Genomic basis of striking fin shapes and colours in the fighting fish

Le Wang[1#], Fei Sun[1#], Zi Yi Wan[1$], Baoqing Ye[1$], Yanfei Wen[1], Huiming Liu[1], Zituo Yang[1], Hongyan Pang[1], Zining Meng[2], Bin Fan[3], Yuzer Alfiko[4], Yubang Shen[5], Bin Bai[1], May Shu Qing Lee[1], Francesc Piferrer[6,*], Manfred Schartl[7,8,*], Axel Meyer[9,*] & Gen Hua Yue[1,10,11*]

[1] Molecular Population Genetics & Breeding Group, Temasek Life Sciences Laboratory, Singapore 117604, Singapore

[2] School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

[3] Department of Food and Environmental Engineering, Yangjiang Polytechnic, Yangjiang 529500, China

[4] Biotech Lab, Wilmar International, Jakarta, Indonesia

[5] Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Shanghai Ocean University, Shanghai 201306, China

[6] Institute of Marine Sciences (ICM), Spanish National Research Council (CSIC), 08003 Barcelona, Spain

[7] Developmental Biochemistry, Biocenter, University of Wuerzburg, 97074 Wuerzburg, Germany

[8] The Xiphophorus Genetic Stock Center, Department of Chemistry and Biochemistry, Texas State University, San Marcos, Texas 78666, USA

[9] Department of Biology, University of Konstanz, 78457 Konstanz, Germany

[10] Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore

[11] School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore

[#] Contributed equally to this study

[$] Contributed equally to this study

[*] Corresponding authors

FP: piferrer@icm.csic.es

MS: phch1@biozentrum.uni-wuerzburg.de

AM: axel.meyer@uni-konstanz.de

GHY: genhua@tll.org.sg

Abstract

Resolving the genomic basis underlying phenotypic variations is a question of great importance in evolutionary biology. However, understanding how genotypes determine the phenotypes is still challenging. Centuries of artificial selective breeding for beauty and aggression resulted in a plethora of colors, long fin varieties, and hyper-aggressive behavior in the air-breathing Siamese fighting fish (*Betta splendens*), supplying an excellent system for studying the genomic basis of phenotypic variations. Combining whole genome sequencing, QTL mapping, genome-wide association studies and genome editing, we investigated the genomic basis of huge morphological variation in fins and striking differences in coloration in the fighting fish. Results revealed that the double tail, elephant ear, albino and fin spot mutants each were determined by single major-effect loci. The elephant ear phenotype was likely related to differential expression of a potassium ion channel gene, *kcnh8*. The albinotic phenotype was likely linked to a cis-regulatory element acting on the *mitfa* gene and the double tail mutant was suggested to be caused by a deletion in a *zic1*/*zic4* co-enhancer. Our data highlight that major loci and cis-regulatory elements play important roles in bringing about phenotypic innovations and establish Bettas as new powerful model to study the genomic basis of evolved changes.

Key words: Domestication, evolution, major-effect loci, cis-regulation; *mitfa*; *zic1*/*zic4*

Introduction

Already in the "On the Origin of Species" (1859) and later in the "*The Variation of Animals and Plants under Domestication*" (1868) Charles Darwin recognized that the same processes of selection that act in nature also apply to selective breeding where they are sped up orders of magnitude by breeders' goals to obtain particular traits. Obviously, Darwin could not know the genomic basis that underlies the selected traits. This began to change, particularly in the last decade, as genome sequences could be obtained for more species. Yet, we are still only at the beginning of understanding how the genotype controls and determines the phenotype (Frazer, et al. 2009). Both, mutations in coding sequences and polymorphisms in non-coding sequences are now known to play important roles in generating phenotypic variation (Wittkopp and Kalay 2012; Andersson, et al. 2015; Petit, et al. 2017; Kemble, et al. 2019). Beyond a handful of genetic or developmental model systems (Lehner 2013), organisms under artificial selection not only provide outstandingly useful phenotypes but also permit studying the connection between evolutionary changes and their genomic bases. Knowledge about the bridge between genotypes and phenotypes poses a question, whether the same genomic mechanisms, as Darwin suggested, both in evolution as well as in animal breeding, are at work bringing about innovations (Shapiro, et al. 2004; Karlsson, et al. 2007; Rubin, et al. 2010). Importantly, assays including CRISPR-Cas mediated genome modification that allow establishing functional associations between genotype and phenotype only became recently available.

The Siamese fighting fish (*Betta splendens*), one of the most popular ornamental fishes worldwide, is well known for its aggressive behaviour (Simpson 1968), extremely diverse colour patterns, and huge variation in fin shapes (Lucas 1968). It belongs to the anabantoid fishes, characterized by a modified gill skeleton that forms the labyrinth organ, which permits air breathing as these fish tend to live in oxygen-deprived waters. Males build bubble nests,

and perform complex courting and parental care behaviours (Lucas 1968; Rüber, et al. 2004; Monvises, et al. 2009). It is a short lived species and its generation interval is only 5-6 months, with each spawning producing up to several hundred eggs (Monvises, et al. 2009). The initiation of domestication of fighting fish has been documented to have occurred as early as six hundred years ago, with the purpose of using these fish in staged fighting contests by the Siamese in the current Thailand, leading to the 'Plakat' betta (Smith 1945). Selection on other display traits, mainly including coloration and fin shapes, has a more recent origin traced back to the middle of the nineteenth century, and was prompted by the use of these fish in exhibition contests (Lucas 1968). In the past decades since these fish became the object of a worldwide aquarium hobby, most artificial selection in fighting fish has focused on modifying the spectacular body colorations and the overgrowth of fins (supplementary **fig. S1**). Therefore, Siamese fighting fish constitute an unparalleled system for identifying genetic variants underlying both simple and highly complex morphological traits to increase the understanding of the genetic basis of phenotypic variation, between "natural mutants" and domesticated "sports"

Results and discussion

Genome assembly and annotation

We sequenced a homozygous yellow single-tail female and a homozygous transparent double-tail male fighting fish **(**supplementary **fig. S2** and **table S1)**, each with over 120-fold genome coverage. Genome assembly sizes for the female and male were 424.9 Mb and 411.1 Mb, respectively (supplementary **note 1**). The contig and scaffold N50 sizes for the female were 21.3 kb and 2.1 Mb, respectively, and for the male, 17.3 kb and 1.9 Mb, respectively **(**supplementary **table S2)**. Both the female and male assemblies showed complete and single-

copy BUSCO scores of > 95 %, indicative of high quality **(**supplementary **table S3** and **fig. S3 & S4)**. A total of 22,977 protein-coding genes were predicted. Transposable elements and other types of repetitive elements together accounted for only 10.7 % of the genome sequences **(**supplementary **table S4)**, a relatively small fraction of the genome compared to other teleosts (Malmstrøm, et al. 2017). We also annotated 160,595 CNEs and 1,059 lncRNAs in the genome, with a total length of 34.1 Mb (~ 8.0 %) and 1.47 Mb (~ 0.35 %), respectively. In comparison to nine other fish genomes, the fighting fish therefore has, besides the pufferfishes, the shortest mean intergenic regions and the lowest overall proportion of non-coding elements within genes, indicating that it has a very compact genome **(**supplementary **fig. S5)**. Using two high-density linkage maps, 94.9 % and 95.3 % of scaffolds were anchored on 21 linkage groups corresponding to the chromosomes of the female and male karyotypes, respectively **(**supplementary **tables S5-S7** and **fig. S6)**. Our genome assembly is an important addition to those published recently for the fighting fish (Fan, et al. 2018; Prost, et al. 2020). Therefore, they supply a useful tool for downstream genetic and genomic studies.

Genetic diversity and population structure of the fighting fish

We sequenced the whole genomes of domesticated fish of diverse coloration and fin traits as well as several wild fish exhibiting ancestral phenotypes, as indirect ancestors (supplementary **table S8**). Domesticated species tend to lose genetic diversity compared to their wild ancestors. Here, based on whole-genome resequencing data, we observed that the overall genetic diversity in domesticated fighting fish was reduced approximately 10fold when compared to the wild ancestors (nucleotide diversity: 0.0003 vs 0.0025, $P < 10^{-8}$ for *t*-test, estimated using VCFtools (Danecek, et al. 2011); unbiased nucleotide diversity: 0.0004 vs 0.0033, $P < 10^{-8}$ for *t*-test, estimated using pixy (Korunes and Samuk 2021)). Remarkably, the average number of rare SNPs with a cut-off value of minor allele frequency of 0.01 was even more decreased, nearly 80 times (4,290 vs 349,824, $P < 10^{-7}$ for *t*-test; **fig. 1A & B)**. Such rapid loss of rare alleles

during domestication is likely due to genetic bottlenecks during establishment of strains and random genetic drift during domestication, rather than resulting from intense artificial selection imposed by selective breeding (Hyten, et al. 2006).

Population structure analyses based on both principal component and admixture analyses consistently showed that the domesticated fish have significantly diverged from their indirect wild ancestors, after several hundred years of selective breeding (**fig. 1C** and supplementary **fig. S7**). Principal component analysis in the different breeding lines of domesticated fish showed that, except for the elephant ear phenotypes, there was no clear differentiation among the studied traits (**fig. 1D)**. In admixture analysis, the most likely number of genetic clusters for hypothesized ancestral groups within domesticated fish was inferred to be three (**fig. 1E)**. We observed that fish exhibiting the same particular trait, e.g., the halfmoon tail, the double tail and the Plakat fin shape, are not always assigned to the same genetic clusters (**fig. 1F**). The results imply that these traits are more independent of their genetic background than those determined by a number of minor-effect loci, as in some domesticated animal breeds (Do, et al. 2013; Al-Mamun, et al. 2015; Wei, et al. 2015) and is likely determined by a single or a few loci with major effects.

Genetic control of diverse pigment patterns

The fighting fish is famous for its diversity of striking pigment patterns generated by artificial selection (supplementary **fig. S1**). First, we examined if pigment related traits are monogenic or polygenic in several test crosses focussing first on red pigments. We noted that xanthophore density differed markedly between body segments **(fig. 2A** and supplementary **fig. S8)**. QTL mapping revealed a major locus at LG6 for red pigment distribution in the caudal fin, with 20.6 % of its phenotypic variation explained (PVE) by this QTL. Aside from this major QTL, three additional QTLs with significant but smaller effects were identified at LG2, LG8 and

LG10, with PVE of 6.0 %, 5.5 % and 6.5 %, respectively **(fig. 2B)**. For pigmentation in the head, we identified one significant and two suggestive QTLs at LG4, LG11 and LG13, with PVE of 10.6 %, 5.9 % and 6.8 %, respectively **(fig. 2B)**. These data show that the distribution of red pigments is a polygenic trait. Interestingly, the QTLs found for tail had no overlap with those for head, implying the distribution of red pigments in different body sections is determined by different genetic loci. Although we did not identify the genes underlying red pigment distribution, our study provides first insights for a better understanding of various pigmentation patterns from a polygenic perspective.

The second pigmentation trait we investigated was dorsal fin spotting **(fig. 2C)**. We phenotyped 156 fish from the F_2 family, RM2, and found that this conforms to a pattern of Mendelian inheritance **(**supplementary **fig. S9a)**. Genome-wide association mapping based on ~ 25K SNPs revealed only one major locus on LG 11 responsible for this trait **(fig. 2C** and supplementary **fig. S9b)**. All fish with dorsal fin spots were homozygous at the most differentiating SNPs, suggesting that this trait is recessive. Using this dataset, the locus was restricted to a region of ~ 800 kb harbouring ~ 100 genes **(**supplementary **fig. S9c)**. This genomic region will be the focus for further investigation.

Finally, we studied the albino phenotype, which is characterized by a total lack of black pigments in the fins and body, except for the eyes, regardless of presence of the other colours. This recessive trait follows a monogenic Mendelian inheritance pattern **(**supplementary **fig. S10a)**. We mapped this trait to a locus on LG4 by using RAD-tag markers on our test crosses **(**supplementary **fig. S10b & c)**. Recombination analysis based on 293 fish revealed that this locus spans a genomic region of ~ 438 kb **(**supplementary **fig. S10d)**, with 18 predicted genes **(**supplementary **S11a)**. Because the albino fish lacked melanin expression in the skin **(fig. 2D & E)**, we compared the expression pattern of these genes in albino and wild-type pigmented fish, and found that only *microphthalmia-associated transcription factor a* (*mitfa*) within this

region, was differentially expressed **(**supplementary **fig. S11b)**. However, the expression of *mitfa* was also decreased in the eye of the albino fish, which typically shows black pigmentation **(**supplementary **fig. S11d)**. Studying the expression of the *mitfa* gene, we found a paralog of *mitfa,* in the eye. The expression of *mitfb* was higher than that of *mitfa* in the eye of both albino and wild-type fish, and, interestingly, *mitfb* was more highly expressed in eyes of albino than wild-type fish **(**supplementary **fig. S11d)**. It is likely that *mitfb* has critical functions for retinal pigment formation and shows compensatory effects on *mitfa* in albino fish phenotype. Interestingly, this mechanism is consistent with the nacre mutant of zebrafish, which is also a mutation of *mitfa* and also has black eyes (Lister, et al. 1999; Lister, et al. 2001).

To verify whether *mitfa* is associated with the albino phenotype, we knocked out this gene using CRISPR/Cas9 system by targeting the coding sequences **(**supplementary **fig. S12)**. Two G0 CRISPants without detectable wild-type alleles completely lost melanin pigmentation, while the mosaic fish with both wild-type and mutated haplotypes were markedly reduced in the density of melanin containing cells, compared to wild-type controls, which consistently presented normal melanin pigmentation 48 hpf **(fig. 2F** and supplementary **fig. S13)**. This CRISPRant phenotype matches the phenotype of the albino mutant, suggesting that *mitfa* is the altered gene in this fighting fish mutant. However, we did not find any mutation in introns and exons of *mitfa*, implying that the mutant phenotype is associated with variation in cis-regulatory element acting on this gene. Comparison between homozygous albino and wild-type pigmented fish revealed a cluster of indels and SNPs about 25 kb upstream of *mitfa*, including a 366-bp deletion in the albino mutant. Genotyping this deletion in ~ 1,000 fish revealed that this deletion was strictly correlated with the albino phenotype **(**supplementary **fig. S14 and table S9)**. These data suggest that the 366-bp deletion is a distant cis-regulatory element and could underlie the albino phenotype.

Taken together, in the fighting fish, a single or very few major loci can bring about phenotypic innovations for some colours. Thus, these traits are more easily affected by selection than polygenic traits. Certainly, many of the other colour strains in the fighting fish are likely to be determined by major effect loci. Further studies on these traits will provide valuable information to understand the mechanism of how selection affects phenotypic innovations.

Genetic basis of elephant ear and double-tail varieties

Another striking feature of domesticated fighting fish is the overgrowth of almost all types of fins. Using 47 sequenced fish including nine "elephant ear" phenotypes (supplementary **table S8 and fig. S1f**), we firstly mapped the locus for the elephant ear mutation (**fig. 3A**), a recessive trait following Mendelian inheritance characterized by elongated pectoral fins (Lucas 1968). Using F_{ST} scans based on whole-genome resequencing data, this locus locates to a 1.3-Mb region on LG9 (**fig. 3B, C & D**). We further refined the haplotypes between elephant ear and wild-type fish and annotated 55 protein coding genes in this region, of which six are known to play important roles in fin development and regeneration (**fig. 3E**). Examining their expression patterns in pectoral fins at one month post hatching when the elephant ear mutation becomes fully apparent, we identified three interesting candidate genes: potassium voltage-gated channel subfamily H member 8 (*kcnh8*), homeobox even-skipped homolog protein 1 (*evx1*) and collagen alpha-1(XVI) chain (*col16a1*) that were significantly down-regulated in elephant ear phenotypes when compared to wild-type fish (**fig. 3F**), an observation that agrees with the recessive inheritance pattern of this trait (Lucas 1968). A previous study suggested that *evx1* is required for joint formation in zebrafish fin dermoskeleton, but, apparently has no role in fin length (Schulte, et al. 2011). Though important in fin regeneration and typically affected by domestication processes (Anastasiadi and Piferrer 2019), there was no obvious evidence that collagen genes are responsible for overgrowth of fins (Durán, et al. 2011; Anastasiadi and

Piferrer 2019). In particular, we found one paralog of *kcnh8* at LG14, implying the functions of these two paralogs might have diverged with one fulfilling the general neural function and the other one regulating fin growth, a situation resembling what has been observed regarding the expression patterns of potassium channels in both zebrafish and goldfish long fin mutants (Perathoner, et al. 2014; Lanni, et al. 2019; Stewart, et al. 2019; Kon, et al. 2020). Interestingly, in one swordtail species, *Xiphophorus hellerii*, differential expression of *kcnh8* was associated with development of a male ornamental trait, a ventral outgrowth of the caudal fin, called sword (Schartl, et al. 2020). In zebrafish and goldfish longfin mutants, mutations in paralogous potassium channel genes *kcnh2a, kcnk5b* and *kcc4a*, cause overgrowth of different types of fins (Perathoner, et al. 2014; Lanni, et al. 2019; Stewart, et al. 2019; Kon, et al. 2020). Mutations disrupting ion channels and ion-dependent signaling often are related to abnormal organ development and regeneration via bioelectrical regulation (McLaughlin and Levin 2018). As discussed above, expression alteration and subfunctionalization of *kcnh8* encoding a potassium voltage-gated channel is more likely related to the formation of the elongated pectoral fins (elephant ear) breed. We found a fixed missense mutation in the last exon (2912A/G, exon16) of *kcnh8*, but not in the other candidate genes. This amino acid change (H/R) is neither evolutionary conserved across teleosts **(**supplementary **fig. S15)**, nor predicted to likely affect protein function, with a score of 1.00 as estimated by SIFT (Sim, et al. 2012). However, a single SNP in coding sequences might be less effective altering gene expression (Cowper-Sal, et al. 2012). Except for this SNP, there are still some SNPs and short sequence variations in the non-coding sequences within and closely flanking these candidate genes that may affect expression. Therefore, the elephant ear breed is more likely caused by mutations that affect expression. In addition, it is also worth mentioning that the *FKBP prolyl isomerase 14* (*fkbp14*), encoding a chaperone and calcium binding protein, shows a similar expression pattern with *kcnh8* and the statistical significance for differential expression is only slightly

over 0.05 (fold change, 2.04 and *P* value for *t*-test, 0.07). In zebrafish, inhibition of *fkbp14* function was shown to cause outgrowth of the caudal fin margin (Kujawski, et al. 2014). In swordtails, the expression pattern of a paralogous gene *fkbp9* was also observed to be associated with the development of sword of the tail fin in males (Schartl, et al. 2020). These data imply that *fkbp14* is another potential candidate gene for "elephant ear" phenotypes. Taken together, our results suggest that a variety of potassium channel and/or calcium binding genes play critical roles to generate favored ornamental phenotypes of overgrowth of various fin types that are observed in artificial selective breeding in the fighting fish (Stewart, et al. 2019; Kon, et al. 2020; Schartl, et al. 2020).

Double tail is one of the most well-known and most appreciated among various fin varieties of fighting fish. This mutant presents a unique ventralized pattern of dorsal trunk and tail, and features a doubling of the number of fin rays for both dorsal and caudal fins **(fig. 4A & B** and supplementary **fig. S16)**. Double tail fighting fish was found to be a recessive homozygote (*st*) and we mapped the locus responsible for double tail to a ~ 130-kb region on LG1 (*st* vs *ST*) by RAD sequencing and fine mapping by examination of recombinants in 502 fish **(**supplementary **fig. S17)**. Sequence analysis revealed that this locus harbours three genes: zinc finger transcription factors Zic1 and Zic4 (*zic1*and *zic4*) and phospholipid scramblase 1 (*plscr1*), and overlaps with the *Da* locus of medaka for a double-tail mutant that contains only *zic1* and *zic4* (Moriyama, et al. 2012). Consistent with medaka, expressions of both *zic1* and *zic4* were suppressed in double tail **(**supplementary **fig. S18)** and no mutation was identified in the coding sequences (Kawanishi, et al. 2013). We further sequenced the genomes of both homozygous single- and double-tail fish and found in double tail no large sequence variation except for a ~ 180-bp deletion ~ 15-kb downstream of *zic4* **(fig. 4C** and supplementary **fig. S19a)**. This deletion was located in a cluster of CNEs and coincided with predicted CNE.006008 **(**supplementary **fig. S19b)**. Genotyping at this locus showed that the deletion was completely

correlated with phenotypes in >1000 examined fish **(fig.** 4**D** and supplementary **table S10)**. In medaka both genes, *zic1* and *zic4* were verified to be responsible for double tail (Moriyama, et al. 2012). However, the mechanism by which these genes induce this phenotype is still unclear. It was assumed that a transposon, *Albatross* (> 41 kb), inserted into the common regulatory region of both *zic1* and *zic4*, ultimately leads to the double-tail mutation of medaka (Moriyama, et al. 2012). Therefore, we hypothesized that the deletion in CNE.006008, an enhancer, is responsible for the double tail phenotype in fighting fish. To test this hypothesis, first we inserted the CNE.006008 locus and its closely flanking sequences of ~ 100 bp separately from single- and double-tail fish into ZED vectors (Bessa, et al. 2009) and injected them into one-cell stage embryos. We observed that the wild-type *ST* allele significantly enhanced GFP expression in embryos at 24 hpf, when both *zic1* and *zic4* show differential expression between double-tail and wild-type fish (Moriyama, et al. 2012), while no visible GFP expression was detected for the *st* allele **(fig. 4E** and supplementary **table S11)**. The efficiency of the two alleles as candidate enhancers was further examined using a Dual-Luciferase Reporter Assay which showed that the *ST* allele enhanced luciferase expression by ~10 × relative to *st* allele in Singapore grouper embryonic cell line **(fig. 4F)**.

Finally, we deleted this enhancer using the CRISPR-Cas9 system in fighting fish. Considering the efficiency of tested gRNAs and the cluster of CNEs that could have unpredicted functions, we limited the modification to the CNE.006008 region and did not involve the other CNEs **(**supplementary **fig. S20)**. Genetic analysis revealed that none of these fish had completely deleted CNE.006008, suggesting non-simultaneous cutting at multiple targeted gRNA positions. These mosaic fish (n = 7) had significantly more fin rays than the non-injected controls **(***P* < 0.01; **fig. 4G).** We screened one modified fish, where > 80% of sequenced clones were mutants with deleted sequences up to 56% of *st* allele **(**supplementary **fig. S21)**. Although this fish was not a pure knockout, we observed that the number of fin rays

of both dorsal and caudal fins was significantly higher than in the single tail and approaching that of the double tail **(fig. 4H)**. Taken together, deletion of the candidate co-enhancer of *zic1* and *zic4*, CNE.006008 was found to be the causative mutation for double-tail fighting fish.

To date, double-tail mutants are only reported in the fighting fish and medaka, and both are caused by mutations in the co-regulatory regions of *zic1* and *zic4* (Moriyama, et al. 2012). As shown in medaka loss of function of either gene is not able to cause a double-tail phenotype (Moriyama, et al. 2012). Furthermore, loss of functions of both *zic1* and *zic4* causes fatal Dandy-Walker malformation like disease in animals (Grinberg, et al. 2004; Blank, et al. 2011). Thus, selection on the co-regulatory regions of multiple effector genes in a single locus becomes more efficient to bring about such phenotypic innovations. The occurrences of those kinds of mutations are scarcer than those determined by a single gene, which likely explains why only two double-tail cases in a number of domesticated teleosts have been observed so far. In comparison, traits that are determined by single genes are much more common. Microphthalmia-associated transcription factors have been extensively reported responsible for the albinism of a number of domesticated animals, such as dogs (Karlsson, et al. 2007), pigs (Chen, et al. 2016), ducks (Zhou, et al. 2018) and quails (Minvielle, et al. 2010). Interestingly, most of them are caused by mutations in the regulatory sequences (Karlsson, et al. 2007; Hauswirth, et al. 2012; Chen, et al. 2016; Zhou, et al. 2018; Hofstetter, et al. 2019). Mutations in coding sequences are more likely to alter protein functions. In particular for pleiotropic factors, such mutations will be more harmful to the organisms than those occurring in regulatory regions, which only affect the expression level (Wittkopp and Kalay 2012; Petit, et al. 2017). In this regard, mutations in the coding sequences of microphthalmia-associated transcription factors have been extensively reported to lead to various defects in animals (Tassabehji, et al. 1994; Levy, et al. 2006). This type of mutation is more prone to be eliminated by artificial selection. Above all, mutations in cis-regulatory elements provide valuable raw

materials for selection during domestication and play critically important roles in phenotypic variation.

Conclusion

In this study, we sequenced the genomes of several fighting fish breeds and studied the genomic basis of most striking colour and fin shape variants in this species. We found that phenotypes including some colorations and fin shapes were determined by major-effect loci, indicating that major loci can bring about phenotypic innovations rapidly. Using CRISPR/Cas-9 induced-mutations, we verified that both double-tail and albino phenotypes resulted from mutations in a regulatory element near *zic1*/*zic4* and mutations in coding regions of *mitfa*, respectively. Our findings suggest that cis-regulatory elements play critically important roles in generating phenotypic variation during domestication by artificial selection as well as natural selection at evolutionary time scales. There are still other breeds varying in fin shapes and sizes as well as in pigment patterns and in aggression worthy of further investigation by the CRISPR/Cas9 methods we developed here. This will facilitate that this species will become a new model system since it is amenable for dissecting the genetic architecture underlying morphological and behavioural evolutionary innovations.

Materials and Methods

Genome sequencing and assembly

Genomic DNA from one highly inbred yellow single-tail female and one transparent double-tail male were used for construction of both short-insert (~270 bp, 350 bp and 550 bp) and jumping libraries (3 kb, 5 kb, 10 kb, 15 kb and 20 kb) (supplementary **table S1 and fig. S2**).

In addition, one female (albino and double-tail) and one male (also homozygous for melanin pigmentation and single-tail) were sequenced with 550 bp insert libraries. Genomic DNA was isolated with MagAttract HMW DNA Kit (Qiagen). Heterozygosity vs homozygosity of these fish was assessed at 10 microsatellite loci development in a previous study (Chailertrit, et al. 2014). The yellow female showed a low genetic heterozygosity of 0.1. Genotypes of these fish were determined by both test crosses and genetic markers developed as described below. All sequencing was carried out on Illumina Nextseq 500. Raw reads were cleaned using the program *process_shortreads* (-r -c -q -t 150) in the Stacks software package (Catchen, et al. 2013). Genomes of the highly inbred yellow single-tail female and the transparent double-tail male of varying libraries were assembled using ALLPATHS-LG (Gnerre, et al. 2011) with default parameters, while the other two unique libraries were assembled using ABYSS2.0 (Jackman, et al. 2017) with default parameters. Gaps were filled with paired reads using GapFiller (Nadalin, et al. 2012). The genome size of the fighting fish was estimated based on k-mer frequencies, *de novo* RADtags mapping and Q-PCR method (supplementary **note 1**). Completeness of the genome assembly was evaluated using BUSCO (Simão, et al. 2015) and by the mapping rate of transcripts and de novo RADtags.

RNA sequencing and analysis

Three mRNA libraries were separately constructed for one male and two females of three months' age. Total RNA was isolated from brain, eye, skin, gill, muscle, intestine, spleen, liver, heart, kidney and gonad, and then equal amounts from each tissue were pooled for library construction using Illumina TruSeq RNA sample preparation kit (Illumina). Moreover, RNA samples of another mature male and female, derived from the above pooling strategy, were used for total RNA library construction with rRNA depletion, using NEBNext® Ultra™ RNA Library Prep Kit (NEB). Raw sequences were cleaned with *process_shortreads* (-r -c -q -t 150)

in Stacks package (Catchen, et al. 2013). Transcripts of individuals were assembled using Trinity (Haas, et al. 2013) with default parameters and then used for genome annotation.

RADseq and SNP genotyping

Samples from both mapping families and cultured strains with specific traits (see below), were genotyped using RADseq (Baird, et al. 2008) with some modifications as described in our previous study (Bai, et al. 2018). High-quality genomic DNA of 500 ng was digested with restriction enzyme *PstI-HF* (NEB) and ligated to barcoded adaptors with T4 DNA Ligase (NEB). DNA was then sheared with a peak of 500 bp for library construction. All libraries were sent to NextSeq500 (Illumina) for 150-bp single end sequencing. Parental and offspring samples were sequenced with an average of 16.9 M and 6.1 M reads, respectively, for accurate SNP calling (supplementary **table S1**). Raw reads were filtered using *process_radtags* (-r -c -q -t 150) in Stacks package (Catchen, et al. 2013). BWA-MEM (Bessa, et al. 2009) was used for reference-based mapping with default parameters and only reads with unique targets were retained. SNPs were discovered and genotyped using Stacks package (Catchen, et al. 2013) with parameters as described in our previous study (Bai, et al. 2018).

Linkage mapping and chromosomal-level genome assembly

Two F_2 families: BM1 and RM2 were used for linkage mapping. These two F_2 families: BM1 (92 fish) and RM2 (274 fish) were generated with two pairs of F_1 parents (i.e. BM1female×BM1male and RM2female×RM2male), respectively, which were the offspring of P parents: DtY2female and F0B1male (see details about their phenotypes in the subsection "QTL mapping and genome-wide association studies for traits of interest"). Genotyping of the two F_2 mapping families was conducted using RAD sequencing as described above. SNPs were firstly filtered for Mendelian segregation distortion using $\chi 2$ tests ($P < 0.05$). The cut-off value of missing genotypes across families was < 15%, which left 80 and 213 samples for BM1 and

RM2, respectively, for linkage mapping. Linkage group assignment and marker ordering were carried out using Lep-MAP3 (Rastas 2017) with LOD cut-off of 10. Both sex-averaged and sex-specific maps were constructed. The constructed linkage maps were used to build a chromosome-level genome assembly. RAD sequences of mapped markers were aligned to scaffolds to examine the occurrence of chimeric assemblies using ALLMAPS (Tang, et al. 2015), as linkage maps are not likely to generate among-chromosome grouping errors (Small, et al. 2016). If there were more than three markers from the same scaffolds mapped to different linkage groups, the scaffolds were split at the longest gaps between mismatched fragments. The new scaffolds were then anchored onto genetic maps to generate chromosome-level assemblies using ALLMAPS (Tang, et al. 2015) with default parameters.

Genome annotation

Annotation of the highly inbreed female fighting fish genome was conducted using MAKER (Cantarel, et al. 2008). The sequences were softmasked using RepeatMasker (Chen 2004) based on the repeat libraries obtained from RepeatModeler (http://www.repeatmasker.org), Repbase (Jurka, et al. 2005) and MAKER (Cantarel, et al. 2008) sequence repeat databases. Both evidence-based and *ab initio* gene models were used for annotation. Transcriptomes of fighting fish and protein sequences of zebrafish, medaka, stickleback, fugu and Nile tilapia from Ensembl database (release 86) were used for evidence. SNAP (Korf 2004) and Augustus (Stanke and Waack 2003) were iteratively used for *ab initio* gene models training. Predicted protein sequences were annotated by blast to nr and RefSeq databases (Pruitt, et al. 2005) with BLASTP (E-value < 1E-10).

Prediction of conserved non-coding elements (CNEs)

Identification of CNEs was according to a previous method (Brawand, et al. 2014). In brief, the fighting fish genome was used as reference for pairwise whole-genome alignment with

zebrafish (*D. rerio*), medaka (*O. latipes*), stickleback (*G. aculeatus*), fugu (*T. rubripes*) and Nile tilapia (*O. niloticus*) (downloaded from Ensembl database, release 86) using LASTZ (Harris 2007). Multiple alignments were generated with MULTIZ (Blanchette, et al. 2004) using the tree topology among the six species based on the phylogenetic study. Conserved sequences at least in one pair of alignments were predicted using PhastCons (Siepel, et al. 2005) under both conserved and non-conserved models (coverage = 0.3 and length = 45 bp). The predicted CNEs were then filtered by comparison to the coding sequences, non-coding RNAs, pseudogenes and transposable elements of the six studied species and also the transcripts of fighting fish (E-value < $1E^{-10}$). Only the elements of > 30 bp and with repetitive content < 50% were retained. For studies on CNEs related to candidate genes, we manually aligned the genomic loci with more reference fish including non-model species to refine the CNEs and identify candidate regulatory elements that might be lineage- or species-specific, using the above standards.

Whole-genome resequencing and genetic diversity analysis

Six wild (four from Thailand and two from Cambodia, respectively) and 28 randomly selected domesticated fish were sequenced with 500 bp insert libraries (supplementary **table S8**). Nine 'elephant ear' mutants, i.e. with phenotype of overgrowth of pectoral fin, were further sequenced to identify the genetic locus (supplementary **table S8**). Raw sequencing reads were filtered using the above method. Sequence mapping and variant calling were carried out using BWA-mem (Bessa, et al. 2009) and Picard/GATK v4.0 best practices workflows (DePristo, et al. 2011). SNPs were filtered with the following parameters: 'QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0', and indels with 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0 || SOR > 10.0'. We further filtered the variants with 'minDP 7, --max-missing 0.925' using VCFtools (Danecek, et al. 2011). A total of 6,735,573 genotypes were obtained for further analysis. Genetic diversity was estimated with

447 a 100-kb window size using VCFtools (Danecek, et al. 2011). Unbiased estimation of
448 nucleotide diversity taking into consideration both segregating and non-segregating sites were
449 computed using the program pixy (Korunes and Samuk 2021). Population structure was firstly
450 analysed with PCA using Plink2.0 (Purcell, et al. 2007). The program Admixture (Alexander,
451 et al. 2009) was then used to infer the genetic clusters at individual level.

452 ***QTL mapping and genome-wide association studies for traits of interest***

453 Three pigmentation traits, including distribution of red pigments in different body
454 compartments (Fig. 2A), dorsal fin spotting (fig. 2C) and albino phenotype (fig. 2D), and two
455 fin morphology traits (i.e. elephant ear and double tail, fig. 3A and fig. 4A) were studied. Before
456 setting up mapping families by crossing parents with different phenotypes, test crosses were
457 generated to examine phenotypic segregation of double tail, albino and dorsal fin spotting.
458 Parents that were homozygous for the two studied traits (i.e. double-tail vs. single-tail and
459 melanin vs. albino), were selected as P (parental) generation (DtY2female and F0B1male) to
460 set up mapping families (supplementary **fig. S2**). Two F_2 families: BM1 and RM2 were
461 produced by crossing F_1 parents (BM1female and BM1male) and (RM2female and RM2male),
462 respectively, which were the offspring of P generation: DtY2female and F0B1male. The two
463 F_2 mapping families were genotyped using RAD sequencing as described above.

464 In the two F_2 families (i.e. BM1 containing 92 fish, and RM2 containing 274 fish), four traits:
465 distribution of red pigment, dorsal fin spotting, albino and double tail, were recorded for each
466 individual. In detail, F0B1male was a homozygous single-tail and melanin pigmented fish
467 (wild type at both loci), while DtY2female was a homozygous double-tail and albino (loss of
468 melanin) fish. Both traits (i.e. double-tail and albino) showed a recessive Mendelian inheritance
469 pattern with a segregation ratio of 3:1 in F_2 populations. Genome-wide association studies
470 (GWAS) were separately conducted based on the two F_2 families to map the two traits. To

narrow down the genomic regions responsible for the two traits, 136 additional fish of confirmed phenotypes, from China, Thailand, Malaysia, Singapore, and Indonesia were genotyped using both RAD sequencing and SNP/indel markers developed from resequencing data, as described above. Genotypes were used to identify recombinants in the major loci determining double-tail and albino. For dorsal fin spotting, parents, DtY2female and F0B1male showed spotted and non-spotted dorsal fin, respectively, while F_1 parents, RM2female and RM2male showed spotted and non-spotted dorsal fin, respectively. Due to a phenotypic interaction of iridescent pigmentation patterns and the albino condition, which also segregated in the F_2 mapping crosses, only 156 individuals from the RM2 family could be phenotyped. This trait presented a segregation ratio of 1:1 for spotted vs non-spotted dorsal fin. We carried out both quantitative trait loci (QTL) mapping and GWAS for this trait using these 156 phenotyped fish. As the results in QTL mapping and GWAS were consistent, we only present the results of the GWAS. Finally, for distribution of red pigment, we observed red pigments in DtY2female, but not in F0B1male, while all F_1 fish presented red pigments. In both F_2 families, we observed the distribution pattern of red pigments varied evidently not only across different body compartments of each individual, but also among individuals throughout the whole family. We developed a method to quantify and record red pigments in different body areas (supplementary **note 2**). QTL mapping was conducted in the large RM2 family with 211 phenotyped offspring, rather than BM1 with only 77 phenotyped siblings, to map and estimate the effects of the loci.

GWAS was performed with compressed mixed linear model implemented in the GAPIT R package, with the kinship relatedness matrix and sex as covariates (Lipka, et al. 2012). This mixed model incorporates and estimates component variance of both kinship relatedness matrix and sex, using VanRaden algorithm. *P* value was calculated for each marker and the statistical significance threshold was determined at 0.05 level with Bonferroni corrections ($P = 0.05/N$,

where N is the number of total markers used for association test). QTL mapping was conducted using the Haley-Knott regression method (Haley and Knott 1992) implemented in the R package qtl (Broman, et al. 2003). The linkage map of RM2 family was used for QTL mapping, with interval mapping (IM) algorithm. LOD thresholds for both chromosome- and genome-wide significance were estimated with permutation tests for 1,000 times. To screen the 'elephant ear' locus, a genome-wide F_{ST} scan was performed between elephant ear mutants and the remaining resequenced samples, 9 and 28 fish respectively, using 30-kb window size with a step of 15kb. Window-size F_{ST} values were then Z transformed ($ZF_{ST} = (F_{ST} - \mu F_{ST})/\sigma F_{ST}$) to compare among chromosomes. Variants within and flanking this locus were retrieved and analysed to refine the haplotypes. Fixed or nearly fixed variants were annotated and protein coding genes within this locus were individually analysed by literature mining. Genes associated with fin development and regeneration were kept for expression analysis using real-time RT-PCR to identify candidate genes.

Developing and validating trait-associated DNA markers

In order to quickly differentiate among genotypes, we developed indel markers for fast PCR assays for double tail and albino traits. The shortest genomic region resulting from association mapping was used for marker screening. Homozygous individuals with regard to the above traits were resequenced for marker discovery using GATK pipeline (DePristo, et al. 2011) according to our previous study (Wang, et al. 2015). Indels were firstly validated by manual alignment of genome sequences between homozygous mutant and wild type. Primers of these indels of suitable length were then designed for PCR assays. The associations between phenotypes and these discovered markers were further tested in domesticated fish from different strains (from ~ 500 to ~ 1000 individuals for different traits) to examine recombination between markers for fine mapping.

Gene expression analysis using RT-PCR and qRT-PCR

Gene expression was studied by reverse transcription PCR (RT-PCR). Total RNA from independent tissues or embryos was isolated using TRIzol reagent (Invitrogen). Two micrograms of total RNA were treated with DNase I (Roche) and then used for cDNA synthesis using Reverse Transcriptase M-MLV (Promega). Expression of genes of interest in different tissues and in different developmental stages was firstly examined using RT-PCR with gene specific primers. The relative expression of candidate genes was then studied using real-time RT-PCR (qRT-PCR) using KAPA™ SYBR® FAST qPCR Kits (KapaBiosystems) with CFX96 Touch™ Real-Time PCR Detection System (Bio-Rad). Three replicates were performed for each sample and cDNA from 50 ng of total RNA was used for each reaction. *Beta actin* or *EF1A* was used as endogenous reference according to their expression stability. The $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen 2001) was used to quantify relative gene expression.

Luciferase reporter assay

Candidate enhancer regions with closely flanking sequences from different alleles were cloned and constructed into the enhancer region of pGL3-Promoter vector that contains a basal SV40 promoter sequence (Promega). The reporter gene constructs together with pRL Renilla Luciferase Control Reporter Vector (Promega) were co-transfected into a Singapore grouper embryonic cell line (Chew-Lim, et al. 1994) using TurboFect or Lipofectamine 3000 (Thermo Fisher). Luciferase activity was measured at 48 h post transfection using Dual-Luciferase Reporter Assay System (Promega). Three independent transfections were carried out in 6-well plates with each measured in triplicates.

Enhancer reporter assay

Candidate enhancer sequences of different alleles with closely flanking sequences were constructed into Zebrafish Enhancer Detection (ZED) Vector (Bessa, et al. 2009) using Gateway Recombination Cloning Technology (Thermo Fisher Scientific). T7-Transposase (Khattak, et al. 2014) (Addgene) was transcribed using mMESSAGE mMACHINE T7 kit (Life Technologies), according to the manufacturer's instructions. A final concentration of 40 ng/µL ZED constructs, 50 ng/µL transposase mRNA and 0.05% phenol red were co-injected into one-cell stage embryos. The embryos were imaged for GFP and internal control RFP expression at different time points using a Leica MZFLIII microscope. The elements were considered as candidate enhancers if there were more than 20% of injected embryos showing consistent expression pattern of GFP at the presence of RFP (Bessa, et al. 2009; Sharma, et al. 2015).

Knockout using CRISPR/Cas9

CRISPR/Cas9 was used to introduce mutations into the genes or elements of interest. Guide RNA (gRNA) was designed using E-CRISP (Heigwer, et al. 2014). gRNA sequences were blasted against the reference genome to avoid off-targets. Template of gRNA was assembled using PCR according to a previous method (Vejnar, et al. 2016). In brief, gRNA was designed with common flanking adaptors as follows: 5'-TAATACGACTCACTATA[GGN(18)]GTTTTAGAGCTAGAA-3'. A universal primer was used to assemble gRNA temple with direct PCR, with the following sequences: 5'-AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTT AACTTGCTATTTCTAGCTCTAAAAC-3'. gRNA was transcribed using HiScribe T7 High Yield RNA Synthesis Kit (NEB) with 150 ng of purified DNA template and gRNA was subsequently purified using miRNeasy Mini Kit (Qiagen), according to the manufacturer's instructions. Cas9 Nuclease NLS (NEB) and gRNA with a final concentration of 100 ng/µL and 200 ng/µL, respectively, were co-injected into one-cell stage embryos. Both phenotypes and genotypes were screened for candidate mutants. DNA fragments spanning the targeted

sequences of gRNAs were amplified using fragment-specific primers. PCR products were purified using QIAquick PCR Purification Kit (Qiagen) for mutant screening using T7 endonuclease assay (NEB). PCR products that showed cleavage in T7 endonuclease assay were then validated by TA cloning and Sanger sequencing. We developed a whole protocol for transgenic and CRISPR knockout technology for the fighting fish, a species of particular mating and brooding behaviors (supplementary **note 3**).

Ethics declarations

All procedures for handling of fish were according to the instructions of the Institutional Animal Care and Use Committee (IACUC) of Temasek Life Sciences Laboratory, Singapore (Approval no. TLL (F)-16-003).

Availability of data and materials

The genome sequences and related annotations of fighting fish are hosted by the web server of Temasek Lifesciences Laboratory and will be released upon acceptation of this manuscript. Sequences used for whole genome-sequencing, RNA sequencing and RAD sequencing will be available upon publication with the DDBJ Sequencing Read Archive (SRA) through BioProject ID PRJDB7253- PRJDB7255.

Acknowledgements

We thank Betta hobbyists from both Betta Club Singapore and Wild Betta Club Singapore for providing samples and photographs. We acknowledge the fish facility, biocomputing facility and IT department of Temasek Lifesciences Laboratory for help in fish culture, high performance computing and data processing. We are grateful to Dr. José Bessa and Dr. Ying Yang for providing ZED and luciferase vectors, respectively. Our grateful thanks are also

589 extended to Dr. Jinlu Wu for providing the Singapore grouper embryonic cell lines. This project
590 was supported by the Temasek Lifesciences Laboratory, Singapore.

591 **Author contributions**

592 GHY initiated the project “Genetics and genomics of the fighting fish”. GHY, LW, MS and
593 AM conceived the study. GHY, FP, MS and AM supervised the whole study. LW and FS
594 performed genome sequencing. LW, ZYW and FS performed RNA sequencing. LW, FS, ZYW,
595 BY and BB performed RAD sequencing. FS, LW, HL, YW and HP set up mapping families
596 and cultured fish. LW assembled and annotated genomes and transcriptomes. LW, BY and
597 ZYW analysed sequences. LW and BY constructed linkage maps. FS, LW, YW and BY
598 conducted fine mapping of QTL. LW, ZY, FS and ZYW constructed vectors. LW, FS, ZYW
599 and YW performed knockout and injections. FS, LW and ZYW examined gene expression. FS
600 and YW contributed to genotyping and sequencing of phenotypes. ZM, BF, YA, YS and ML
601 collection and phenotyping of samples. LW, BY, FS and GHY wrote the paper with inputs
602 from the other authors. GHY, LW, FP, MS and AM interpreted the findings in biological
603 context and commented on the manuscript. All authors discussed the results and approved the
604 final version of the paper.

605 **Additional information**

606 Supplementary Information

607 **Competing financial interests:**

608 The authors declare no competing financial interests.

609 **References**

610 Al-Mamun HA, Kwan P, Clark SA, Ferdosi MH, Tellam R, Gondro C. 2015. Genome-wide
611 association study of body weight in Australian Merino sheep reveals an orthologous
612 region on OAR6 to human and bovine genomic regions affecting height and weight.
613 *Genetics Selection Evolution* 47:66.
614 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in
615 unrelated individuals. *Genome Research* 19:1655-1664.
616 Anastasiadi D, Piferrer F. 2019. Epimutations in developmental genes underlie the onset of
617 domestication in farmed European sea bass. *Molecular Biology and Evolution* 36:2252-
618 2264.
619 Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng
620 HH, Clarke L, Couldrey C. 2015. Coordinated international action to accelerate genome-
621 to-phenome with FAANG, the Functional Annotation of Animal Genomes project.
622 *Genome Biology* 16:1-6.
623 Bai B, Wang L, Zhang YJ, Lee M, Rahmadsyah R, Alfiko Y, Ye BQ, Purwantomo S, Suwanto
624 A, Chua N-H. 2018. Developing genome-wide SNPs and constructing an ultrahigh-
625 density linkage map in oil palm. *Scientific Reports* 8:691.
626 Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA,
627 Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD
628 markers. *PLoS ONE* 3:e3376.
629 Bessa J, Tena JJ, de la Calle-Mustienes E, Fernández-Miñán A, Naranjo S, Fernández A,
630 Montoliu L, Akalin A, Lenhard B, Casares F. 2009. Zebrafish enhancer detection (ZED)
631 vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory
632 regions in zebrafish. *Developmental Dynamics* 238:2409-2417.

633 Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom
634 K, Clawson H, Green ED. 2004. Aligning multiple genomic sequences with the threaded
635 blockset aligner. *Genome Research* 14:708-715.
636 Blank MC, Grinberg I, Aryee E, Laliberte C, Chizhikov VV, Henkelman RM, Millen KJ. 2011.
637 Multiple developmental programs are altered by loss of Zic1 and Zic4 to cause Dandy-
638 Walker malformation cerebellar pathogenesis. *Development* 138:1207-1216.
639 Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW,
640 Bezault E. 2014. The genomic substrate for adaptive radiation in African cichlid fish.
641 *Nature* 513:375-381.
642 Broman KW, Wu H, Sen Ś, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses.
643 *Bioinformatics* 19:889-890.
644 Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M.
645 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model
646 organism genomes. *Genome Research* 18:188-196.
647 Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool
648 set for population genomics. *Molecular Ecology* 22:3124-3140.
649 Chailertrit V, Swatdipong A, Peyachoknagul S, Salaenoi J, Srikulnath K. 2014. Isolation and
650 characterization of novel microsatellite markers from Siamese fighting fish (*Betta*
651 *splendens*, Osphronemidae, Anabantoidei) and their transferability to related species, *B.*
652 *smaragdina* and *B. imbellis*. *Genetics and Molecular Research* 13:7157-7162.
653 Chen L, Guo W, Ren L, Yang M, Zhao Y, Guo Z, Yi H, Li M, Hu Y, Long X. 2016. A de novo
654 silencer causes elimination of MITF-M expression and profound hearing loss in pigs.
655 *BMC Biology* 14:1-15.
656 Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences.
657 *Current Protocols in Bioinformatics* 5:4.10. 11-14.10. 14.

Chew-Lim M, Ngoh G, Ng M, Lee J, Chew P, Li J, Chan Y, Howe J. 1994. Grouper cell line for propagating grouper viruses. *Singapore Journal of Primary Industries* 22:113-116.

Cowper-Sal R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, Moore JH, Lupien M. 2012. Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics* 44:1191-1198.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156-2158.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43:491-498.

Do DN, Strathe AB, Ostersen T, Jensen J, Mark T, Kadarmideen HN. 2013. Genome-wide association study reveals genetic architecture of eating behavior in pigs and its implications for humans obesity by comparative mapping. *PLoS ONE* 8:e71509.

Durán I, Marí-Beffa M, Santamaría J, Becerra J, Santos-Ruiz L. 2011. Actinotrichia collagens and their role in fin formation. *Developmental Biology* 354:160-172.

Fan G, Chan J, Ma K, Yang B, Zhang H, Yang X, Shi C, Chun-Hin Law H, Ren Z, Xu Q. 2018. Chromosome-level reference genome of the Siamese fighting fish Betta splendens, a model species for the study of aggression. *GigaScience* 7:giy087.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* 10:241-251.

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. *Proceedings of the National Academy of Sciences, USA* 108:1513-1518.

683 Grinberg I, Northrup H, Ardinger H, Prasad C, Dobyns WB, Millen KJ. 2004. Heterozygous
684 deletion of the linked genes ZIC1 and ZIC4 is involved in Dandy-Walker malformation.
685 *Nature Genetics* 36:1053-1055.
686 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles
687 D, Li B, Lieber M. 2013. De novo transcript sequence reconstruction from RNA-seq
688 using the Trinity platform for reference generation and analysis. *Nature Protocols*
689 8:1494-1512.
690 Haley CS, Knott SA. 1992. A simple regression method for mapping quantitative trait loci in
691 line crosses using flanking markers. *Heredity* 69:315-324.
692 Harris RS. 2007. Improved Pairwise Alignmnet of Genomic DNA: The Pennsylvania State
693 University, PA, USA.
694 Hauswirth R, Haase B, Blatter M, Brooks SA, Burger D, Drögemüller C, Gerber V, Henke D,
695 Janda J, Jude R. 2012. Mutations in MITF and PAX3 cause "splashed white" and other
696 white spotting phenotypes in horses. *PLoS Genetics* 8:e1002653.
697 Heigwer F, Kerr G, Boutros M. 2014. E-CRISP: fast CRISPR target site identification. *Nature*
698 *Methods* 11:122-123.
699 Hofstetter S, Seefried F, Häfliger IM, Jagannathan V, Leeb T, Drögemüller C. 2019. A non-
700 coding regulatory variant in the 5′-region of the MITF gene is associated with white-
701 spotted coat in Brown Swiss cattle. *Animal Genetics* 50:27-32.
702 Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan
703 PB. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of*
704 *the National Academy of Sciences, USA* 103:16666-16671.
705 Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H,
706 Coombe L, Warren RL. 2017. ABySS 2.0: resource-efficient assembly of large genomes
707 using a Bloom filter. *Genome Research* 27:768-777.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* 110:462-467.

Karlsson EK, Baranowska I, Wade CM, Hillbertz NHS, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas EJ. 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genetics* 39:1321-1328.

Kawanishi T, Kaneko T, Moriyama Y, Kinoshita M, Yokoi H, Suzuki T, Shimada A, Takeda H. 2013. Modular development of the teleost trunk along the dorsoventral axis and zic1/zic4 as selector genes in the dorsal module. *Development* 140:1486-1496.

Kemble H, Nghe P, Tenaillon O. 2019. Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. *Evolutionary applications* 12:1721-1742.

Khattak S, Murawala P, Andreas H, Kappert V, Schuez M, Sandoval-Guzmán T, Crawford K, Tanaka EM. 2014. Optimized axolotl (*Ambystoma mexicanum*) husbandry, breeding, metamorphosis, transgenesis and tamoxifen-mediated recombination. *Nature Protocols* 9:529-540.

Kon T, Omori Y, Fukuta K, Wada H, Watanabe M, Chen Z, Iwasaki M, Mishina T, Shin-ichiro SM, Yoshihara D. 2020. The genetic basis of morphological diversity in domesticated goldfish. *Current Biology* 30:2260-2274.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5:59.

Korunes KL, Samuk K. 2021. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*.

Kujawski S, Lin W, Kitte F, Börmel M, Fuchs S, Arulmozhivarman G, Vogt S, Theil D, Zhang Y, Antos CL. 2014. Calcineurin regulates coordinated outgrowth of zebrafish regenerating fins. *Developmental Cell* 28:573-587.

Lanni JS, Peal D, Ekstrom L, Chen H, Stanclift C, Bowen ME, Mercado A, Gamba G, Kahle KT, Harris MP. 2019. Integrated K+ channel and K+ Cl-cotransporter functions are required for the coordination of size and proportion during development. *Developmental Biology* 456:164-178.

Lehner B. 2013. Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics* 14:168-178.

Levy C, Khaled M, Fisher DE. 2006. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends in Molecular Medicine* 12:406-414.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28:2397-2399.

Lister JA, Close J, Raible DW. 2001. Duplicate mitf genes in zebrafish: complementary expression and conservation of melanogenic potential. *Developmental Biology* 237:333-344.

Lister JA, Robertson CP, Lepage T, Johnson SL, Raible DW. 1999. Nacre encodes a zebrafish microphthalmia-related protein that regulates neural-crest-derived pigment cell fate. *Development* 126:3757-3767.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25:402-408.

Lucas GA. 1968. A study of variation in the Siamese Fighting Fish, Betta splendens, with emphasis on color mutants and the problem of sex determination: Iowa State University.

Malmstrøm M, Matschiner M, Tørresen OK, Jakobsen KS, Jentoft S. 2017. Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific Data* 4:160132.

McLaughlin KA, Levin M. 2018. Bioelectric signaling in regeneration: Mechanisms of ionic controls of growth and form. *Developmental Biology* 433:177-189.

Minvielle F, Bed'Hom B, Coville J-L, Ito Si, Inoue-Murayama M, Gourichon D. 2010. The" silver" Japanese quail and the MITF gene: causal mutation, associated traits and homology with the" blue" chicken plumage. *BMC Genetics* 11:15.

Monvises A, Nuangsaeng B, Sriwattanarothai N, Panijpan B. 2009. The Siamese fighting fish: well-known generally but little-known scientifically. *ScienceAsia* 35:8-16.

Moriyama Y, Kawanishi T, Nakamura R, Tsukahara T, Sumiyama K, Suster ML, Kawakami K, Toyoda A, Fujiyama A, Yasuoka Y. 2012. The medaka zic1/zic4 mutant provides molecular insights into teleost caudal fin evolution. *Current Biology* 22:601-607.

Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13:S8.

Perathoner S, Daane JM, Henrion U, Seebohm G, Higdon CW, Johnson SL, Nüsslein-Volhard C, Harris MP. 2014. Bioelectric signaling regulates size in zebrafish fins. *PLoS Genetics* 10:e1004080.

Petit F, Sears KE, Ahituv N. 2017. Limb development: a paradigm of gene regulation. *Nature Reviews Genetics* 18:245-258.

Prost S, Petersen M, Grethlein M, Hahn SJ, Kuschik-Maczollek N, Olesiuk ME, Reschke J-O, Schmey TE, Zimmer C, Gupta DK. 2020. Improving the chromosome-level genome assembly of the Siamese fighting fish (Betta splendens) in a university master's course. *G3: Genes, Genomes, Genetics* 10:2179-2183.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 33:D501-D504.

781 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De
782 Bakker PI, Daly MJ. 2007. PLINK: a tool set for whole-genome association and
783 population-based linkage analyses. *American Journal of Human Genetics* 81:559-575.
784 Rastas P. 2017. Lep-MAP3: robust linkage mapping even for low-coverage whole genome
785 sequencing data. *Bioinformatics* 33:3726-3732.
786 Rüber L, Britz R, Tan HH, Ng PK, Zardoya R. 2004. Evolution of mouthbrooding and life-
787 history correlates in the fighting fish genus Betta. *Evolution* 58:799-813.
788 Rubin C-J, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman
789 M, Sharpe T, Ka S. 2010. Whole-genome resequencing reveals loci under selection
790 during chicken domestication. *Nature* 464:587-591.
791 Schartl M, Kneitz S, Ormanns J, Schmidt C, Anderson JL, Amores A, Catchen J, Wilson C,
792 Geiger D, Du K. 2020. The developmental and genetic architecture of the sexually
793 selected male ornament of swordtails. *Current Biology* 31:1-12.
794 Schulte CJ, Allen C, England SJ, Juárez-Morales JL, Lewis KE. 2011. Evx1 is required for
795 joint formation in zebrafish fin dermoskeleton. *Developmental Dynamics* 240:1240-1248.
796 Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D,
797 Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction
798 in threespine sticklebacks. *Nature* 428:717-723.
799 Sharma S, Londono D, Eckalbar WL, Gao X, Zhang D, Mauldin K, Kou I, Takahashi A,
800 Matsumoto M, Kamiya N. 2015. A PAX1 enhancer locus is associated with susceptibility
801 to idiopathic scoliosis in females. *Nature Communications* 6:6452.
802 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth
803 J, Hillier LW, Richards S. 2005. Evolutionarily conserved elements in vertebrate, insect,
804 worm, and yeast genomes. *Genome Research* 15:1034-1050.

805 Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting
806 effects of amino acid substitutions on proteins. *Nucleic Acids Research* 40:W452-W457.
807 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
808 assessing genome assembly and annotation completeness with single-copy orthologs.
809 *Bioinformatics* 31:3210-3212.
810 Simpson M. 1968. The display of the Siamese fighting fish, Betta splendens. *Animal Behaviour*
811 *Monographs* 1:1-73.
812 Small C, Bassham S, Catchen J, Amores A, Fuiten A, Brown R, Jones A, Cresko W. 2016. The
813 genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome*
814 *Biology* 17:258.
815 Smith HM. 1945. The freshwater fishes of Siam, or Thailand. *Bulletin of the United States*
816 *National Museum* 188:1-633.
817 Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron
818 submodel. *Bioinformatics* 19:ii215-ii225.
819 Stewart S, Le Bleu HK, Yette GA, Henner AL, Braunstein JA, Stankunas K. 2019. Longfin
820 causes cis-ectopic expression of the kcnh2a ether-a-go-go K+ channel to autonomously
821 prolong fin outgrowth. *BioRxiv* doi: https://doi.org/10.1101/790329.
822 Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J. 2015.
823 ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biology* 16:3.
824 Tassabehji M, Newton VE, Read AP. 1994. Waardenburg syndrome type 2 caused by
825 mutations in the human microphthalmia (MITF) gene. *Nature Genetics* 8:251-255.
826 Vejnar CE, Moreno-Mateos MA, Cifuentes D, Bazzini AA, Giraldez AJ. 2016. Optimized
827 CRISPR–Cas9 system for genome editing in zebrafish. *Cold Spring Harbor Protocols*
828 2016:pdb. prot086850.

Wang L, Huang SQ, Xia JH, Liu P, Wan ZY, Yue GH. 2015. Genome-wide discovery of gene-related SNPs in Barramundi *Lates calcarifer*. *Conservation Genetics Resources* 7:605-608.

Wei C, Wang H, Liu G, Wu M, Cao J, Liu Z, Liu R, Zhao F, Zhang L, Lu J. 2015. Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. *BMC Genomics* 16:194.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13:59-69.

Zhou Z, Li M, Cheng H, Fan W, Yuan Z, Gao Q, Xu Y, Guo Z, Zhang Y, Hu J. 2018. An intercross population study reveals genes associated with body size and plumage color in ducks. *Nature Communications* 9:1-10.

Figures

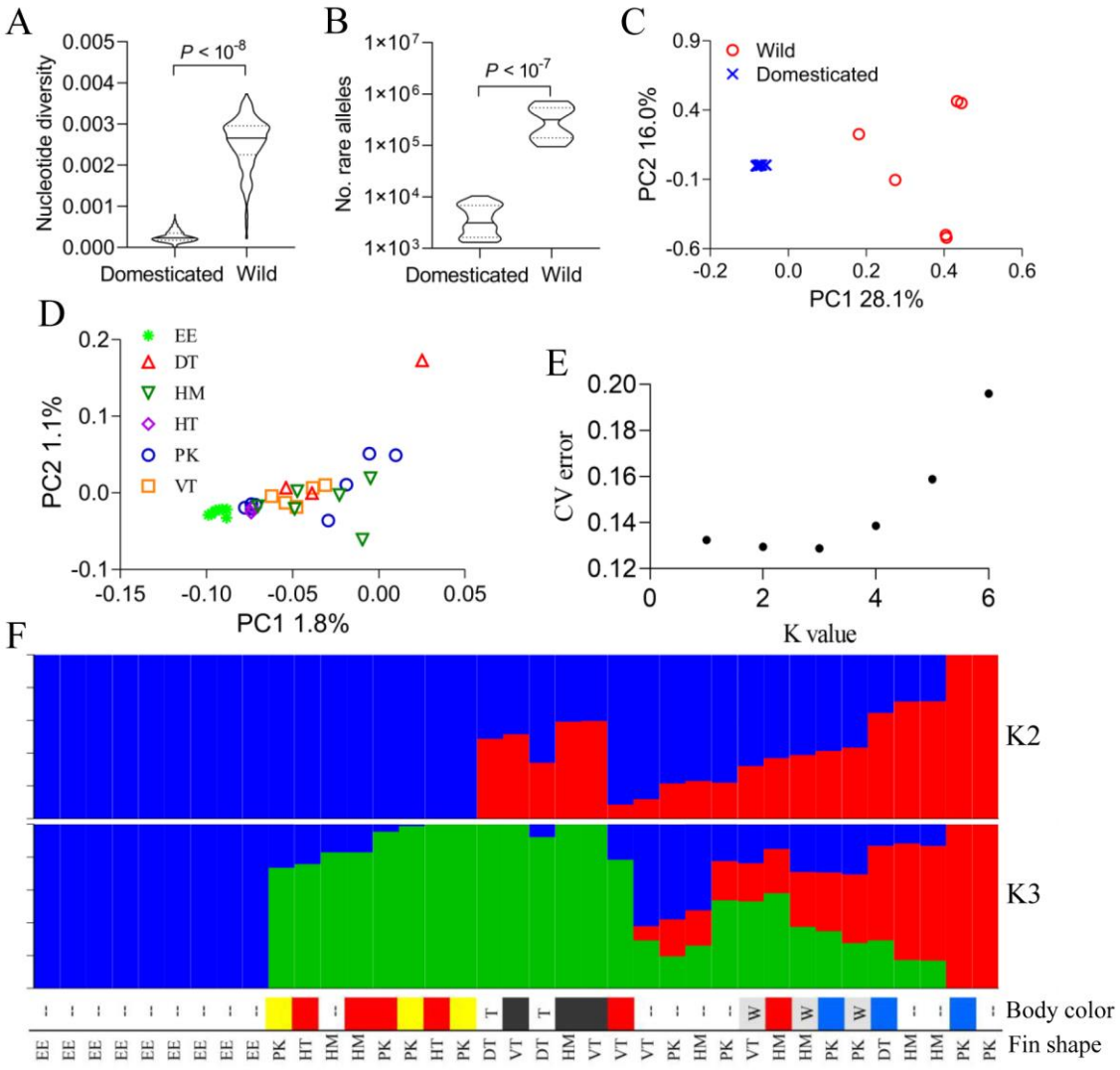
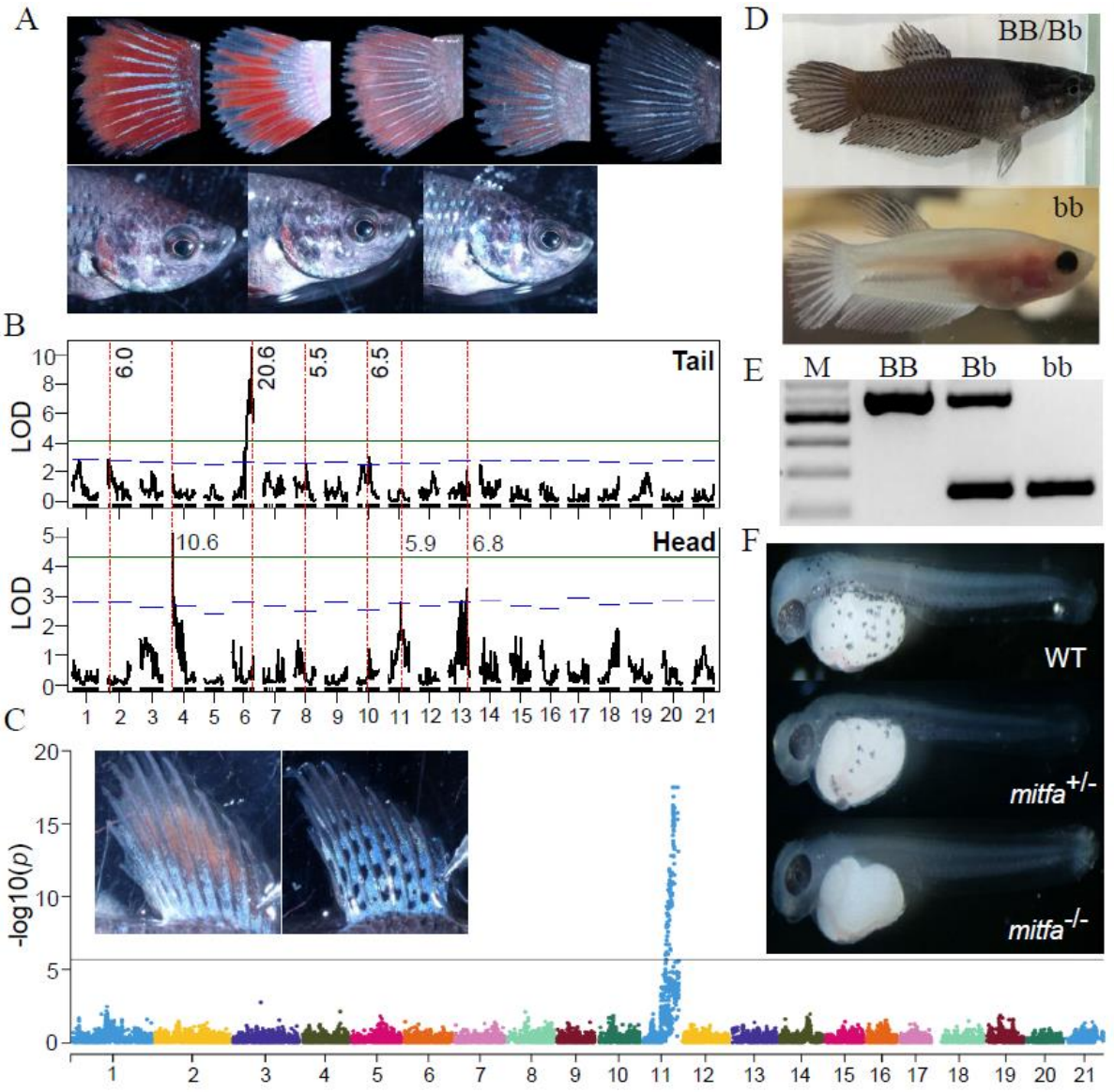


FIG. 1 Genetic diversity and population structure in fighting fish. **(A)** and **(B)** Differences of genetic diversity between domesticated and wild fish measured in nucleotide diversity and number of rare alleles, respectively. *P* values for *t*-test are shown above. **(C)** and **(D)** Population structure among domesticated and wild fish, and within domesticated fish, respectively, revealed by principal component analysis. EE, elephant ear; DT, double tail; HM, halfmoon tail; HT, horse tail; PK, Plakat tail and VT, veil tail. **(E)** The most likely number of genetic clusters (K) is inferred as 3, where shows the lowest cross validation errors. **(F)** Population structure at individual level revealed by admixture analysis, at K= 2 and 3, in domesticated fish. Major traits including body color (-- indicates too complicated color pattern to phenotype, while T and W indicate transparent and white coat color, respectively) and fin shape (codes are corresponding to those in **D**), for each individual, are also shown below.



867

868 **FIG. 2** Genetic mapping of distribution of red pigments (xanthophores) and fin spotting pattern,
869 albino mutant, and validation of *mitfa* gene as the candidate causal gene for albino mutant using
870 CRISPR/Cas9 knockout. **(A)** Variation of the distribution of red pigments in caudal fin and
871 head sections. **(B)** QTL mapping and comparison for distribution of red pigments in caudal fin
872 and head sections, where blue and green horizontal lines indicate LOD cut-off values of
873 chromosome- and genome-wide significance, respectively. Phenotypic variation explained
874 (PVE, %) by each QTL is shown at the top of each QTL region. Comparisons of QTL
875 distributions between the two traits are indicated with vertical dashed lines. **(C)** Spotted vs non-
876 spotted fin pigmentation patterns in fighting fish and genome-wide association study which
877 identified only one locus at LG11 for this trait. **(D)** and **(E)** The melanin (wild-type pigmented)
878 and albino mutant and their corresponding genotypes based on a deletion flanking *mitfa*. **(F)**
879 The wild-type pigmented fighting fish with regular pattern of melanized cells at 48 hpf (WT),
880 mosaic *mitfa*-knockout fish showing less melanized cells at 48 hpf ($mitfa^{+/-}$) and *mitfa*-
881 knockout fish showing no melanized cells throughout the whole embryo at 48 hpf ($mitfa^{-/-}$),
882 where no wild-type haplotypes are detected.

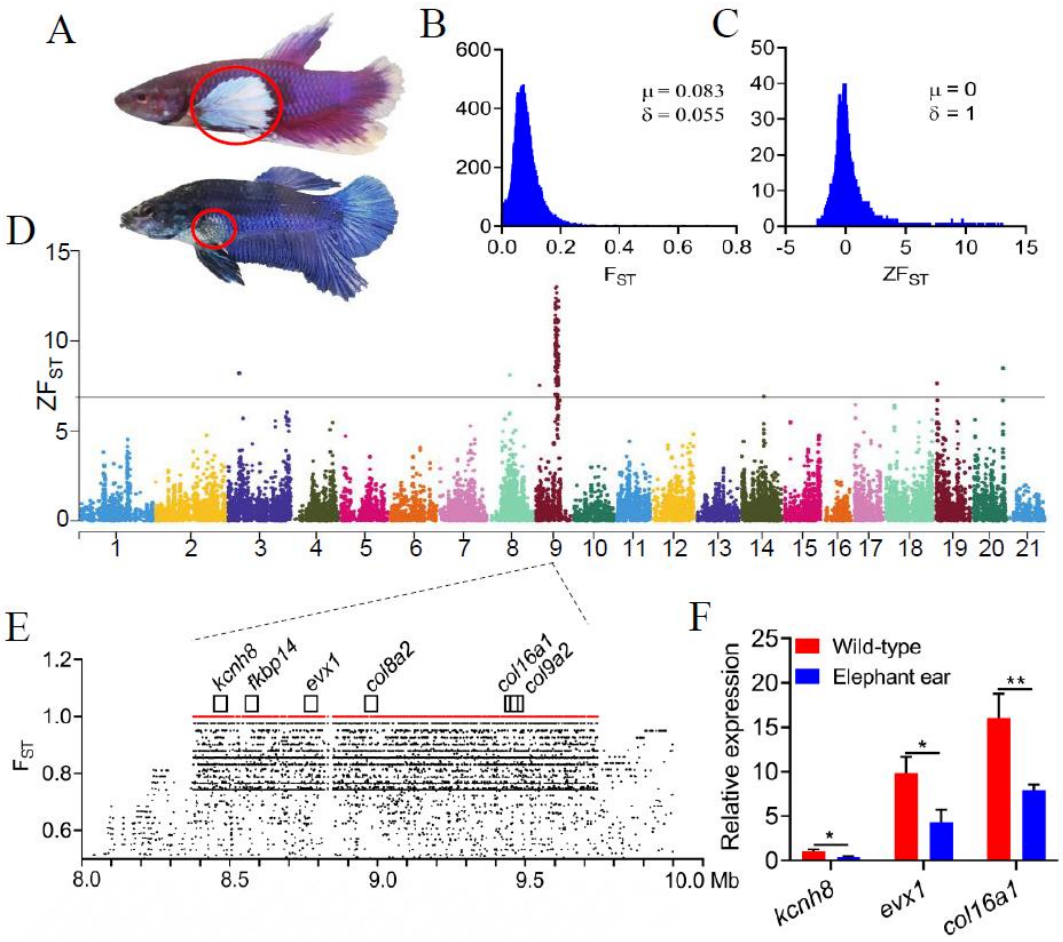


FIG. 3 Mapping and identifying candidate genes for elephant ear mutant of fighting fish. **(A)** Elephant ear mutant showing overgrowth of pectoral fin (highlighted with circle), in contrast to wild-type fish. **(B)** and **(C)** Distribution of F_{ST} and Z-transformed F_{ST} of 30 kb window size for whole genome-wide variants between elephant ear and wild-type samples, respectively. **(D)** Whole genome scan identifies a major locus at LG9 for elephant ear using Z-transformed F_{ST}. Genome-wide significance cut-off value is denoted with horizontal line. **(E)** Six protein coding genes associated with fin development and regeneration are predicted in the elephant ear haplotype with a length of ~ 1.3 Mb. Fixed variants are denoted with red. **(F)** Three genes including *kcnh8, exv1* and *col16a1* are significantly down regulated in elephant ear mutants (* $P < 0.05$, ** $P < 0.01$; n=3, *t*-test).

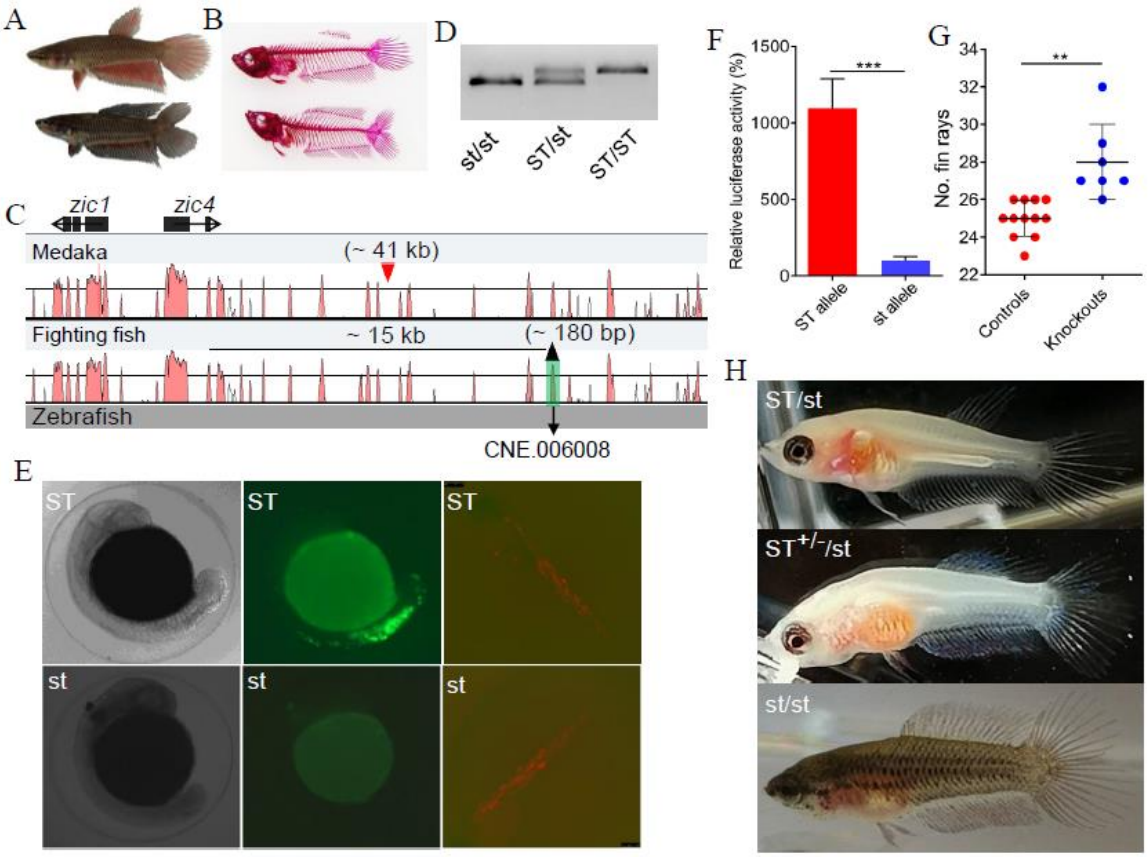


FIG. 4 Deletion in the putative enhancer of *zic1* and *zic4* is associated with double-tail mutant. **(A)** Overview of the wild-type (single-tail) and double-tail mutant fighting fish. **(B)** Skeleton staining shows the numbers of fin rays of both dorsal fin and caudal fin are significantly higher in double tail than in single tail. **(C)** Vista plotting of the genomic locus for double-tail mutation among zebrafish, fighting fish and medaka. Zebrafish is used as reference. Approximately 180-bp deletion located at ~ 15 kb downstream of *zic4* is screened overlapping with predicted CNE.006008 of double-tail allele. The insert position of transposon Albatross (~ 41 kb) in medaka *Da* locus is indicated with red triangle. **(D)** PCR screening of the deletion in single-tail and double-tail fish. **(E)** Representative fighting fish injected with enhancer detection vector ZED constructed with CNE.006008 from single-tail allele (ST) showing GFP expression predominantly in the dorsal fin and caudal fin positions, and those injected with double-tail allele (st) showing no GFP expression in the whole embryos at 24 hpf. RFP that is only detectable, particularly in muscles, since 72 hpf, is used as internal control. **(F)** Relative luciferase activity in Singapore grouper embryonic cell line transfected with pGL3-Promoter constructs including CNE.006008 region separately from the single-tail and double-tail alleles (Mann-Whitney test, *** $P < 0.001$). **(G)** The total number of fin rays of dorsal and caudal fins between genetically modified fish (n = 7) and its corresponding controls (n = 12) in CNE.006008 (Mann-Whitney test, ** $P < 0.01$). **(H)** The knockout fighting fish ($ST^{+/-}$/st), with ~ 60 % of ST allele sequences deleted at CNE.006008, shows much more fin rays both in dorsal fin and caudal fin than the single-tail (ST/st) control, but less than double-tail control (st/st). Heterozygous ST/st fish were used as recipients for the CRISPR/Cas9 injections.