SI Appendix for:

The subgenomes show asymmetric expression of alleles in hybrid lineages of *Megalobrama amblycephala* × *Culter alburnus*

Short title: Asymmetric expression in hybrid fish lineages

Li Ren,[1,b] Wuhui Li,[1,b] Qinbo Qin,[1,b] He Dai,[2,b] Fengming Han,[2] Jun Xiao,[1] Xin Gao,[1] Jialin Cui,[1] Chang Wu,[1] Xiaojing Yan,[1] Guoliang Wang,[3] Guiming Liu,[3] Jia Liu,[1] Jiaming Li,[1] Zhong Wan,[4] Conghui Yang,[1] Chun Zhang,[1] Min Tao,[1] Jing Wang,[1] Kaikun Luo,[1] Shi Wang,[1] Fangzhou Hu,[1] Rurong Zhao,[1] Xuming Li,[2] Min Liu,[2] Hongkun Zheng,[2] Rong Zhou,[1] Yuqin Shu,[1] Yude Wang,[1] Qinfeng Liu,[1] Chenchen Tang,[1] Wei Duan,[1] and Shaojun Liu[1,a]

[1]State Key Laboratory of Developmental Biology of Freshwater Fish, College of Life Sciences, Hunan Normal University, Changsha, China

[2]Biomarker Technologies Corporation, Beijing 101300, China

[3]Beijing Agro-Biotechnology Research Center, Beijing Academy of Agriculture and Forestry Sciences, 100097, Beijing, China

[4]School of Mathematics and Statistics, Central South University, Changsha, 410083, Hunan, P.R. China

[a]Corresponding author

Professor Shaojun Liu

Email: lsj@hunnu.edu.cn

State Key Laboratory of Developmental Biology of Freshwater Fish, Hunan Normal University, Changsha 410081, China

Tel./Fax: +86-073188873074

[b]These authors contributed equally to this work.

Keywords: Hybridization, Whole-genome sequencing, Allelic exchange, Expressed recombinant transcripts, *Cis*- and *trans*-regulatory expression

1 SI methods and results of genome analysis

1.1 Sample preparation for genome and transcriptome sequencing

A wild, water-captured male adult of topmouth culter (*Culter alburnus,* TC) and a gynogenetic individual of blunt snout bream (*Megalobrama amblycephala*, BSB) were obtained for whole-genome sequencing. To obtain the gynogenetic BSB individual, BSB eggs were stimulated by UV-inactivated sperm of *Cyprinus carpio haematopterus* (Gong et al. 2019). One minute after stimulation, the eggs were placed in a water bath at 4-8℃ for cold shock treatment for 30-40 min. The purpose of this treatment was to suppress the discharge of the second polar body, resulting in diploid gynogenetic individuals. Then, we collected 26 fish from the Engineering Center of Polyploid Fish Breeding of the National Education Ministry at Hunan Normal University, China, for RNA-seq. The samples included the following fishes: three 2-year-old TC males whose testes were mature; three 2-year-old BSB females whose ovaries were mature; nine 2-year-old $2nF_1$-$2nF_3$ female hybrids of BSB (♀) × TC (♂) (abbreviated as BT); and nine 2-year-old $2nF_1$-$2nF_3$ female hybrids of TC (♀) × BSB (♂) (abbreviated as TB) (Fig. 1 A-H). Among these hybrid offspring, only "real hybrid" with two inbred parental chromosomes were identified with 45S rDNA and used in our study (Xiao et al. 2016). All tissue samples were excised carefully and subsequently stored at -80°C. The Administration of Affairs Concerning Animal Experimentation guidelines state that approval from the Science and Technology Bureau of China and the Department of Wildlife Administration is not necessary when the fish in question is not rare or near extinction, as is the case in this paper.

1.2 Sequencing and assembly

High-quality DNA was extracted from the muscle of one adult male of TC and one adult female of BSB using the DNeasy Blood and Tissue Kit (Qiagen). We constructed 13 paired-end libraries (220 bp to 17 kb) of TC and 10 paired-end libraries (220 bp-17 kb) of BSB to produce the raw data according to the Illumina standard operating procedure. A total of ~171 Gb (~194 X) and ~181 Gb (~161 X) of reads was obtained for every sequenced locus for TC and BSB, respectively (Supplemental Tables S1, S2). Before the beginning of assembly, sequencing adaptors of raw reads were removed. Then, contaminated reads (mitochondrial, bacterial and viral sequences, etc.) were screened by alignment to the NCBI-NR database using BWA (v 0.7.13) (Li and Durbin 2010) with default parameters. FastUniq (v1.1) (Xu et al. 2012) was used to remove the duplicated read pairs, while the low-quality Illumina reads were removed based on the following criteria: reads with ≥ 10% unidentified nucleotides (N); reads with ≥ 10 nt aligned to the adaptor, allowing ≤ 10% mismatches; and reads with ≥ 50% of bases having a Phred quality < 5.

Illumina reads (44.27 Gb and 51.96 Gb of data for TC and BSB, respectively) from the 220-bp library were selected to estimate genome size. This analysis was performed using

"kmer_freq_stat" software (developed by Biomarker Technologies) and showed a major peak at 48X and 38X with a 21-kmer frequency distribution for TC and BSB, respectively (Supplemental Fig. S1). Based on the formula Genome size = kmer_Number/Peak_Depth, the genome sizes of TC and BSB were estimated to be ~946.91 Mb and ~1.12 Gb, respectively. At the same time, we determined the genome size of TC and BSB with flow cytometry, using zebrafish (1.37 Gb, GRCz10 downloaded from the Ensembl database) as a reference standard and propidium iodide as the stain. Our flow cytometry analysis showed that the genome sizes of TC and BSB were ~ 1.12 Gb and 1.15 Gb, respectively.

Genome assembly was performed with ALLpaths-LG (Gnerre et al. 2011), joining the contigs into scaffolds with SSPACE (v 2.3) (Boetzer et al. 2011) and filling the gaps within scaffolds with GapCloser (v 1.12) in the SOAPdenovo package (Luo et al. 2012), resulting in two preassemblies: 1) a total length of 993.5 Mb and a contig and scaffold N50 size of 19.6 kb and 3.4 Mb, respectively, for TC and 2) a total length of 1.04 Gb and a contig and scaffold N50 size of 33.7 kb and 1.33 Mb, respectively, for BSB. We also sequenced the genomes of TC and BSB with ~6X (mean subread length: 8,523 bp, subread N50: 11,677 bp) and ~10X (mean subread length: 14,092 bp, subread N50: 19,206 bp) coverage, respectively, using long PacBio reads and used these reads with the PBJelly algorithm to close gaps in the ALLpaths-LG assembly above. The addition of long PacBio reads strikingly increased the length of the contig N50 from 19.6 kb to 72.2 kb for TC and from 33.7 kb to 142.7 kb for BSB (Table 1).

1.3 Assessment of assembly quality

To evaluate the integrity of the assembly, we assembled 45,055 and 13,871 unigenes (≥ 500 bp) with RNA-seq reads from multiple tissues using Trinity and aligned these unigenes to the assemblies of TC and BSB using BLAT (Kent 2002), yielding an average of more than 99% (44,787 for TC and 13,837 for BSB) coverage of these transcripts based on sequence similarity. The two assemblies were also analyzed using CEGMA and BUSCO (v. 2.0.1) with default parameters to evaluate their completeness. Our results suggested that more than 99% (456 for TC and 458 for BSB) conserved core eukaryotic genes (CGEs) and more than 98% (244 for TC and 246 for BSB) highly conserved CGEs were found in the TC and BSB assemblies. In addition, more than 90% of metazoan and vertebrate BUSCOs were covered by the assembled genomes (Supplemental Table S3). These evaluations suggested a high-quality assembly for TC and BSB.

1.4 Gene prediction and annotation

Protein-coding genes were predicted using *de novo*, protein homology-based and transcript-based approaches. To elaborate, GENSCAN (v 1.0) (Burge and Karlin 1997), Augustus (v 2.5.5) (Stanke and Waack 2003), GlimmerHMM (v 3.0.1) (Majoros et al. 2004),

GeneID v1.3 (Parra et al. 2000) and SNAP (Johnson et al. 2008) with the default parameters were used for *de novo* gene prediction, and all of these software packages were tested using the zebrafish (GCF_000002035.6) and grass carp (http://www.ncgr.ac.cn/grasscarp/) gene models before gene prediction. Homologous peptides from the zebrafish, grass carp, and common carp (GCF_000951615.1) genomes were aligned to our assembly to identify homologous genes with GeMoMa (v 1.4.2) (Keilwagen et al. 2016). The RNA-seq reads of multiple organs of TC and BSB, including muscle, blood, liver, testis, and ovary, were assembled into transcripts using Trinity (v 2.4.4) (Grabherr et al. 2011) and aligned to the corresponding repeat-masked TC and BSB assemblies with BLAT, and the gene structures were modeled using the Program for Automated Sequential Assignment. Consensus gene models were generated by integrating the *de novo* predictions and protein and transcript alignments using EVidenceModeler (Haas et al. 2008). Functional annotation was performed based on comparisons with the SwissProt, TrEMBL, InterPro, COG and KEGG protein databases. The gene ontology (GO) for each gene was assigned by Blast2GO (Conesa et al. 2005) based on NCBI databases (Supplemental Table S7). Overall, 30,835 (99.40%) and 28,930 (93.72%) of the protein-coding genes could be annotated, for which more than 70% of the predicted exonic regions were covered by aligning RNA-seq reads and assembled transcripts to the genome assembly using TopHat v2.1.2 and BLAT, respectively. Finally, we obtained 30,443 and 29,994 protein-coding genes (Supplemental Table S8). Then, comparisons of coding sequence (CDS) number, CDS length, intron length and gene length were performed in four Cyprinid fish (*Danio rerio*, *Ctenopharyngodon idellus*, BSB and TC) (Supplemental Fig. S3).

1.5 Noncoding RNA prediction

Noncoding RNAs play an important role in life processes, such as the rRNAs and tRNAs involved in mRNA translation. Therefore, the noncoding RNAs in our TC and BSB assemblies were predicted. The rRNA fragments were identified by aligning the rRNA template sequences (Pfam database v 22.0) using BLAST with an *e*-value of $1e^{-10}$ and identity cutoff of 95% or more. The tRNAscan-SE v 2.0 algorithms (Lowe and Chan 2016) with default parameters were applied to predict tRNA genes. The miRNA, snRNA and snoRNA genes were identified by mapping the genome sequences to the Rfam v11.0 database using INFERNAL (v 1.1) (Nawrocki and Eddy 2013). Finally, we identified 335 miRNAs, 628 rRNAs, 891 tRNAs, 106 snRNAs and 203 snoRNAs in TC, while 302 miRNAs, 474 rRNAs, 470 tRNAs, 92 snRNAs and 198 snoRNAs were detected in BSB (Supplemental Table S9).

1.6 Evolutionary analysis

Protein sequences of *D. rerio* (zebrafish), *C. idellus* (grass carp), *C. carpio* (common carp), *Xiphophorus maculatus* (platy fish), *Oryzias latipes* (medaka), *Fugu rubripes* (fugu),

Larimichthys crocea (large yellow croaker) and *Cynoglossus semilaevis* (flatfish) were downloaded from the Ensembl or GenBank database. The proteomes of the above eight species and our proteomes of TC and BSB, comprising 26,003 and 26,566 protein sequences, respectively, were clustered into 27,498 orthologous groups using OrthoMCL (Li et al. 2003) based on an all-to-all BLASTP strategy with an *e*-value of $1e^{-5}$ and a Markov chain clustering (MCL) default inflation parameter of 1.5. Based on the clustering results, TC- and BSB-specific gene clusters were obtained and annotated. To infer phylogenetic relationships, we extracted 796 single-copy clusters from all ten species, and multiple sequence alignment of proteins for each cluster was performed by MUSCLE (Edgar 2004) (Fig. 2A; Supplemental Table S10); all the alignments were combined into one supergene to construct a phylogenetic tree using RAxML (v 7.0.0) (Stamatakis 2014) with 1,000 rapid bootstrap analyses followed by a search of the best-scoring maximum-likelihood (ML) tree in a single run (Fig. 2A). Finally, divergence time was estimated using MCMCTree from the PAML package (Yang 2007), together with the molecular clock model. Two reference calibrated time points provided by the TimeTree database (http://timetree.org/) were used to assess the divergence times of nodes of interest. Expansion and contraction of OrthoMCL-derived homologous clusters were determined by CAFÉ (v 2.1) calculation based on changes in gene cluster size, utilizing the phylogeny and the species divergence time.

1.7 Diversifying selection analysis

Orthologous gene pairs between TC and BSB were determined by best reciprocal BLAST hits with an *e*-value of $1e^{-5}$. The K_s value between the orthologous pairs was calculated by the yn00 program in the PAML package. The all-to-all BLASTP method was used to detect the orthologous genes in the TC-*C. idellus*, TC-BSB, BSB-*C. idellus*, *D. rerio*-TC, *D. rerio*-*C. idellus* and *D. rerio*-BSB pairs with an *e*-value threshold of $1e^{-5}$. Homologous blocks were detected using MCScanX (Wang et al. 2012), and the 4DTV (transversions at fourfold degenerate sites) values of the blocks in the CDS alignments were calculated using the HKY model. The distribution of K_s values was used to determine the events of species divergence (Fig. 2B).

1.8 Genetic map construction and scaffold anchoring

High-quality genomic DNA extracted from 106 individuals of TC (including one male parent, one female parent, and 104 F_1 progenies) was used to construct Illumina sequencing libraries following the manufacturer's protocol. The genomic libraries were sequenced on an Illumina HiSeq 2500 system, and a total of 133.1 M reads were obtained. The average sequencing depths of sequenced loci per parent and per progeny were ~28.0X and ~6.5X, respectively. Genotyping and evaluation of the quality of genetic markers were performed as

described in Sun et al. After filtering, a total of 6,515 high-quality SNP markers were identified and used to construct the genetic map with the double pseudo-testcross strategy using HighMap software (Liu et al. 2014). Using the nearest-neighbor method, a total of 6,377 markers were clustered into 24 linkage groups with a total genetic length of 3,867.4 cM. The SNP markers were then aligned to the TC assembled scaffolds, and only uniquely aligned markers were used to anchor and orient the scaffolds onto the 24 TC pseudochromosomes.

1.9 Hi-C assembly of the BSB genome

According to the Hi-C procedure, nuclear DNA from the blood of the BSB individual was cross-linked and then cut with a restriction enzyme, leaving pairs of distally located but physically interacting DNA molecules attached to one another. The sticky ends of these digested fragments were biotinylated and then ligated to each other to form chimeric circles. Biotinylated circles, which are chimeras physically associated with DNA molecules from the original crosslinking, were enriched, sheared and sequenced. In our study, a total of 202.3 million clean Hi-C read pairs (60.58 Gb) with approximately 55.64-fold coverage of the BSB genome were obtained. Of these reads, 87.32% were mapped to the BSB genome, and 58.53% showed unique alignment to the genome. After filtering, we obtained 88.53 M valid interaction pairs for the chromosome-level assembly. Subsequently, the scaffolds within the BSB assembly were broken into 50-kb fragments and clustered by LACHESIS software (Burton et al. 2013) using valid interaction read pairs. Finally, 2,867 scaffolds with a total length of 1,048 Mb were anchored to the BSB chromosomes, of which 1,901 scaffolds (983.74 Mb in length) were assigned, ordered and oriented to 24 chromosome-level groups (Supplemental Fig. S2). The syntenic relationships among the TC, BSB and *D. rerio* chromosomes were constructed (Fig. 2C; Supplemental Fig. S4).

2 SI methods and results of allelic recombinant analysis

2.1 RNA isolation and transcriptome sequencing

To sequence the transcriptomes of reciprocal cross hybrids and their inbred parents, total RNA was isolated and purified from the liver (24 samples), muscle (6 samples) and gonad (ovary) (6 samples) by a TRIzol extraction method. The RNA concentration was measured using NanoDrop technology. Total RNA samples were treated with DNase I (Invitrogen) to remove any contaminating genomic DNA. The purified RNA was quantified using a 2100 Bioanalyzer system (Agilent, Santa Clara, CA, USA). We fragmented 1 μg of isolated mRNA with fragmentation buffer. The resulting short fragments were reverse transcribed and amplified to produce cDNA. An Illumina RNA-seq library was prepared according to a standard high-throughput method (Dillies et al. 2013). The cDNA library concentration and quality were assessed by the Agilent Bioanalyzer 2100 system, after which the library was sequenced with

a paired-end setting using the Illumina HiSeq 2000/4000/X Ten platform. An RNA-seq experiment was conducted with three biological replicates. Then, the raw reads containing adapters and poly-N tails and of low quality were removed using a custom computational script (see Supplemental Script 1: filtering raw reads). The high-quality reads were used in the next analysis.

2.2 Ortholog identification and allelic recombinant sequence detection by Illumina sequencing

The orthologous gene alignments between BSB and TC were obtained from all-against-all reciprocal BLASTP (v 2.2.26) comparisons with the parameters "-e 10-5 –F F –v 1 –m 8" based on the protein sequences. The alignments were used in MCScanX to determine syntenic blocks, which were then displayed as a schematic diagram created with Circos (v 0.69-6) (Krzywinski et al. 2009) (Fig. 2C; Supplemental Fig. S4). Transcripts lacking annotated CDSs and those with lengths < 100 bp were discarded. A total of 12,322 orthologous gene pairs were selected based on the best match scores and filtering.

Allelic recombinant genes in hybrids originate from the recombination of inbred parental genome sequences. The expression of these proteins can be detected by Illumina sequencing based on a few methods (Francesc et al. 2009; Sabio and Davis 2010; Liu et al. 2016). The high-quality Illumina reads of hybrids in the above analysis were mapped to the mixed genome of two parents using STAR (v 2.4.0) with the parameter "--outFilterMismatchNoverReadLmax 0.01/0.02/0.03 --chimSegmentMin 50/70/90/105/120" (Dobin et al. 2013). Focusing on mapping files, pairwise alignments of orthologous gene pairs were used to assess the differences between hybrids. The "--outFilterMismatchNoverReadLmax 0.02" parameter was selected based on low mutation rates and sequencing errors. The other parameter, "—chimSegmentMin", was related to the standard for recombinant reads, which required at a minimum that fragments be identified in two paired-end reads. Regarding the two styles of paired-end (PE) libraries used in our analysis, the parameter "--chimSegmentMin 70" with 100 PE reads*2 samples and "--chimSegmentMin 105" with 150 PE reads*2 samples was selected to detect the recombinant reads in hybrids. Then, only the reads distributed in orthologous gene pairs between BSB and TC were considered recombinant reads. After removing the repeated recombinant reads in one orthologous gene pair, the distribution of recombinant events in reciprocal cross offspring was enriched in the corresponding genes (Branchetti et al. 2013).

2.3 PCR validation of allelic recombinant genes

From the short fragments obtained by Illumina sequencing, 19 genes were selected based on the above prediction results. The total DNA of BSB, TC, TBF_1, BTF_1, TBF_3 and BTF_3 was extracted in triplicate. The set of primers used in PCRs and clone numbers are provided in

Supplemental Table S14. Amplified products were evaluated using an ABI 3730 DNA Analyzer (Applied Biosystems, Carlsbad, CA, USA). To determine sequence homology and variation among the fragments, sequences were aligned using BioEdit (Huang et al. 2012) and CLUSTALW2 (Huang and Parmacek 2012). The *tgfb1b* sequences of grass carp and zebrafish were downloaded from the NCBI database (accession nos. EU099588.1 and XM_687246.8).

2.4 Library construction and PacBio sequencing

To verify the accuracy of the allelic recombinant genes, the unassembled long reads from PacBio sequencing were considered useful for studying allelic recombinant events. Thus, the long transcripts of reciprocal cross F_3 hybrids were obtained with PacBio SMRT sequencing. The total RNA of TBF_3 and BTF_3 from five tissues, including liver, muscle, gonad, heart and hypophysis, was obtained and mixed in equal amounts. The RNA was reverse transcribed using the SMARTer PCR cDNA Synthesis Kit, and PCR amplification was performed using KAPA HiFi PCR Kits. The PCR product (size = 0.5-6 K and > 6 K) was selected based on the agarose gel electrophoresis method. Then, libraries were constructed from these cDNA products using the SMRTbell Template Prep Kit 1.0. After library preparation, the library template and enzyme mixture were used in the PacBio Sequel™ system for sequencing.

2.5 Allelic recombinant gene detection by genome-wide long-read alignments

After obtaining the sequencing data, low-quality data (adapter sequences, length of subreads < 50 bp, accuracy rate < 0.75) were deleted from the raw data. Sequence reads from the PacBio RS SMRT chip were processed through PacBio's SMRT-Portal analysis suite to generate circular consensus sequences (CCSs). To obtain more accurate reads, the reads (number of cycles of CCS > 1 and accuracy > 0.8) were used to obtain full-length reads (lengths > 300 bp, poly (A) tails, 5' primers and 3' primers) based on the SMRT Iso-Seq analysis pipeline (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/).

We detected allelic recombinant transcripts in TBF_3 and BTF_3 by the following steps: a recombinant full-length read was split into two fragments that were mapped to orthologous gene pairs between BSB and TC. To find the paired fragment alignment of a recombinant long read, we aligned reads to the reference genome (mixed genome of two parents) by GMAP with the parameters identity > 90% and coverage ≥ 99% (Parmacek and Epstein 2013). In addition, we examined all mutant pairs of fragment alignments and defined them as candidate expressed recombinant transcripts if the following conditions were met: (i) both aligned fragments were of sufficient length (> 200 bp); (ii) two fragments were mapped to an orthologous gene pair; and (iii) exon regions were distributed in both aligned fragments. Long repetitive genomic elements and TGS long reads with sequencing errors were eliminated as described by Jason et al. (Nguyen et al. 2013).

The advantages of sequencing technology, including the high throughput of Illumina sequencing (Li et al. 2012; Shen et al. 2011), the long reads in PacBio sequencing (Rhoads and Au 2015), and the high accuracy of Sanger sequencing (Huang et al. 2009; Parmacek and Epstein 2009), and the disadvantages, including short reads in Illumina sequencing (Rong et al. 2012), high error rate in PacBio sequencing (Rong et al. 2012), and the low sequencing data content and short fragments in Sanger sequencing (Huang et al. 2009; Parmacek and Epstein 2009), compelled us to combine the three methods for predicting recombinant events in hybrids. As a comparison of Illumina and PacBio sequencing, the differences in the number and length of sequencing reads between the two methods may be the main causes leading to a large number of allelic recombinant events in Illumina sequencing and few such events in PacBio sequencing. However, the shared allelic recombinant events detected between the methods proved the accuracy of these methods in predicting allelic recombinant events.

3 SI methods and results of expression analysis

3.1 Mapping of RNA-seq data

All Illumina reads of BSB and TC were aligned to the BSB and TC genomes using STAR (v 2.4.0) with the default parameters, respectively (Dobin et al. 2013). In addition, the other RNA-seq reads of reciprocal cross hybrids were aligned to the two reference genomes of BSB and TC. The number of mapped reads in each gene was calculated with some in-house Perl scripts (see Supplemental Scripts 2: calculating mapped reads). Therefore, two mapping results based on alignment to the two reference genomes were obtained in hybrid offspring. The silent genes were screened in comparisons of both parents with each hybrid, with a threshold of no mapped reads detected in hybrids and mapped read counts ≥ 5 in both parents, with three biological replicates. To avoid the negative effect of expression noise, the expression analysis was performed on only filtered genes with mapped read counts ≥ 5 in all three biological replicates in each comparison. To further avoid the interference of expression noise, the expression values of three biological replicates were screened with a mean $\pm$ 2 standard deviation (SD) threshold in each gene (Quackenbush 2002). Moreover, the total expression value was normalized based on the ratio of the number of mapped reads for each gene to the total number of mapped reads for the entire genome. For the transcriptome data of the two parents, 68.10 million clean reads (84.68%) from the liver of BSB and 64.34 million clean reads (84.91%) from the liver of TC were obtained from mapping to the respective reference genome. Additionally, 626.33 million clean reads and 627.92 million clean reads in the hybrid mapped to the two parental genomes, respectively.

3.2 Differential expression analysis

To investigate the expression levels of the two parents and hybrids in one comparison, we

set the "average parental expression" (APE) value (H_{APE}) related to orthologous gene pairs between BSB (B) and TC (T). The value was equal to the average expression level of orthologous gene pairs in the parents ($H_{APE} = (B + T)/2$). Moreover, the real expression level of the hybrid (H_{real}) was equal to the average of the two expression values aligned to the two parental genomes ($H_{real} = (H_B + H_T)/2$). To compare H_{APE} with H_{real}, the clear changes in expression after hybridization were described in one comparison. The expression values of each gene with an FDR ≤ 0.05 and a $\log_2$-transformed ratio of 1.25 were considered differentially expressed genes. The DE analysis was performed with the DESeq2 package in the R program (v 1.10.0) (R Foundation for Statistical Computing, Vienna, Austria) (Varet et al. 2016).

3.3 Analysis of expression dominance

To compare the expression levels of hybrids with those of both inbred parents, the mode of the expression level relative to the total expression level of orthologous genes between BSB and TC was used. The comparison for each gene of the hybrid was performed in three steps: 1) the BSB and hybrid expression values were obtained by alignment with the BSB genome, and significant DE was detected. 2) The TC and hybrid expression values were obtained by alignment with the TC genome, and significant DE was detected. 3) Information on the significant DE between the two parents based on orthologous gene pairs between BSB and TC was obtained. The analysis of significant DE was performed with Fisher's exact tests (FDR) (Robinson et al. 2010), and the distribution of *P*-values was controlled for with an FDR by the BH method at $\alpha = 0.05$ (Francesc et al. 2009). Genes in the results of the above three comparisons were classified as additive, BSB/TC dominant, underdominant, and overdominant based on the magnitude of the expression difference, as described by Gibson et al. (Gibson et al. 2004).

3.4 Species-specific SNP identification

To investigate the expression levels of parent-of-origin genes (allelic expression), the LASTZ pairwise alignment tool (v 1.02.00) (Harris 2007) with default parameters was used to obtain the corresponding loci from the orthologous gene pairs between BSB and TC. The InDel loci were discarded, and only the aligned loci with the best match scores were used for the next analysis. Then, the SNPs were collected using the SNP Calling pipeline with GATK (v 3.8) based on the results of parental transcriptome mapping to the respective genome as described above (McKenna et al. 2010). According to the comparison of SNPs and other loci in orthologous gene pairs between BSB and TC, the differential loci, including heterozygous and homozygous loci, were considered species-specific SNPs, as in Schaefke *et al.* (McManus et al. 2010; Schaefke et al. 2013). To prevent the negative effect of sequencing and mapping, the screening of species-specific SNPs must be checked for consistency with three biological

replicates, and loci possessing read counts ≥1 in all accessions and biological replicates in each comparison are retained. After screening 9,753 gene pairs, 103,190 SNPs, including 268 heterozygous loci in TC and 41 heterozygous loci in BSB, were obtained from 60,909,895 and 52,151,477 clean reads mapped to the BSB and TC reference genomes, respectively (Supplemental Fig. S27).

3.5 Detection of allelic expression levels in hybrids

To describe the allelic expression in hybrids, the hybrid transcriptome mapping results (bam files) described above were used in our next analysis. The map files of each hybrid were divided into two categories based on the two different parental reference genomes. The BSB/TC allelic reads in the hybrid were calculated using some in-house Perl scripts based on corresponding BSB/TC species-specific SNPs in the corresponding map files (see Supplemental Scripts 3: calculating allelic reads with SNPs). In addition, the expression levels of parents in the allelic analysis were also detected based on the respective reference genome related to species-specific SNPs. In the above analysis, to remove the negative effect of mutation sites in the hybrid in biological replicates, if the mapped reads of a species-specific SNP did not comply with the mean ± 2 SD threshold in the three biological replicates, the abnormal value was discarded when estimating the BSB and TC allelic expression levels. Then, the total numbers of BSB and TC alleles in each gene were normalized based on the ratio of the number of mapped reads for each gene to the total number of mapped reads for the entire genome (Quackenbush 2002). The sum of BSB/TC allelic reads for all species-specific SNPs of each gene was used to assess BSB/TC allelic expression.

3.6 Allelic expression silencing and bias

After obtaining BSB/TC allelic expression in hybrids, we detected allelic expression silencing in different tissues and generations of reciprocal cross hybrids. This screening complied with the thresholds in which the total number of BSB and TC allelic reads for each gene in both parents was larger than four (≥ 5), either the number of BSB allelic reads or the number of TC allelic reads in the hybrid was zero in the three biological replicates, and the total number of other BSB or TC allelic reads was greater than four (≥ 5).

Equal expression of both alleles and unbalanced expression of the two alleles were used for simple classification of hybrids. For this analysis, expression levels were considered “equal” if the $\log_2$-transformed ratio of allelic expression of TC and BSB ($|\log_2 (TC/BSB)|$) in hybrid samples was less than 1. Genes with $|\log_2 (TC/BSB)|$ values in the hybrid sample greater than or equal to 1 were considered “unbalanced”, and TC and BSB allelic expression bias was classified based on the plus-minus of these values. Additionally, we chose another APE value for allelic analysis based on the $\log_2$-transformed ratio of allelic expression of TC and BSB in

the two parental samples, which was used as a reference for the beginning of changes in allelic expression.

3.7 *Cis*- and *trans*-regulatory differences underlying expression divergence between BSB and TC

To further investigate the mechanism of expression divergence that arises from changes in TC and BSB alleles, *cis*- and/or *trans*-regulatory patterns were established based on significant differences between TC and BSB in parents and hybrids, as described in McManus *et al.* (McManus et al. 2010). In the above analysis, binomial exact tests and Fisher's exact tests, with an FDR of 5%, were used to identify genes with significant differences. In addition, we further sorted the *cis*- and/or *trans*-regulatory results based on the plus-minus $\log_2$-transformed ratio of allelic expression of TC to BSB in parents or hybrids.

3.8 Correlation analysis of *cis*- and *trans*-regulatory expression, K_a/K_s and allelic recombinant genes

To investigate the underlying regulatory mechanism of *cis*- and *trans*-regulatory expression, correlation analysis was performed using Pearson's rank correlation coefficients and Student's *t* test in GraphPad Prism (v 7.0) between ω and $|\log_2 (TC/BSB)|$ in both the parents and the hybrids. Furthermore, another correlation analysis based on the Spearman method was used to obtain the correlation coefficient between the absolute value of the difference in CDS length and $|\log_2 (TC/BSB)|$ in both the parents and the hybrids. To further investigate *cis*- and *trans*-regulatory expression of genes less likely than others to be influenced by allelic recombinant events, a chi-square test was used to assess the correlation between the distribution of allelic recombinant events and the "*cis* only" gene in each sample.

References

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**: 578-579. doi: 10.1093/bioinformatics/btq683

Branchetti E, Poggio P, Sainger R, Shang E, Grau JB, Jackson BM, Lai EK, Parmacek MS, Gorman RC, Gorman JH et al. 2013. Oxidative stress modulates vascular smooth muscle cell phenotype via CTGF in thoracic aortic aneurysm. *Cardiovasr Res* **100**: 316-324. doi: 10.1093/cvr/cvt205

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen. *J Mol Biol* **268**: 78-94. doi: 10.1006/jmbi.1997.0951

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat*

Biotechnol **31**: 1119. doi: 10.1038/nbt.2727

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674-3676. doi: 10.1093/bioinformatics/bti610

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **14**: 671-683. doi: 10.1093/bib/bbs046

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21. doi: 10.1093/bioinformatics/bts635

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797. doi: 10.1093/nar/gkh340

Francesc P, Andy B, Jeanclaude FR, Martin FH, Pierrick H, Lorenzo C. 2009. Polyploid fish and shellfish: Production, biology and applications to aquaculture for performance improvement and genetic containment. *Aquaculture* **293**: 125-156. doi: 10.1016/j.aquaculture.2009.04.036

Gibson G, Rileyberger R, Harshman L, Kopp A, Vacha S, Nuzhdin S, Wayne M. 2004. Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. *Genetics* **167**: 1791-1799. doi: 10.1534/genetics.104.026583

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* **108**: 1513-1518. doi: 10.1073/pnas.1017351108

Gong DB, Xu LH, Wu C, Wang S, Liu QF, Cao L, Mao ZW, Wang YD, Hu FZ, Zhou R et al. 2019. Two types of gynogenetic blunt snout bream derived from different sperm. *Aquaculture* **511**: 734250. doi: 10.1016/j.aquaculture.2019.734250

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644-652. doi: 10.1038/nbt.1883

Haas BJ, Salzberg SL, Wei Z, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**: R7. doi: 10.1186/gb-2008-9-1-r7

Harris RS. 2007. *Improved pairwise alignment of genomic DNA*. PhD thesis, Pennsylvania State University, University Park, PA.

Huang J, Elicker J, Bowens N, Liu X, Cheng L, Cappola TP, Zhu X, Parmacek MS. 2012.

Myocardin regulates BMP10 expression and is required for heart development. *J Clin Invest* **122**: 3678-3691. doi: 10.1172/JCI63635

Huang J, Min Lu M, Cheng L, Yuan LJ, Zhu X, Stout AL, Chen M, Li J, Parmacek MS. 2009. Myocardin is required for cardiomyocyte survival and maintenance of heart function. *Proc Natl Acad Sci USA* **106**: 18734-18739. doi: 10.1073/pnas.0910749106

Huang J, Parmacek MS. 2012. Modulation of smooth muscle cell phenotype: the other side of the story. *Circ Res* **111**: 659-661. doi: 10.1161/CIRCRESAHA.112.277368

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'donnell CJ, De Bakker PI. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**: 2938-2939. doi: 10.1093/bioinformatics/btn564

Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. 2016. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* **44**: e89. doi: 10.1093/nar/gkw092

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664. doi: 10.1101/gr.229202

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res*. doi: 10.1101/gr.092759.109

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595. doi: 10.1093/bioinformatics/btp698

Li J, Bowens N, Cheng L, Zhu X, Chen M, Hannenhalli S, Cappola TP, Parmacek MS. 2012. Myocardin-like protein 2 regulates TGFbeta signaling in embryonic stem cells and the developing vasculature. *Development* **139**: 3531-3542. doi: 10.1242/dev.082222

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178-2189. doi: 10.1101/gr.1224503

Liu D, Ma C, Hong W, Huang L, Liu M, Liu H, Zeng H, Deng D, Xin H, Song J. 2014. Construction and analysis of high-density linkage map using high-throughput sequencing data. *Plos One* **9**: e98855. doi: 10.1371/journal.pone.0098855

Liu S, Luo J, Chai J, Ren L, Zhou Y, Huang F, Liu X, Chen Y, Zhang C, Tao M et al. 2016. Genomic incompatibilities in the diploid and tetraploid offspring of the goldfish x common carp cross. *Proc Natl Acad Sci USA* **113**: 1327-32. doi: 10.1073/pnas.1512955113

Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* **44**: W54-W57. doi: 10.1093/nar/gkw413

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Qi P, Liu Y. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**: 18-18. doi: 10.1186/2047-217X-1-18

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**: 2878-2879. doi: 10.1093/bioinformatics/bth315

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303. doi: 10.1101/gr.107524.110

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**: 816-825. doi: 10.1101/gr.102491.109

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933-2935. doi: 10.1093/bioinformatics/btt509

Nguyen AT, Gomez D, Bell RD, Campbell JH, Clowes AW, Gabbiani G, Giachelli CM, Parmacek MS, Raines EW, Rusch NJ et al. 2013. Smooth muscle cell plasticity: fact or fiction? *Circ Res* **112**: 17-22. doi: 10.1161/CIRCRESAHA.112.281048

Parmacek MS, Epstein JA. 2009. Cardiomyocyte renewal. *N Engl J Med* **361**: 86-88. doi: 10.1056/NEJMcibr0903347

Parmacek MS., Epstein JA. 2013. An epigenetic roadmap for cardiomyocyte differentiation. Circ Res **112**: 881-883. doi: 10.1161/CIRCRESAHA.113.301134

Parra G, Blanco E, Guigo R. 2000. GeneID in Drosophila. *Genome Res* **10**: 511-515. doi: 10.1101/gr.10.4.511

Quackenbush J. 2002. Microarray data normalization and transformation. *Nat Genet* **32**: 496-501. doi: 10.1038/ng1032

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**: 278-289. doi: 10.1016/j.gpb.2015.08.002

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140. doi: 10.1093/bioinformatics/btp616

Rong L, Liu J, Qi Y, Graham AM, Parmacek MS, Li S. 2012. GATA-6 promotes cell survival by up-regulating BMP-2 expression during embryonic stem cell differentiation. *Mol Biol Cell* **23**: 3754-3763. doi: 10.1091/mbc.E12-04-031

Sabio G, Davis RJ. 2010. cJun NH2-terminal kinase 1 (JNK1): roles in metabolic regulation of insulin resistance. *Trends Biochem Sci* **35**: 490. doi: 10.1016/j.tibs.2010.04.004

Schaefke B, Emerson JJ, Wang TY, Lu MY, Hsieh LC, Li WH. 2013. Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. *Mol Biol Evol* **30**: 2121-2133. doi: 10.1093/molbev/mst114

Shen D, Li J, Lepore JJ, Anderson TJ, Sinha S, Lin AY, Cheng L, Cohen ED, Roberts JD, Jr.,

Dedhar S et al. 2011. Aortic aneurysm generation in mice with targeted deletion of integrin-linked kinase in vascular smooth muscle cells. *Circ Res* **109**: 616-628. doi: 10.1161/CIRCRESAHA.110.239343

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313. doi: 10.1093/bioinformatics/btu033

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**: ii215-225. doi: 10.1093/bioinformatics/btu033

Varet H, Brillet-Gueguen L, Coppee JY, Dillies MA. 2016. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLoS One* **11**: e0157022. doi: 10.1371/journal.pone.0157022

Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**: e49-e49. doi: 10.1093/nar/gkr1293

Xiao J, Hu F, Luo K, Li W, Liu S. 2016. Unique nucleolar dominance patterns in distant hybrid lineage derived from *Megalobrama Amblycephala* × *Culter Alburnus*. *BMC Genet* **17**: 150. doi: 10.1186/s12863-016-0457-3.

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *Plos One* **7**: e52249. doi: 10.1371/journal.pone.0052249

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591. doi: 10.1093/molbev/msm088