) (Additional file 2: Table S10).

Phylogenetic relationships

The availability of genomic dataset improves the capability to precisely examine the evolutionary history and phylogeny of marine medaka. We clustered the *O. melastigma* gene models with the genes from 17 other vertebrate genomes and used 391 single-copy gene families with one-to-one correspondence in the different genomes to reconstruct a high-confidence phylogenetic

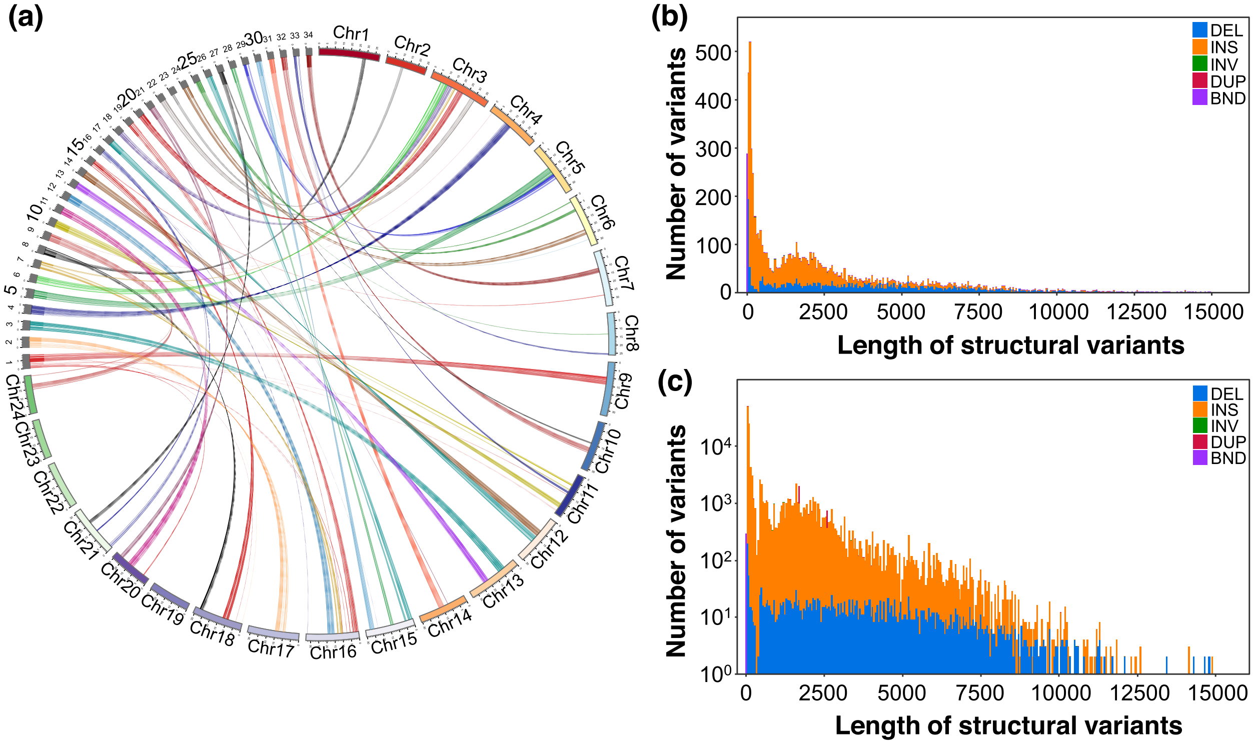

Fig. 2 a The synteny plot is showing the pairwise synteny comparison between the Japanese medaka chromosomes and the 34 scaffolds of the new version of marine medaka. Labels with Chr1 to Chr24 represents the chromosome of Japanese medaka, while labels with 1 to 34 represent the scaffolds of marine medaka. **b**, **c** Length profile of SV calls. The top right panel has SVs up to 15 kb binned per 50 bp; the bottom right panel, up to 15 kb binned per 50 bp with a log-transformed number of variants

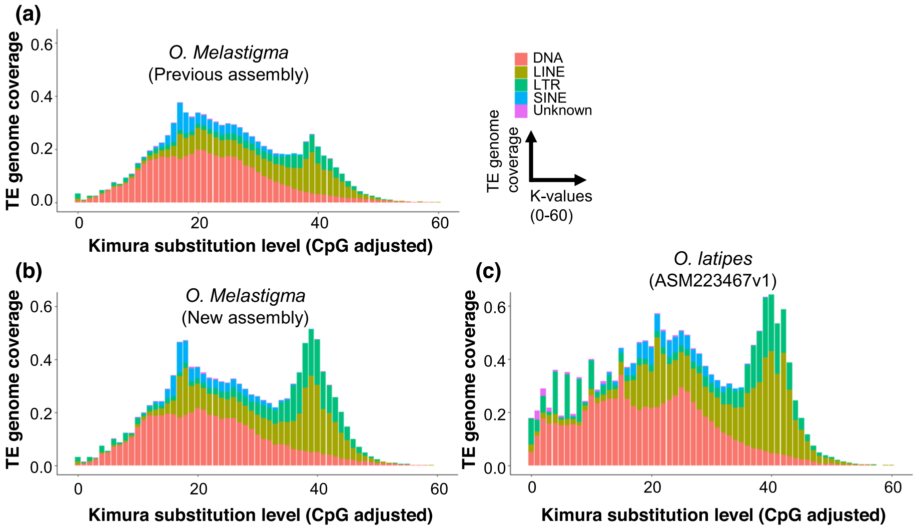

Fig. 3 Landscapes of repeat sequence among three assemblies (**a**) previous *O.melastigma* assembly. **b** new version of *O.melastigma* assembly. **c** *O.latipes*. Kimura distance-based copy divergence analysis of transposable elements in *O. latipes*, previous *O. melastigma* and current assembly (Kimura substitution level – CpG adjusted). The graphs represent genome coverage (y-axis) for each type of TEs (DNA transposons, SINE, LINE, LTR retrotransposons and Unknown) in the different genomes analyzed, clustered according to Kimura distances to their corresponding consensus sequence (x-axis, *K*-value from 0 to 60). Copies clustering on the left of the graph do not diverge very much from the consensus sequence of the element and potentially correspond to recent copies, while sequences on the right might correspond to ancient/degenerated copies

tree and estimate the divergence times with four calibration points (Fig. 4). As a species of the genus *Oryzias*, *O. melastigma* had the closest relationship with *O. latipes*. According to the TimeTree database, the estimated divergence time between *O. latipes* and *O. melastigma* was around 37.3 million years ago. The relationship among other vertebrate genomes is also in agreement with previous estimates [45, 46].

Evolutionary adaptation of marine medaka

Conspicuous expansion or contraction in the size of special gene families is usually connected with the adaptive divergence of closely related species [47, 48]. Based on the result of gene cluster analysis, we undertook a computational analysis of gene family sizes to study gene family expansion and contraction among *O. melastigma* and related species (Additional file 1: Fig. S8). The result showed that there were 25,223 gene families inferred to be present in the most recent common ancestor (MRCA) of mammals. By comparing with the ancestor of *O. melastigma* and *O. latipes*, we found a total of 44 gene families that are significantly (*P*-value< 0.05) expanded in *O. melastigma* and 46 gene families that are significantly contracted (Additional file 2: Table S11). Based on Kyoto Encyclopedia of Genes and Genomes (KEGG) of genes from most of these expanded gene families, we found high enrichment of KEGG pathways includes: calcium signaling pathway (Adj. *P*-value = 3.48e-18), ABC transporters (Adj. *P*-value = 5.78e-16), cell adhesion molecules (CAMs)(Adj. *P*-value = 5.10e-13), circadian entrainment (Adj. *P*-value = 1.18e-12), long-term potentiation (Adj. *P*-value = 2.49e-09), systemic lupus erythematosus (Adj. *P*-value = 1.10e-07) and so on (Table 4; Additional file 2: Table S12). Furthermore, for significantly expanded gene families, we conducted the gene ontology (GO) enrichment analyses and found enrichment for GO terms such as 'ATPase activity' (Adj. *P*-value = 1.15e-31), 'transmembrane transport' (Adj. *P*-value = 2.27e-25), 'ATPase activity, coupled to movement of substances' (Adj. *P*-value = 1.74e-23), 'phospholipid-translocating ATPase activity' (Adj. *P*-value = 3.38e-23), 'phospholipid transport' (Adj. *P*-value = 3.38e-23), 'calcium ion binding' (Adj. *P*-value = 3.99e-13), 'ion channel activity' (Adj. *P*-value =1.33e-08), 'voltage-gated sodium channel activity' (Adj. *P*-value = 6.91e-08), 'ion transport' (Adj. *P*-value =4.37e-06), etc. (Additional file 2: Table S13; Fig. 5).

The evolutionary adaptations of *O. melastigma* populations could have been accompanied by dramatic changes in the environment, such as oil contamination, heavy metals, temperature variation, salinity and pH of seawater. These changes resulted in powerful selective pressures for new genotypes that were better suited in harsh environments. So, signals of very recent positive selection were identified, which provide information about the genetic adaptations of *O. melastigma* to local environmental conditions. We found 274 highly significant (Adj. *P*-value < 0.005) positively selected genes (PSGs) in *O. melastigma* through a likelihood ratio test (Additional file 2: Table S14). KEGG and GO

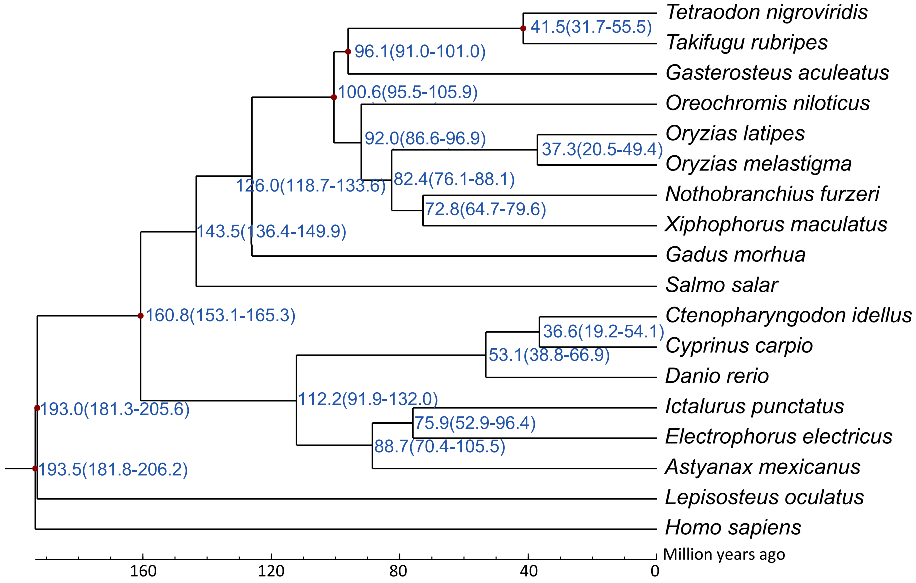

Fig. 4 Phylogenetic tree of *O. melastigma* and related species. Estimates of divergence times (millions of years) calculated from the rate of sequence similarity are indicated at each node. Note: The numbers on the nodes represent the divergence times from present (million years ago, Mya), and the numbers in brackets represent the confidence interval values

Table 4 Functional annotation of the most significantly expanded and contracted gene families in *O. melastigma*

Gene families	KEGG terms	Input no.	Background no.	*P*-value
Expanded gene families	Calcium signaling pathway	33	470	3.48E-18
	ABC transporters	18	66	5.78E-16
	Cell adhesion molecules (CAMs)	24	345	5.10E-13
	Circadian entrainment	26	251	1.18E-12
	Long-term potentiation	18	162	2.49E-09
	Systemic lupus erythematosus	14	118	1.10E-07
	Cocaine addiction	12	100	9.23E-07
	Nicotine addiction	12	100	9.23E-07
	Amyotrophic lateral sclerosis (ALS)	12	106	1.60E-06
	Alzheimer's disease	19	293	3.09E-06
	Dilated cardiomyopathy	16	229	9.50E-06
	Measles	15	216	2.08E-05
	Amphetamine addiction	12	150	4.78E-05
	Asthma	7	45	5.37E-05
	Staphylococcus aureus infection	10	107	6.44E-05
	Axon guidance	18	427	0.000275
	Intestinal immune network for IgA production	7	69	0.000764
	Hypertrophic cardiomyopathy (HCM)	12	205	0.000774
	Viral myocarditis	10	152	0.001042
	Allograft rejection	7	80	0.001658
	Alcoholism	12	227	0.001718
	Autoimmune thyroid disease	7	92	0.003543
	Glutamatergic synapse	12	266	0.006097
Contracted gene families	Tight junction	12	306	4.30E-10
	Systemic lupus erythematosus	5	118	0.000282
	Ascorbate and aldarate metabolism	3	30	0.000866
	Axon guidance	7	427	0.000959
	Drug metabolism - cytochrome P450	3	37	0.000959
	Porphyrin and chlorophyll metabolism	3	39	0.000959
	Drug metabolism - other enzymes	3	42	0.000964
	Metabolism of xenobiotics by cytochrome P450	3	43	0.000964
	Steroid hormone biosynthesis	3	50	0.001068
	Alcoholism	5	227	0.001068
	Pentose and glucuronate interconversions	3	51	0.001068
	Chemical carcinogenesis	3	51	0.001068
	Retinol metabolism	3	68	0.002299
	Starch and sucrose metabolism	3	75	0.002839

annotations results showed the involvement of PSGs in the 'DNA damage repair' functional category (e.g., *DDB2*, *RAD23B*, *CHST11*, *MRE11* and *XRCC3*) (Additional file 2: Table S15; Additional file 2: Table S16). Interestingly, *RAD23B* was located in 'nucleotide excision repair', *DDB2* in 'nucleotide excision repair' and 'p53 signaling pathway', *CHST11* in 'biosynthesis of amino acids', *MRE11* and *XRCC3* were located in 'homologous recombination' functional categories. Furthermore, many of the PSGs (e.g., *PIGB*, *VAMP8*, *RFWD2*, *RPE* and *MYADM*) in the *O. melastigma* genome were enriched in the transporter activities (including 'peptide

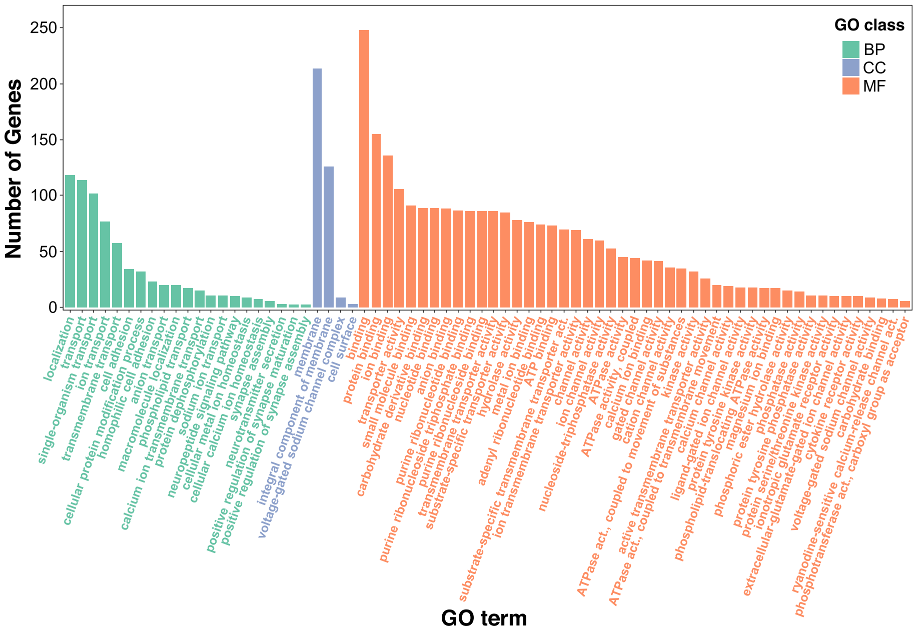

Fig. 5 Gene Ontology (GO) enrichment results of expanded gene families (Adjusted *P-value* < 0.05). The horizontal axis displays the detailed GO annotation corresponding to each functional type (biological processes, cellular components and molecular functions). The vertical axis displays the number of expanded genes

transporter activity', 'peptide transport', 'amide transport', 'nitrogen compound transport', 'ion homeostasis' and 'cation homeostasis').

Discussion

This genome sequencing projects of *O. melastigma* aimed to generate a high-quality reference assembly that can serve as a foundation for various downstream analyses, such as gene finding, variant identification, and comparative and functional assays. Several commonly used second-generation genome sequencing approaches provide gigabases of data [49, 50]. Although these approaches offer higher sequencing depth per sample, the short-read sequencing approach limits the assembly of longer contigs, especially when sequencing the complex heterozygous genomes. Other limitations include that repetitive genomic region in complex genomes is often poorly assembled with short-read sequencing approach [50, 51]. Therefore, in current study, PacBio sequencing provides longer reads which make it easy to sequence through extended repetitive regions and to detect large-scale mutations [52]. The sequence data from both PacBio and 10X Genomics linked-read sequencing can be used to extend contigs and/or fill in the gaps between neighbouring contigs [53]. So, this de novo genome assembly of *O. melastigma* was done by combining the above mentioned methods to cover the regions that would be problematic for short-read sequencing methods. This combined approach has been illustrated to be suitable for highly heterozygous genomes [28, 30, 54, 55].

The advent of PacBio sequencing resulted in the first long-read based genome assembly of *O. melastigma*. We compared the quality of our new assembly with the previous assembly of *O. melastigma* based on the short Illumina sequencing reads [25]. The contig N50 size of our new assembly was substantially higher than the previous genome assembly of *O. melastigma*, highlighting the exclusive benefits of long-read sequencing methods in assembling complex genomes. Similarly, Weisenfeld et al., (2017) have also reported that using long-read genome sequencing enables a genome assembly to achieve both high sequence contiguity as well as high scaffold contiguity. Besides, the new genome assembly of *O. melastigma* revealed numerous errors and filled gaps within and surrounding many genomic regions in the previous assembly. These errors are not limited to intergenic repetitive DNA regions known to be hard to assemble with short reads [56, 57], but also located within functional regions of genes. For example, we compared the syntenic blocks among previous assembly chromosomes with our de novo contigs. The results showed several misassembles of pseudo-chromosomes and several conserved orthologous genes were lacked in the previous assembly. Moreover, we compared the distribution of long terminal repeat retrotransposon (LTR-RT) copies among new assembly and previous assemblies [25, 41].

Our genome assembly can detect more and longer LTR-RT compared with the previous assemblies, demonstrating that single-molecule sequencing of complex genomes can overcome the complications of the short-read sequencing by producing the longer reads spanning repetitive genomic regions.

The extreme conditions in the polluted sea-water may influence the osmoregulation, growth and developmental processes in *O. melastigma*, possibly causing DNA, RNA, and protein damages. Notably, we were able to identify several unique genes, expanded gene families, and genes that underwent positive selection possibly linked to evolutionary adaptations of *O. melastigma*. GO and KEGG functional assays revealed that the molecular functions of those genes are involved in; i) homologous recombination (HR) and nucleotide excision repair pathways, which are essential mechanism to recognize and accurately repair the bulky DNA double-strand breaks (DSBs) [58], ii) an important p53 pathway which is a critical factor that helps to conserve the stability of the genome by preventing mutations caused by cellular stress or DNA damage [59], and iii) the transmembrane transportation and ion transmembrane transportation processes to maintain cell homeostasis for continuous and proper functioning of the cell. For example, in this study we identified *DDB2,* a gene in the nucleotide excision repair and p53 pathways, was positively selected and was known to perform critical functions related to DNA damage repair (such as chromatin remodeling, cell cycle arrest and homologous recombinational) caused by the ionizing radiation or carcinogenic benzo(a) pyrene metabolite [60–63]. Similarly, we found nine *ABCC* expanded genes enriched in ATP-binding cassette (ABC) transporter pathways, where they may regulate the transportation of diverse substances (such as drugs, sterols, ions, sugars, peptides, lipids and proteins) [64, 65]. These results and pieces of evidence from previous studies suggested that expansion/duplication and subsequent positive selection of genes are essential mechanisms for evolutionary adaptation in animals.

Taken together, our improved *de-novo* genome assembly of *O. melastigma* (with more complete and accurately assembled genes of interests) will serve as an ideal reference for future studies based on genome evolution. Moreover, comparative genomic results and functional annotation of expanded and positively selected genes will provide a solid foundation for further investigation of molecular responses of *O. melastigma* to marine environmental stressors.

Conclusions

Marine medaka is considered as a model organism to illustrate the toxicological impacts on the marine ecosystem. In this study, we demonstrated the deployment of long-read sequence technology to generate high-quality, accurate and near to complete draft genome of marine medaka. Comparison with the previous published marine medaka genome based on second-generation sequencing platform and assembled with the assistant of related species indicates that our long-read assembly provides superior performance in terms of contig length, gene contents, gaps filling and repeat sequence detection. Our assembly has a length of 844Mbp, which corresponds to 98.75% of the estimated size of the genome. The results of our study highlighted that the use of single-molecule sequencing reads could overcome the limitations of short-read sequencing. Using this version of the genome, we identified gene families that underwent significant expansion and genes showed the signature of positive selection are enriched in DNA damage repair and cellular transportation of diverse substances pathways, which reflect the evolutionary adaptations of *O. melastigma*. The highly contiguous marine medaka genome and comparative genomic analyses will increase our understanding of mechanisms of its extraordinary adaptation capability and significantly accelerate researches in marine ecotoxicology.

Methods

Sample preparation and sequencing

All animal procedures were carried out in strict compliance with the National Institute of Health Guidelines for the Care and Use of Laboratory Animals and were approved by the animal welfare and ethics committee of Xiamen University. The marine medaka was provided by the State Key Laboratory of Marine Environmental Science, Xiamen University, the State Key Laboratory in Marine Pollution, City University of Hong Kong. Our laboratory established a self-propagating population of marine medaka (bigg-433). The total of 8 mature (five-month-old) male marine medaka were collected, and instantly anaesthetised with dry ice bath for 10s. Muscles of two deeply anaesthetized mature male marine medaka were dissected, and their DNA was extracted for genome sequencing. We used DNA from one mature male of marine medaka for PacBio sequencing, and another mature sibling male for Illumina sequencing. DNA from one fish was insufficient to construct all libraries for sequencing. We also dissected the brain, heart, gill, gonad, muscle from the other six male marine medaka, and extracted the RNA for RNA sequencing.

PacBio library construction and sequencing

Genomic DNA was isolated from the muscles of marine medaka. The qualified genomic DNA was fragmented randomly by ultrasonication (Covaris) and concentrated using the AMPure PB magnetic beads. Then, followed by PacBio SMRTbell 20 kb Library Preparation procedures to construct a 20-kb insert size library. Finally, we sequenced the DNA library on the PacBio Sequel platform, yielding about 68.61 Gb pacbio data (mean read length ≥ 7.9 Kb) (Additional file 2: Table S1). Subreads were filtered with the default parameters. We used falcon [66] to do self-correction and assembly for pacbio data. Then we corrected and polished the assembly to generate high-quality consensus sequences efficiently by Arrow in SmrtLink v5.0.1 [67].

10X genomics library construction, sequencing and extending scaffolds

10X Genomics provides an integrated microfluidics-based platform for generating linked reads and customized software for their analysis [53, 68]. A 10X Genomics library was constructed according to manufacturer's instructions, and a lane of Illumina HiSeqX ten 150 bp paired-end reads was generated with a coverage of about 117.85X. We used BWA mem [36] to align the 10X Genomics linked-reads to consensus sequences gained by PacBio using default settings. Then, we used fragScaff [69] for scaffolding.

A standard protocol to correct PacBio long reads was adopted as a second-generation sequencing platform (like Illumina) to assemble a genome with an error rate of less than 1%. To achieve this goal, one paired-end Illumina sequence library was constructed with an insert size of 350 bp, and sequencing was carried out on the Illumina HiSeqX ten platform according to the manufacturer's instructions; 146.91 Gb (172X coverage) sequencing data were produced. The following criteria filtered raw sequence data generated by the Illumina platform: filtered reads with adapters, filtered reads with N bases more than 10%, and filtered reads with more than 20% of low-quality bases (≤ 5). We used BWA [36] to align all the short clean data to the assembly, then used Pilon [70] with default settings to correct assembled errors.

Repeat prediction

ab initio repeat annotation of marine medaka genome was first carried out by successively using RepeatScout [51], TRF (Tandem Repeats Finder) [71] and LTR_FINDER [72]. The marine medaka repeat library was finally constructed by RepeatMasker [73] through the combined database between ab initio repeat library and the Repbase transposable element library [74]. The identification and classification of genomic repeats were conducted by Piler [75].

Gene and non-coding RNA prediction

EVidenceModeler [76] was used to generate a nonredundant and complete gene set based on ab initio predictions from AUGUSTUS [77], GlimmerHMM [78], SNAP [79], GeneID [80] and Genscan [81], homology annotation with the universal single-copy genes from related species (*Danio rerio*, *Oryzias latipes*, *Xiphophorus maculatus*, *Tetraodon nigroviridis*, *Takifugu rubripes*, *Gasterosteus aculeatus*, *Gadus morhua*, and *Oreochromis niloticus*) (Additional file 2: Table S17) and RNA-seq alignment data. For RNA-seq data derived from the brain, heart, gill, gonad, muscle, we removed adaptor sequences and filtered low-quality reads by using Trimmomatic [82]. The clean reads were de novo assembled and annotated with the Trinity [83], PASA program [84] and Cufflinks [85] after mapping to the new assembly by tophat [86]. Then combined RNA-seq prediction to correct the EVidenceModeler result by PASA and add UTR and alternative splicing information. These results were integrated into a final set of protein-coding genes for annotation.

We then generated functional assignments of the marine medaka genes with BLAST [87] and GeneWise [88] by aligning their protein-coding regions to sequences in public protein databases, including SwissProt [89], NCBI non-redundant protein database, Pfam [90], Gene Ontology [91], KEGG [92] and InterPro [93].

The rRNA fragments were predicted by aligning the rRNA sequences of related species because of high conservation. The tRNA genes were searched by tRNAscan-SE [94]. Additionally, miRNA and snRNA were identified by using INFERNAL [95] to search from the Rfam database [96]. CPC2 [97] and CPAT [98] identified the lncRNAs. Transcripts encoding ORFs longer than 100 amino acids were filtered, and the remaining transcripts were further screened by BLASTX (e-value <1e-10) against the SwissProt and Nr database.

Gene family identification

We downloaded genome and annotation data of *Nothobranchius furzeri*, *Salmo salar*, *Cyprinus carpio*, *Ictalurus punctatus*, *Ctenopharyngodon idellus*, *Oryzias latipes*, *Xiphophorus maculatus*, *Oreochromis niloticus*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Gasterosteus aculeatus*, *Gadus morhua*, *Danio rerio*, *Astyanax mexicanus*, *Lepisosteus oculatus*, *Homo sapiens* and *Electrophorus electricus* (see Additional file 2: Table S17). We chose the longest transcript to represent each gene and removed gene models encode less than 30 amino acids. The similarities among proteins were obtained by blastp [87] with an *E*-value cutoff of 1e-5. Gene family clustering was conducted using OrthoMCL [99] based on the set of predicted genes of *O. melastigma* and the protein sets of the above 17 species. This analysis yielded 25,227 gene families.

Phylogenetic tree construction and phylogenomic dating

A phylogenetic tree was constructed based on a concatenated sequence alignment of 391 single-copy gene families from marine medaka and the 17 other related species. These single-copy gene families were firstly aligned by MUSCLE [100], then concatenated to a super alignment matrix. In the end, ML phylogenic tree was constructed using RaxML [101]. PAML MCMCTree [102] estimated divergence times. The Markov chain Monte Carlo (MCMC) process was run with a sample number of 1,000,000, a sampling frequency of two after a burn-in of 1000 iterations. Other parameters used the default settings of MCMCTree. The following constraints were used for time calibrations: (i) the *Tetraodon nigroviridis* and *Gasterosteus aculeatus* divergence time [149–166 million years ago (Mya)]; (ii) the *Oryzias latipes* and *Gasterosteus aculeatus* divergence time (97–151 Mya); (iii) the *Lepisosteus oculatus* and other 16 fish species divergence time (291–338 Mya); and (iv) the *Homo sapiens* and other 17 species divergence time (416-422Mya). Estimation of gene family expansion and contraction were done using CAFÉ [103].

Detecting positive selection in the genome

Sequence alignments were conducted using the MUSCLE [100] tool for single-copy gene families. Both nonsynonymous (dN) and synonymous substitution rates (dS) and dN/dS ratio (ω) of every lineage were estimated using the branch-site model analysis with codeml program in PAML [104–106]. Based on a maximum likelihood ratio test (LRT), we identified genes under positive selection in marine medaka. These genes were identified as positively selected according to the chi-squared test (P-value < 0.01, FDR < 0.05, df = 1), and containing amino acid sites that were selected with a Bayes probability higher than 95%.

Calling of variants

PacBio subreads were aligned to new assembly (contig level) using NGMLR (v0.2.7) [107] to generate the BAM file. The BAM file was sorted by SAMtools [37], then used as the input of Sniffles (version 1.0.11) [107] to identify structural variant. Jcvi [108] was used to detect the syntenic blocks.

SAMtools package [37] was used to perform SNP calling based on bam file (generated earlier, the same Illumina data used to correct assembly errors). Raw vcf files were filtered, and SnpEff [109] software was used to annotate the variable sites.

Go annotation

Significantly overrepresented GO terms in this study were identified using the topGO [110] package in R programming language, and the FDR correction was applied. Significantly overrepresented GO terms were identified with corrected P-values of ≤0.05.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-07042-7.

Additional file 1: Figure S1. Length distribution of gaps in the previous version (left) and new assembly (right). There are 51,440 and 1,331 gaps in the previous version and new assembly. Moreover, the maximum gap length of them was 892,371 bp and 8,013 bp separately. **Figure S2**. Length distribution of contigs in the previous version (A) and new assembly (B). There are 59,791 and 2,589 contigs in previous version (contig N50 28,594 bp) and new assembly (contig N50 707,795 bp. Furthermore, the maximum contig length of them were 268,000 bp and 5,175,882 bp separately. **Figure S3**. The read depth of the region around breakpoint of new *de novo* contig439. Mapping of PacBio long reads to *de novo* contig439 to showed if it is continuous near 2.57Mb of the contig. **Figure S4**. The read depth of the region around breakpoint of new *de novo* contig1840. Mapping of PacBio long reads to *de novo* contig1840 to showed if it is continuous near 2.31Mb of the contig. **Figure S5**. The read depth of the region around breakpoint of new *de novo* contig1980. Mapping of PacBio long reads to *de novo* contig1980 to showed if it is continuous. **Figure S6**. The length distribution of long terminal repeats (LTR) families for new assembly and previous assembly. **Figure S7**. The distribution of gene family types which include single-copy orthologs, multiple-copy orthologs, unique and other orthologs in each species. **Figure S8**. Estimation of gene family expansion and contraction using CAFÉ. Clock calibrated phylogenetic tree showing the number of gene families significantly (P-value ≤ 0.01) expanded (green), contracted (red). MRCA: most recent common ancestor.

Additional file 2: Table S1. Genomic characteristics statistics of *Oryzias melastigma* (Kmer=17). **Table S2**. Sequencing data used for the *Oryzias melastigma* genome assembly. **Table S3**. The mapping rate and coverage rate of short read sequences. **Table S4**. Statistics of variants calling. **Table S5**. Number of SNP effects by region in the marine medaka genome. **Table S6**. Genome completeness as measured by CEGMA and BUSCO. **Table S7**. Statistical of predicted functional genes in public protein databases. **Table S8**. The number of all kinds of non-coding RNA. **Table S9**. Summary statistics of repeat elements. **Table S10**. Significantly over-represented Gene Ontology (GO) terms among *O. melastigma*-specific genes compared with *Oryzias latipes, Nothobranchius furzeri* and *Xiphophorus maculatus*. "X" is the number of *O. melastigma*-specific genes assigned to that GO term. The GO terms with corrected P-value bellow 0.05 are selected as significantly enriched groups. **Table S11**. The list of 44 expanded gene families and 46 contracted gene families that appeared unique to *Oryzias melastigma*. **Table S12**. KEGG pathway results of expanded gene families. **Table S13**. GO functional enrichment results for expanded gene families. **Table S14**. Positively selected genes in the *O. melastigma*. **Table S15**. Gene Ontology (GO) enrichment of positively selected genes (PSGs) in the *O. melastigma*. **Table S16**. KEGG pathway descriptions of those positively selected genes in *O. melastigma*, which showed significant P-value (0.05). **Table S17**. Species included in the comparative genomics in this study.

Abbreviations

ABC: ATP-binding cassette; SMRT sequencing: Single Molecule, Real-Time sequencing; ppt: Parts per thousand; PE: Paired-end; MP: Mate-pair; BWA: Burrows-Wheeler Aligner; SNP: Single-nucleotide polymorphisms; BUSCO: Benchmarking Universal Single-Copy Orthologs; CEGMA: Core Eukaryotic Genes Mapping Approach; RACA: Reference-assisted chromosome assembly; SV: Structural variation; LTR-RT: Long terminal repeat retrotransposon; LINE: Long interspersed nuclear element; SINE: Short interspersed nuclear element; LTR: Long terminal repeat; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; MRCA: most recent common ancestor; CAMs: Cell adhesion molecules; PSG: Positively selected gene; HR: Homologous recombination; DBS: Double-strand break; PASA: Program to Assemble Spliced Alignments; UTR: Untranslated region; PAML: Phylogenetic Analysis by Maximum Likelihood; MCMC: Markov chain Monte Carlo; Mya: Million years ago; FDR: False discovery rate

Acknowledgments

We thank the Coastal State Key Laboratory of Marine Environmental Science at Xiamen University for providing fish lines used in this study.

Authors' contributions

YS and PL conceived and designed the study. PL and XN conducted lab experiments and data acquisition. PL and YS conducted the analyses, prepared figures and tables, and wrote the draft of the manuscript. HS gave statistical inputs and critically revised the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No.31671318), National Key R&D Program of China (2016YFC0502901), Natural Science Foundation of Fujian Province of China (No.2015 J05074) and the Fundamental Research Funds for the Central Universities (No. 20720150068, No. 20720190106). The funding bodies did not participate in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All sequence data that support the findings of this study have been deposited in GenBank with the following accession numbers: WKFB00000000 for whole-genome sequence assembly under BioProject accession PRJNA556761; SRX8937101 for sequences of the 350 bp library, SRX8937099 for those from the 500-700 bp library, SRX8937100 for those from the 20 kb long-read PacBio library; SRX8911616 -to- SRX89116120 for RNA-Seq data set for heart, muscle, gonad, brain and gill transcriptome. The web links corresponding to the genome and annotation datasets for *Nothobranchius furzeri*, *Salmo salar*, *Cyprinus carpio*, *Ictalurus punctatus*, *Ctenopharyngodon idellus*, *Oryzias latipes*, *Xiphophorus maculatus*, *Oreochromis niloticus*, *Takifugu rubripes*, *Tetraodon nigroviridis*, *Gasterosteus aculeatus*, *Gadus morhua*, *Danio rerio*, *Astyanax mexicanus*, *Lepisosteus oculatus*, *Homo sapiens* and *Electrophorus electricus* are listed in Additional file 2: Table S17.

Ethics approval and consent to participate

All animal procedures were carried out in strict compliance with the National Institute of Health Guidelines for the Care and Use of Laboratory Animals and were approved by the animal welfare and ethics committee of Xiamen University.

Consent for publication

Not applicable.

Competing interests

All authors declared that they have no competing interest.

Author details

[1]College of the Environment and Ecology, Xiamen University, Xiamen 361102, China. [2]State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Fujian Agriculture and Forestry University, Fuzhou 350002, China. [3]Fudan University, Shanghai 200240, China.

Received: 3 February 2020 Accepted: 31 August 2020
Published online: 16 September 2020

References

1. Boudjellaba D, Dron J, Revenko G, Démelas C, Boudenne J-L. Chlorination by-product concentration levels in seawater and fish of an industrialised bay (gulf of Fos, France) exposed to multiple chlorinated effluents. Sci Total Environ. 2016;541:391–9. https://doi.org/10.1016/j.scitotenv.2015.09.046.
2. Li H, Lin L, Ye S, Li H, Fan J. Assessment of nutrient and heavy metal contamination in the seawater and sediment of Yalujiang estuary. Mar Pollut Bull. 2017;117:499–506. https://doi.org/10.1016/j.marpolbul.2017.01.069.
3. de Groot AJ. Metals and sediments: a global perspective met contam aquat sediments. In: metal contaminated aquatic sediments. New York: Routledge; 2018. p. 1–20. https://doi.org/10.1201/9780203747643.
4. Chen X, Li L, Wong CKC, Cheng SH. Rapid adaptation of molecular resources from zebrafish and medaka to develop an estuarine/marine model. Comp Biochem Physiol Part C Toxicol Pharmacol. 2009;149:647–55. https://doi.org/10.1016/j.cbpc.2009.01.009.
5. Inoue K, Takei Y. Diverse adaptability in *Oryzias* species to high environmental salinity. Zool Sci. 2002;19:727–34. https://doi.org/10.2108/zsj.19.727.
6. French RA, Jacobson AR, Kim B, Isley SL, Penn L, Baveye PC. Influence of ionic strength, pH, and cation valence on aggregation kinetics of titanium dioxide nanoparticles. Environ Sci Technol. 2009;43:1354–9. https://doi.org/10.1021/es802628n.
7. Jeon J, Kannan K, Lim HK, Moon HB, Ra JS, Kim SD. Bioaccumulation of perfluorochemicals in pacific oyster under different salinity gradients. Environ Sci Technol. 2010;44:2695–701. https://doi.org/10.1021/es100151r.
8. You C, Jia C, Pan G. Effect of salinity and sediment characteristics on the sorption and desorption of perfluorooctane sulfonate at sediment-water interface. Environ Pollut. 2010;158:1343–7. https://doi.org/10.1016/j.envpol.2010.01.009.
9. Campos-Garcia J, Martinez DST, Alves OL, Leonardo AFG, Barbieri E. Ecotoxicological effects of carbofuran and oxidised multiwalled carbon nanotubes on the freshwater fish Nile tilapia: nanotubes enhance pesticide ecotoxicity. Ecotoxicol Environ Saf. 2015;111:131–7. https://doi.org/10.1016/j.ecoenv.2014.10.005.
10. Correia AT, Rebelo D, Marques J, Nunes B. Effects of the chronic exposure to cerium dioxide nanoparticles in *Oncorhynchus mykiss*: assessment of oxidative stress, neurotoxicity and histological alterations. Environ Toxicol Pharmacol. 2019;68:27–36. https://doi.org/10.1016/j.etap.2019.02.012.
11. Horie Y, Kanazawa N, Yamagishi T, Yonekura K, Tatarazako N. Ecotoxicological test assay using OECD TG 212 in marine Java Medaka (*Oryzias javanicus*) and freshwater Japanese Medaka (*Oryzias latipes*). Bull Environ Contam Toxicol. 2018;101:344–8. https://doi.org/10.1007/s00128-018-2398-1.
12. Zhang Y, Feng J, Gao Y, Liu X, Qu L, Zhu L. Physiologically based toxicokinetic and toxicodynamic (PBTK-TD) modelling of Cd and Pb exposure in adult zebrafish *Danio rerio*: Accumulation and toxicity. Environ Pollut. 2019;249:959–68. https://doi.org/10.1016/j.envpol.2019.03.115.
13. Villares R, Puente X, Carballeira A. Ulva and Enteromorpha as indicators of heavy metal pollution. Hydrobiologia. 2001;462:221–32. https://doi.org/10.1023/A:1013154821531.
14. Reish DJ. Effects of metals and organic compounds on survival and bioaccumulation in two species of marine gammaridean amphipod, together with a summary of toxicological research on this group. J Nat Hist. 1993;27:781–94. https://doi.org/10.1080/00222939300770471.
15. Liu Z, Li X, Tai P, Sun L, Yuan H, Yang X. Toxicity of ammonia, cadmium, and nitrobenzene to four local fishes in the Liao River, China and the derivation of site-specific water quality criteria. Ecotoxicol Environ Saf. 2018;147:656–63. https://doi.org/10.1016/j.ecoenv.2017.09.008.
16. Dong S, Kang M, Wu X, Ye T. Development of a promising fish model (*Oryzias melastigma*) for assessing multiple responses to stresses in the marine environment. Biomed Res Int. 2014;2014:1–17. https://doi.org/10.1155/2014/563131.
17. Naruse K. Classification and phylogeny of fishes of the genus *Oryzias* and its relatives. Fish Biol J Medaka. 1996;8:1–9. https://doi.org/10.18999/fisbjm.8.1.
18. Yip WP. Relating Estradiol and telomeres to longevity in marine medaka *Oryzias melastigma*. 2011. p. 190. http://lbms03.cityu.edu.hk/theses/c_ftt/mphil-bch-b40865356f.pdf. Accessed 20 Dec 2019.
19. Lee C, Kwon BO, Hong S, Noh J, Lee J, Ryu J, et al. Sub-lethal and lethal toxicities of elevated CO_2 on embryonic, juvenile, and adult stages of marine medaka *Oryzias melastigma*. Environ Pollut. 2018;241:586–95. https://doi.org/10.1016/j.envpol.2018.05.091.
20. Huang Q, Fang C, Wu X, Fan J, Dong S. Perfluorooctane sulfonate impairs the cardiac development of a marine medaka (*Oryzias melastigma*). Aquat Toxicol. 2011;105:71–7. https://doi.org/10.1016/j.aquatox.2011.05.012.
21. Chen X, Li L, Cheng J, Chan LL, Wang DZ, Wang KJ, et al. Molecular staging of marine medaka: a model organism for marine ecotoxicity study. Mar Pollut Bull. 2011;63:309–17. https://doi.org/10.1016/j.marpolbul.2011.03.042.
22. Sun L, Zuo Z, Chen M, Chen Y, Wang C. Reproductive and transgenerational toxicities of phenanthrene on female marine medaka (*Oryzias melastigma*). Aquat Toxicol. 2015;162:109–16. https://doi.org/10.1016/j.aquatox.2015.03.013.
23. Hong H, Shen R, Liu W, Li D, Huang L, Shi D. Developmental toxicity of three hexabromocyclododecane diastereoisomers in embryos of the marine medaka *Oryzias melastigma*. Mar Pollut Bull. 2015;101:110–8. https://doi.org/10.1016/j.marpolbul.2015.11.009.

24. Wang J, Wang W. Salinity influences on the uptake of silver nanoparticles and silver nitrate by marine medaka (*Oryzias melastigma*). Environ Toxicol Chem. 2014;33:632–40. https://doi.org/10.1002/etc.2471.
25. Kim HS, Lee BY, Han J, Jeong CB, Hwang DS, Lee MC, et al. The genome of the marine medaka *Oryzias melastigma*. Mol Ecol Resour. 2018;18:656–65. https://doi.org/10.1111/1755-0998.12769.
26. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell. 2017;30:149–61. https://doi.org/10.1007/s13577-017-0168-8.
27. Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, et al. Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res. 2018;5:50. https://doi.org/10.1038/s41438-018-0071-9.
28. Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, et al. Long-read sequence assembly of the gorilla genome. Science. 2016;352:aae0344. https://doi.org/10.1126/science.aae0344.
29. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50. https://doi.org/10.1038/ng.3802.
30. Das A, Ianakiev P, Baten A, Nehleen R, Ehsan T, Ahmed O, et al. Genome of *Tenualosa ilisha* from the river Padma, Bangladesh. BMC Res Notes. 2018;11:921. https://doi.org/10.1186/s13104-018-4028-8.
31. Conte MA, Kocher TD. An improved genome reference for the African cichlid, *Metriaclima zebra*. BMC Genomics. 2015;16:724. https://doi.org/10.1186/s12864-015-1930-5.
32. Xu S, Xiao S, Zhu S, Zeng X, Luo J, Liu J, et al. A draft genome assembly of the Chinese sillago (*Sillago sinica*), the first reference genome for Sillaginidae fishes. Gigascience. 2018;7:giy108. https://doi.org/10.1093/gigascience/giy108.
33. Marcionetti A, Rossier V, Bertrand JAM, Litsios G, Salamin N. First draft genome of an iconic clownfish species (*Amphiprion frenatus*). Mol Ecol Resour. 2018;18:1092–101. https://doi.org/10.1111/1755-0998.12772.
34. Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AWC, Pippel M, et al. The axolotl genome and the evolution of key tissue formation regulators. Nature. 2018;554:50–5. https://doi.org/10.1038/nature25458.
35. Smith JJ, Timoshevskaya N, Ye C, Holt C, Keinath MC, Parker HJ, et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. Nat Genet. 2018;50:270–7. https://doi.org/10.1038/s41588-017-0036-1.
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.
37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.
38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351.
39. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7. https://doi.org/10.1093/bioinformatics/btm071.
40. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496:498–503. https://doi.org/10.1038/nature12111.
41. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007;447:714–9. https://doi.org/10.1038/nature05846.
42. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484:55–61. https://doi.org/10.1038/nature10944.
43. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, et al. The genome sequence of Atlantic cod reveals a unique immune system. Nature. 2011;477:207–10. https://doi.org/10.1038/nature10342.
44. Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, et al. Reference-assisted chromosome assembly. Proc Natl Acad Sci. 2013;110:1785–90. https://doi.org/10.1073/pnas.1220349110.
45. Valenzano DR, Benayoun BA, Singh PP, Zhang E, Etter PD, Hu CK, et al. The African turquoise killifish genome provides insights into evolution and genetic architecture of lifespan. Cell. 2015;163:1539–54. https://doi.org/10.1016/j.cell.2015.11.008.
46. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200–5. https://doi.org/10.1038/nature17164.
47. Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, et al. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet. 2011;43:913–8. https://doi.org/10.1038/ng.889.
48. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. Science. 2010;330:641–6. https://doi.org/10.1126/science.1197005.
49. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. Biomed Res Int. 2012;2012:1–11. https://doi.org/10.1155/2012/251364.
50. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics. 2012;13:341. https://doi.org/10.1186/1471-2164-13-341.
51. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005;21(supll-1):i351–8. https://doi.org/10.1093/bioinformatics/bti1018.
52. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biol. 2013;14:405. https://doi.org/10.1186/gb-2013-14-6-405.
53. Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, et al. A hybrid approach for *de novo* human genome sequence assembly and phasing. Nat Methods. 2016;13:587–90. https://doi.org/10.1038/nmeth.3865.
54. Cai H, Li Q, Fang X, Li J, Curtis NE, Altenburger A, et al. A draft genome assembly of the solar-powered sea slug *Elysia chlorotica*. Sci Data. 2019;6:190022. https://doi.org/10.1038/sdata.2019.22.
55. Kawamoto M, Jouraku A, Toyoda A, Yokoi K, Minakuchi Y, Katsuma S, et al. High-quality genome assembly of the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 2019;107:53–62. https://doi.org/10.1016/J.IBMB.2019.02.002.
56. Palazzo AF, Gregory TR. The case for junk DNA. PLoS Genet. 2014;10:e1004351. https://doi.org/10.1371/journal.pgen.1004351.
57. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13:36–46. https://doi.org/10.1038/nrg3117.
58. Takata M, Sasaki MS, Sonoda E, Morrison C, Hashimoto M, Utsumi H, et al. Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. EMBO J. 1998;17:5497–508. https://doi.org/10.1093/emboj/17.18.5497.
59. Prives C, Hall PA. The p53 pathway. J Pathol. 1999;187:112–26. https://doi.org/10.1002/(SICI)1096-9896(199901)187:1<112::AID-PATH250>3.0.CO;2-3.
60. Zhu Q, Battu A, Ray A, Wani G, Qian J, He J, et al. Damaged DNA-binding protein down-regulates epigenetic mark H3K56Ac through histone deacetylase 1 and 2. Mutat Res. 2015;776:16–23. https://doi.org/10.1016/j.mrfmmm.2015.01.005.
61. Zou N, Xie G, Cui T, Srivastava AK, Qu M, Yang L, et al. DDB2 increases radioresistance of NSCLC cells by enhancing DNA damage responses. Tumor Biol. 2016;37:14183–91. https://doi.org/10.1007/s13277-016-5203-y.
62. Christmann M, Boisseau C, Kitzinger R, Berac C, Allmann S, Sommer T, et al. Adaptive upregulation of DNA repair genes following benzo(a)pyrene diol epoxide protects against cell death at the expense of mutations. Nucleic Acids Res. 2016;44:10727–43. https://doi.org/10.1093/nar/gkw873.
63. Wittschieben BØ, Iwai S, Wood RD. DDB1-DDB2 (xeroderma pigmentosum group E) protein complex recognizes a cyclobutane pyrimidine dimer, mismatches, apurinic/apyrimidinic sites, and compound lesions in DNA. J Biol Chem. 2005;280:39982–9. https://doi.org/10.1074/jbc.M507854200.
64. Goffeau A, De Hertogh B. ABC Transporters. In: Encyclopedia of Biological Chemistry: Academic Press; 2013. p. 7–11. https://doi.org/10.1016/B978-0-12-378630-2.00224-3.
65. Dean M, Rzhetsky A, Allikmets R. The human ATP-binding cassette (ABC) transporter superfamily. Genome Res. 2001;11:1156–66. https://doi.org/10.1101/gr.184901.
66. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015;12:780–6. https://doi.org/10.1038/nmeth.3454.
67. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT

sequencing data. Nat Methods. 2013;10:563–9. https://doi.org/10.1038/nmeth.2474.

68. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27:757–67. https://doi.org/10.1101/gr.214874.116.
69. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, Christiansen L, et al. In vitro, long-range sequence information for *de novo* genome assembly via transposase contiguity. Genome Res. 2014;24:2041–9. https://doi.org/10.1101/gr.178319.114.
70. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963. https://doi.org/10.1371/journal.pone.0112963.
71. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80. https://doi.org/10.1093/nar/27.2.573.
72. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265–8. https://doi.org/10.1093/nar/gkm286.
73. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2015. http://www.repeatmasker.org.
74. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7. https://doi.org/10.1159/000084979.
75. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics. 2005;21(suppl_1):i152–8. https://doi.org/10.1093/bioinformatics/bti1003.
76. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9:R7. https://doi.org/10.1186/gb-2008-9-1-r7.
77. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. 2006;34(suppl_2):W435–9. https://doi.org/10.1093/nar/gkl200.
78. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. Bioinformatics. 2004;20:2878–9. https://doi.org/10.1093/bioinformatics/bth315.
79. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59. https://doi.org/10.1186/1471-2105-5-59.
80. Guigo R. Assembling genes from predicted exons in linear time with dynamic programming. J Comput Biol. 1998;5:681–702. https://doi.org/10.1089/cmb.1998.5.681.
81. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268:78–94. https://doi.org/10.1006/jmbi.1997.0951.
82. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
83. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52. https://doi.org/10.1038/nbt.1883.
84. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66. https://doi.org/10.1093/nar/gkg770.
85. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5. https://doi.org/10.1038/nbt.1621.
86. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11. https://doi.org/10.1093/bioinformatics/btp120.
87. Kent WJ. BLAT-the BLAST-like alignment tool. Genome Res. 2002;12:656–64. https://doi.org/10.1101/gr.229202..
88. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14:988–95. https://doi.org/10.1101/gr.1865504.
89. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28:45–8. https://doi.org/10.1093/nar/28.1.45.
90. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30. https://doi.org/10.1093/nar/gkt1223.
91. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9. https://doi.org/10.1038/75556.
92. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.
93. Zdobnov EM, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17:847–8. https://doi.org/10.1093/bioinformatics/17.9.847.
94. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955. https://doi.org/10.1093/nar/25.5.955.
95. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25:1335–7. https://doi.org/10.1093/bioinformatics/btp157.
96. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 2005;33:D121–4. https://doi.org/10.1093/nar/gki081.
97. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45:W12–6. https://doi.org/10.1093/nar/gkx428.
98. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41:e74. https://doi.org/10.1093/nar/gkt006.
99. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89. https://doi.org/10.1101/gr.1224503.
100. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7. https://doi.org/10.1093/nar/gkh340.
101. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3. https://doi.org/10.1093/bioinformatics/btu033.
102. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91. https://doi.org/10.1093/molbev/msm088.
103. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22:1269–71. https://doi.org/10.1093/bioinformatics/btl097.
104. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21. https://doi.org/10.1093/sysbio/syq010.
105. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22:2472–9. https://doi.org/10.1093/molbev/msi237.
106. Yang Z, Nielsent R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19:908–17. https://doi.org/10.1093/oxfordjournals.molbev.a004148.
107. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8. https://doi.org/10.1038/s41592-018-0001-7.
108. Tang H, Krishnakumar V, Li J. jcvi: JCVI utility libraries. Zenodo. 2015. https://doi.org/10.5281/zenodo.31631.
109. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012;6:80–92. https://doi.org/10.4161/fly.19695.
110. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. Bioconductor. 2013. http://www.bioconductor.org/packages/2.11/bioc/html/topGO.html. Accessed 20 Dec 2019.

Publisher's Note