Research Article

Prediction of the Sex-Associated Genomic Region in Tunas (Thunnus Fishes)

Yoji Nakamura,1 Kentaro Higuchi,2 Kazunori Kumon,3 Motoshige Yasuike,1 Toshinori Takashi,2 Koichiro Gen,2 and Atushi Fujiwara4

1Bioinformatics and Biosciences Division, Fisheries Stock Assessment Center, Fisheries Resources Institute, Japan Fisheries Research and Education Agency, 2-12-4 Fuku-ura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan
2Tuna Aquaculture Division, Aquaculture Research Department, Fisheries Technology Institute, Japan Fisheries Research and Education Agency, 1551-8 Taira-machi, Nagasaki 851-2213, Japan
3Amami Field Station, Tuna Aquaculture Division, Aquaculture Research Department, Fisheries Technology Institute, Japan Fisheries Research and Education Agency, 955-5 Hyousakiyamahara, Setouchi, Kagoshima 894-2414, Japan
4Aquatic Breeding Division, Aquaculture Research Department, Fisheries Technology Institute, Japan Fisheries Research and Education Agency, 422-1 Nakatsuhamaura, Minami-ise, Mie 516-0193, Japan

Correspondence should be addressed to Yoji Nakamura; yojnakam@affrc.go.jp

Received 12 July 2021; Revised 28 October 2021; Accepted 13 November 2021; Published 14 December 2021

Copyright © 2021 Yoji Nakamura et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Fish species have a variety of sex determination systems. Tunas (genus Thunnus) have an XY genetic sex determination system. However, the Y chromosome or responsible locus has not yet been identified in males. In a previous study, a female genome of Pacific bluefin tuna (T. orientalis) was sequenced, and candidates for sex-associated DNA polymorphisms were identified by a genome-wide association study using resequencing data. In the present study, we sequenced a male genome of Pacific bluefin tuna by long-read and linked-read sequencing technologies and explored male-specific loci through a comparison with the female genome. As a result, we found a unique region carrying the male-specific haplotype, where a homolog of estrogen sulfotransferase gene was predicted to be encoded. The genome-wide mapping of previously resequenced data indicated that, among the functionally annotated genes, only this gene, named sult1st6y, was paternally inherited in the males of Pacific bluefin tuna. We reviewed the RNA-seq data of southern bluefin tuna (T. maccoyii) in the public database and found that sult1st6y of southern bluefin tuna was expressed in all male testes, but absent or suppressed in the female ovary. Since estrogen sulfotransferase is responsible for the inactivation of estrogens, it is reasonable to assume that the expression of sult1st6y in gonad cells may inhibit female development, thereby inducing the individuals to become males. Thus, our results raise a promising hypothesis that sult1st6y is the sex determination gene in Thunnus fishes or at least functions at a crucial point in the sex-differentiation cascade.

1. Introduction

Fish species have a substantial variety of sex determination systems compared to other animals, ranging from environmentally to genetically modulated ones. In the environmental sex determination system, individual sex depends mainly on nongenetic factors such as temperature, growth rate, and density, while in the genetic system, specific genes dominantly control sex differentiation. Regarding sexuality, it is estimated that more than 98% of fish are gonochoristic [1]. Therefore, in most fish, some responsible loci should exist biasedly between males and females or differentially affect gonadal phenotypes, even if the sex is primarily determined by environmental factors. In heterogametic (XY) males or heterogametic (ZW) females, sex chromosomes may be cytologically evident markers as carrying the sex determination loci, but they are often homomorphic in fishes [2], undistinguishable from autosomes by direct microscopic
observation. In other words, fish sex chromosomes are not highly differentiated, implying that turnovers may have often occurred over the course of evolution [3]. Recently, sex determination genes have been reported in several fish species, including DMY/Dmrt1 in medaka [4, 5] and Chinese tongue sole [6], Amhr2 in fugu [7], sDy in rainbow trout [8], Hsd17b1 in yellowtail [9], and amh in Patagonian pejerrey [10] and northern pike [11]. These studies suggest that a master gene for sex determination in fish can be easily replaced by another gene, which may be followed by the above-mentioned turnover of the sex chromosome. In particular, the cases in fugu and yellowtail are prominent examples, where only a single-nucleotide nonsynonymous polymorphism in a gene (Amhr2 for fugu and Hsd17b1 for yellowtail) between the males and females is critical for sex determination. Such observations suggest that new sex determination genes can emerge by a small amount of nucleotide substitutions in a short period of evolutionary time. Thus, the sex determination genes of fish are often different for each lineage and are not always identified by deductive approaches from model organisms (that is, case studies are required to identify the genes).

Regarding tunas, fish of the genus Thunnus South 1845, it is assumed that Pacific bluefin tuna (Thunnus orientalis, PBT) has an XY genetic sex determination system [12]. However, the chromosome or locus responsible has not yet been identified. One reason for this is that no distinguishable sex chromosomes were observed in the cytological analysis. It has been reported that little tuna, Euthynnus affinis, may have a ZW sex determination system [13]. These observations suggest that the sex determination gene of Thunnus species has only recently emerged and, therefore, that the sex chromosomes are not highly divergent. Another reason is the technical difficulty in handling: since tunas are large and difficult fish to culture, in vivo experiments for identifying sex-associated loci are costly and time-consuming. Identification of the sex-associated loci of tunas is considered to have a high impact in the fishery industry as well as in ichthyology: tunas have a high market value worldwide, and tuna target fishing and aquaculture have been conducted in many countries. In aquaculture, the control of the sex ratio in tanks or ocean nets is very important for breeding, as keeping more females than males leads to an increased production of fertilized eggs. Since tunas do not show any clear sexual dimorphism, the direct observation of the gonad has been of interest as an alternative for sex identification. In this study, we resequenced and reassembled the PBT male genome using long-read sequencing technologies and examined the sex-associated regions using previously published data [17]. We have found a sex-associated gene that is paternally inherited and is involved in estrogen inactivation. The current results provide the most promising candidate for the sex determination gene of tuna.

2. Materials and Methods

2.1. Genome Sequencing and Assembly. A three-year-old male F2 individual of Pacific bluefin tuna, which was cultured at the Seikai National Fisheries Research Institute (currently Fisheries Technology Institute), was collected for genome sequencing. A frozen heart tissue was prepared from the specimen, and DNA was extracted using a DNeasy Blood & Tissue Kit (Qiagen, Venlo, Netherlands). Then, an SMRTbell template library was prepared following the manufacturer’s protocol and sequenced using the Pacific Biosciences (PacBio) Sequel platform (P6C4 chemistry)
(Pacific Biosciences, Inc., CA). From the same tissue sample, the DNA was extracted also using the standard phenol-chloroform protocol. A Chromium 10X library was prepared using Chromium Genome Library Kit v2, Genome Gel Bead Kit, Genome Chip Kit v2, and 17 Multiplex Kit (10X Genomics, CA) following the manufacturer’s protocol, and the linked reads were sequenced using the Illumina NovaSeq 6000 platform (Illumina, Inc., CA). The PacBio reads were first assembled by HGAP4 (SMRT Link v5.1.0.26412) [18]. The contigs obtained were corrected using NovaSeq linked-reads by Pilon v1.23 [19] and Tigmint v1.1.2 [20] and finally scaffolded by ARKS v1.0.2 [21] in combination with LINKS v1.8.5 [22]. The scaffold sequences obtained were submitted to the DNA Data Bank of Japan under accession nos. BOUD01000001-BOUD01000948.

The final scaffold sequences were assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) version 4.0.5 [23] with a set of Actinopterygii orthologs (actinopterygii_odb10). The genes around the sex-associated region were predicted based on the protein sequences of model fish species in the Ensembl database (release 99) [24], namely, *Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*, and *Tetraodon nigroviridis*, by Exonerate v2.4.0 [25].

The female genome sequence and the southern bluefin tuna (SBT) male genome sequence were obtained from the GenBank database (accession nos. BKCK00000000 and CAJUDC01000000, respectively). The nucleotide sequence comparison was performed by MUMmer v4 [26]. The repeat library of PBT genome was constructed by RepeatModeler Open-1.0, and the repetitive regions were masked by RepeatMasker Open-4.0 (http://www.repeatmasker.org). The duplicated regions were further detected by self-to-self BLASTN [27] in each of the PBT male and female genomes, and the sequences detected were used also for masking. To examine the phased genomic regions, the linked reads were assembled using Supernova2 (v2.1.1) [28].

### 2.2. Genome-Wide SNV/CNV Analysis

The sequenced reads for 31 PBT individuals (15 males and 16 females) [16] were downloaded from the NCBI SRA database (accession no. DP3005507). The read sequences were extracted using SRA Toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) and then preprocessed using Trimmomatic v2.2.5 [30], according to Suda et al. Mapping to the scaffold sequences and single-nucleotide variant calling were conducted using MapCaller v0.9.9.37 [31]. After filtering out inadequate sites (not biallelic, depth < 20), genome-wide association (GWA) between males and females was estimated through 2 × 3 contingency tables using the R package, rrBLUP [32], with minor allele frequency > 0.05. Statistical significance was defined as $P < 0.05$ after Bonferroni’s correction based on the total number of variant sites examined. For relative depth analysis, raw read depth per nucleotide site was counted using Samtools (samtools depth) [33] and then divided by the median depth. Copy number variations (CNVs) were detected by CNVcaller [34], where the above-mentioned mapping data were scanned by a 500bp window and regional copy numbers were obtained by merging them in the neighboring windows. The difference in the regional copy number between males and females was tested using the Mann-Whitney U test with Bonferroni’s correction. Furthermore, the regions carrying either of the three types of copy numbers (0, 0.5, or 1) among the 31 samples were selected, and the GWA was estimated in the same way as the SNV-based GWA scan mentioned above.

### 2.3. Sequence Analysis of Sex-Associated Genes

The transcriptome data of southern bluefin tuna [15] were downloaded from the NCBI SRA database (accession no. SRP059929). The read sequences were extracted using SRA Toolkit, preprocessed by Trimomatic, and assembled using Trinity v2.8.4 [35]. From the contig sequences obtained, protein-coding sequences were predicted using TransDecoder v5.0.0 [35]. Homologs of PBT genes were searched using BLASTP [36] from the contig sequences, and the transcriptional variants were manually checked based on the sequence alignment using MAFFT v7.310 [37]. Using the Trinity contig sequences as a reference, gene expression was measured using RSEM v1.3.1 [38]. For molecular phylogenetic analysis, homologs in other fish genomes were downloaded from the Ensembl database. The coding sequences were then aligned using MAFFT, where the deduced protein sequences were aligned and converted into the corresponding codon alignments. The maximum-likelihood tree was constructed using RAxML-NG v0.9.0 [39] and visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

### 3. Results

#### 3.1. Genome Assembly and Validation of Sex-Associated Markers

We constructed the genome sequence of male PBT by de novo assembly, yielding a total of 948 scaffolds (>1000 bp) and totaling 827 Mb (Table 1).

Using BUSCO for quality assessment, this assembly completely captured 96.3% of Actinopterygii orthologs (3505 out of 3460). On the other hand, the score for the female genome was 89.3% (3253 out of 3640), consistent with the previous report. Using the sex-associated PCR primer sequences reported, we performed a BLAST search against the current male genome and found that two scaffolds, namely, M175 and M44, carried the regions amplified by the primer sets (two male-specific sets, scaf64_3726404_F/R and scaf64_3726411_F/R, and a common primer set, scaf64_3724591_F/R) (Table 2).

In scaffold M175 (232,647 bp in total), the regions amplified by the three primer sets were closely located. The estimated amplicon sizes were congruent with those previously observed only in males: 113 bp for scaf64_3724604_F/R, 143 bp for scaf64_3726411_F/R, and 142 bp for scaf64_3724591_F/R, respectively. Scaffold M44 (4,583,724 bp in total) had a single region amplified by scaf64_3724591_F/R, and the estimated amplicon size was 149 bp, which was equivalent to that observed previously in both males and females. Thus, based on the estimated PCR products, M175 carried all of the male-specific regions and...
the relative depths of SLRM175 were close to 1 in both males and females. These observations strongly suggest that, in shotgun sequencing libraries, the DNA content derived from SLRM175 in the male samples is always half that derived from the average scaffolds, reflecting the ploidy or copy number. The majority of chromosomal regions, except for highly repetitive ones or those of polyploidized genomes, should exist as singletons in a diploid manner (2n), and the relative depths for many regions would be approximately 1. Therefore, our observations suggest that SLRM175, with a relative depth of ~0.5, should exist in a haploid manner (1n) only in males, while SLRM44 should exist in a diploid manner, like many other regions, in both males and females.

3.3. Genome-Wide Association Analysis. Since the results obtained until now appeared to be inconsistent with those previously reported (see Discussion), we validated the previous mapping and GWA results based on the female genome. We mapped the 31 resequenced read sets to the female reference genome and reperformed GWA analysis. As a result, the previous result was roughly reproduced, with significant peaks observed in multiple scaffolds, including F64 (Figures 5(a) and 5(c)). However, when the mapped reads between the female and male genomes were compared, over 10% of the reads mapped to female SLRF64 in the 15 male samples were found to be potentially derived from SLRM175 (Figure 6). This observation implies that the reads derived from the male-specific region, such as SLRM175, would have been incorrectly mapped to paralogous regions in the female genome in the previous study. To avoid the effect of such cross mapping, we performed the GWA scan again after the duplicated regions were thoroughly masked by RepeatMasker and BLASTN. This result was in contrast to that from the unmasked reference: no significant peaks were observed in the masked sequences (Figures 5(b) and 5(d)). Similarly, GWA scans were performed using the current male genome, but no significant peaks were found in the masked sequences (Supplementary Figure 4).

We then tested another approach to identify the regions carrying male-specific haplotypes, such as SLRM175. Using the mapping data for the 31 individuals, we computed copy number variations (CNVs) across the male genome and explored the regions whose copy numbers were significantly different between males and females. As a result, only one candidate was found in M175 (18,001–78,750 bp), very close to SLRM175 and the copy number in males was significantly larger than that in females (Figure 7(a)). All of the males had 0.5 copies of the detected region and all of the females had no copy (i.e., absent), in concordance with the relative depth analysis results (Figure 4). We computed genome-wide associations for the haploid regions by converting the copy numbers (0, 0.5, and 1) to three quasigenotypes (0/0, 0/1, and 1/1). As a result, only one significant peak for this region was detected in scaffold M175 (Figure 7(b)).

3.4. Protein-Coding Genes in Scaffold M175. We scanned the M175 sequence and predicted 12 protein-coding genes (Figure 8 and Supplementary Table 1). Of these, two genes,
g3 (32,614–64,628 bp) and g4 (76,852–77,963 bp), were located in the haploid region (18,001–78,750 bp) detected by CNV analysis. The deduced amino acid sequence of g3 was similar to that of estrogen sulfotransferase encoded by sult1st6 (e.g., with ~83% identity to ENSORLP00000007539 of Oryzias latipes). Therefore, the homolog of PBT was named sult1st6y. In the case of g4, the deduced amino acid sequence was similar to that of ENSTRUP00000072273 (gene ID: ENSTRUG00000033234) of Takifugu rubripes, with ~70% identity. As a “novel gene” in the Ensembl database, the function was unknown. However, the domain “Integrase, catalytic core” was predicted by Pfam and PROSITE and found to overlap with the Gene3D domain “Ribonuclease H superfamily.” The neighboring 10 genes (g1, g2, and g5–g12) were novel or related to transposable elements like g4 (Supplementary Table 1), and thus, only g3 was a functionally annotated gene. We found no homologs of sult1st6y (g3) in the scaffolds M44/F64. Instead, a paralog was found in male scaffold 30 (M30, 6594208–6602041 bp), named, sult1st6a, and the nucleotide and protein identities to sult1st6y were 92% and 89%, respectively. Male scaffold M30 corresponded to a part of female scaffold 47 (F47) (Supplementary Figure 5), where sult1st6a was also found. In the vicinity of sult1st6a, six genes were encoded, namely, paqr4b (progestin and adipoQ receptor family member IVb), pelo (pelota mRNA surveillance and ribosome rescue factor), ppp1cab (serine/threonine-protein phosphatase alpha-2 isoform), atp2a1l (sarcoplastic/endoplasmic reticulum calcium ATPase 1), mto1 (mitochondrial tRNA translation optimization 1), and adprh (ADP-ribosylarginine hydrolase). The conserved synteny (paqr4b-pelo-ppp1cab-atp2a1l-sult1st6-mto1-adprh) was observed between PBT and other fish species (Supplementary Table 2).

To determine whether these sult1st6 genes are present in other tuna species, we downloaded the RNA-seq data of SBT, which were obtained from 10 male testis samples and 10 female ovary samples. We assembled the reads using Trinity and conducted a homology search against the contigs using PBT sult1st6y/sult1st6a sequences. As a result, we found the homologs of sult1st6y/sult1st6a in SBT, although the Trinity contigs were originally classified into a single cluster. We conducted phylogenetic analysis using PBT and SBT sult1st6 and the homologs in the Ensembl database (Figure 9 and Supplementary Table 2). The phylogenetic tree showed that each SBT homolog formed a clade with each of PBT’s sult1st6y/sult1st6a, respectively, suggesting that SBT also had two paralogs of sult1st6 and the duplication to sult1st6y and sult1st6a occurred in the common ancestor of PBT and SBT. The SBT sult1st6
paralogs were confirmed also by the published male genome annotation (accession nos. XM_042430888 for sult1st6y on chromosome 13 and XM_042397350 for sult1st6a on chromosome 20, respectively). Using the transcript contigs as a reference, we mapped the reads in each of the 20 samples to those and measured the gene expression levels (Figure 10). As a whole, the expression of SBT’s sult1st6y/ sult1st6a was low compared to that of the control (0–2% of β-actin in FPKM). In detail, the expression of sult1st6a was observed in both the testis and ovary samples, although it tended to be low in the latter (0.003–0.05% of β-actin in FPKM, except for 0% for SRR2079766) (Figure 10(a)). The expression of sult1st6y was observed in all 10 testis samples, but none (i.e., FPKM = 0, undetectable) in ovary samples (Figure 10(b)). Here, we categorized the samples with FPKM > 0 as “expressed” and those with FPKM = 0 as “not expressed.” Thus, the expression of sult1st6y was significantly biased between the testes and ovaries (P < 0.05, Fisher’s exact test).

4. Discussion

In this study, the male genome of Pacific bluefin tuna was sequenced using long-read sequencing technologies. The scaffold sequences assembled were more contiguous than those published in 2013, which were obtained from pyrosequencing and short paired-end reads. Completeness assessment indicates that the current version is more complete than that of the recent female genome, although the number of scaffolds (948) is larger than that for the female genome (444). Therefore, the current genome is expected to be useful for the validation and extension of previous results regarding

![Figure 2: Sequence comparison between male scaffold M44 and female scaffold F64. The 6.5 kb SLRs are indicated with arrows, respectively.](image1)

![Figure 3: Mapped read depths around SLRs in resequenced PBT samples. For each of the female (accession no. DRR177382) and male (accession no. DRR177383) samples, SLR is highlighted in gray. The read depths for all the 31 samples are shown in Supplementary Figure 3.](image2)
**Figure 4:** Comparison of relative read depths between males and females for SLR\textsubscript{M44}/SLR\textsubscript{M175}. Each of the male/female samples is indicated by the accession number (DRR number).

**Figure 5:** Genome-wide association with the sex of PBT using the female reference sequences. For the unmasked (a and c) and masked (b and d) references, Manhattan plot and Q-Q plot are shown, respectively. In Manhattan plots, genome-wide significance cutoffs based on Bonferroni’s correction are indicated by a red line, and the significant peak in scaffold F64 is indicated by a triangle. P values lower than 10\textsuperscript{-15} are treated as 10\textsuperscript{-15} for convenience.
the sex identification of PBT. In the previous study, three PCR primer sets were designed from the sequence of female scaffold F64, so that the amplification patterns were different between PBT males and females. In this study, we first searched for the locus amplified by the primer sets in the current male genome and found two regions: one in scaffold M44 and the other in M175. Based on the results of the large-scale sequence comparison, M44 was found to be the counterpart of F64, while M175 was male-specific. Since M44 and F64 were well aligned with each other, including 6.5 kb SLRs (SLR_{M44}/SLR_{F64}), it is unlikely that large-scale misassemblies are present in M44/F64. The PCR amplification pattern in SLR_{M44} estimated using primer sequences, was identical to that observed in females. Regarding M175, all the PCR amplicons with male-specific sizes were observed within the 10.7 kb region (SLR_{M175}). It is worth noting that the PCR primer sets were originally designed based on the SNVs predicted in F64 by GWA analysis. Therefore, according to the previous results, M44 and M175 should be allelic to each other. However, these
sequences are quite dissimilar to each other. Furthermore, based on the mapping result of resequenced data, the relative read depths to that for the average scaffold data were \((\text{SLR}_M^{44}, \text{SLR}_M^{175}) \approx (1, 0)\) in females but \((\text{SLR}_M^{44}, \text{SLR}_M^{175}) \approx (1, 0.5)\) in males. These estimates indicate that \(\text{SLR}_M^{44}\) should exist in a diploid manner in both males and females, whereas \(\text{SLR}_M^{175}\) should exist in a haploid manner only in males. If \(\text{SLR}_M^{44}\) was allelic to \(\text{SLR}_M^{175}\) in males, the relative read depths (\(\text{SLR}_M^{44}, \text{SLR}_M^{175}\)) should be \((0.5, 0.5)\). The result of relative read depth is congruent with that of phased assembly by Supernova2, where two pseudohaplotype scaffolds carrying homozygotic \(\text{SLR}_M^{44}\) were obtained, but an unphased one was obtained for \(\text{SLR}_M^{175}\). These results indicate that \(\text{SLR}_M^{44}\) and \(\text{SLR}_M^{175}\) are located at different loci in the PBT male genome, where \(\text{SLR}_M^{175}\) is the male-specific structural variant (i.e., inserted in males or deleted in females) and is inherited paternally. Moreover, in the SBT male genome, the counterparts of \(M_44\) and \(M_175\) were located on different chromosomes, respectively. Thus, \(\text{SLR}_F^{64}\) (or \(\text{SLR}_M^{44}\)) reported in the previous study was not associated with PBT sex, although the PCR primer sets designed in the previous study. Therefore, it is possible that the current male genome contains other scaffolds carrying sex-associated structural variants. To check this possibility, we conducted a genome-wide scan based on a CNV detection scheme using mapping data. As a result, we observed that only \(\text{M}_{175}\) had a 61-kb region (18,001–78,750 bp) where the CNV was significantly different between males and females. Since this region is very close to \(\text{SLR}_M^{44}\), eventually the candidate sex-associated locus was narrowed down to a single region around or including \(\text{SLR}_M^{175}\). Regarding the CNV detected, we found that PBT females have no copy, while males have 0.5 copies of the region, indicating that this region exists in a haploid manner in males or, equivalently, that this region is paternally inherited. Thus, we conclude that this region (61-kb region + \(\text{SLR}_M^{175}\)) is the only candidate for the Y-linked locus in PBT. In \(\text{M}_{175}\), we predicted a total of 12 protein-coding genes, two of which \((g_3, g_4)\) were included in the predicted haploid region and one \((g_3)\) of which was annotated as encoding estrogen sulfotransferase \(\text{SULT1ST6y}\), while the other \((g_4)\) was functionally unknown. We note that \(\text{g}_4\) has the domain "Integrase, catalytic core" or "Ribonuclease H superfamily" which is a feature of the retroviral integrase superfamily [40]. Regarding the other 10 predicted genes in \(\text{M}_{175}\), the functions are unknown. In particular, \(g_1, g_2, \) and \(g_3\) may be involved in transposons according to domain prediction. Therefore, around the male-specific haploid region, only the centermost \(g_3\), namely, \(\text{sult1st6y}\), is functionally annotated and
sandwiched between possibly transposable elements, implying that the g3 locus may have been unstable and transposed from the ancestral locus.

The current study suggests that $sult1st6$ is the most convincing gene involved in the sex determination or differentiation of PBT. The encoding protein, SULT1ST6, belongs to the family of cytosolic sulfotransferases (SULTs) and has been studied as a detoxifying enzyme in mammals [41]. Regarding fish species, the SULT genes have been investigated mainly in zebrafish, and 20 distinct paralogs
SULT1s, 3 SULT2s, 5 SULT3s, 1 SULT4, 1 SULT5, and 1 SULT6) have been characterized. Among these, SULT1ST6, one of the nine SULT1 paralogs, displays strict substrate specificity for estrogens (e.g., estrone). This gene is thought to be involved in organogenesis, such as the development of the eye and muscle in zebrafish [42]. On the other hand, there is no report about the relationship between estrogen sulfotransferase and sex development. In fish, steroid hormones are considered to be the key factor for sex determination [2, 43, 44], wherein sex can change depending on the levels of estrogen or androgen production in gonad cells. The gene directly contributing to steroidogenesis has also been reported as a sex determination gene in yellowtails [9]. Since endogenous estrogens are inactivated by sulfation [45], it may not be surprising that the function of sult1st6 might be explained by a sort of neofunctionalization among possible evolutionary mechanisms of duplicated genes [46]: sult1st6y could have been immune from the original role of sult1st6 as a redundant copy, leading to the gain of a new function related to sex development under relaxed selection. Here, it is assumed that some mutations may have occurred in the regulatory region of sult1st6y, which changes the expression pattern but not the enzymatic function itself as estrogen sulfotransferase. In fact, it has been reported that a copy of duplicated genes may be a sex determination gene in fishes [8, 10, 47]. We found that homologs of sult1st6y and sult1st6a were present in the southern bluefin tuna. Phylogenetic analysis suggested that gene duplication occurred in the common ancestor of PBT and SBT. The duplications of sult1st6 were observed also in other fish lineages, but the emergence of sult1st6y and sult1st6a was specific to the tuna lineage. In particular, the tuna’ paralogs are located on different chromosomes based on the SBT male genome data, while those in other fish species are tandemly or closely located on the same chromosome. We speculate that the tissues in which sult1st6y is expressed and the timing of its expression may have been differentiated from those of sult1st6a in the proto-paternal line after the gene duplication in the ancestor of PBT and SBT. Actually, SBT sult1st6y was expressed in male’s testis but not in female’s ovary according to the transcriptome. This significant expression bias suggests that sult1st6y might have some sort of male-specific role in SBT, in particular, related to testis development. Regarding sult1st6a, the quantitative expression profile implies that this paralog as well as sult1st6y might be involved in testis development. In other words, sult1st6a might be bifunctionally utilized in the tissues assumed for fish sult1st6 and testis. On the other hand, since sult1st6a was poorly expressed in some male samples, it may not be always essential for testis development. Thus, we speculate that sult1st6a may be expressed in concert with sult1st6y to some extent, and the role of sult1st6a may be minor or secondary to that of sult1st6y. It should be noted that the expression of sult1st6a was at least detectable in...
females (FPKM > 0 in nine out of 10 samples), although it was at a much lower level than in males. Therefore, the observation that the FPKMs of sult1st6y were zero in all the female samples cannot be explained by random factors, such as variations among individuals or loss of mRNA in library preparation, raising a hypothesis that sult1st6y might be originally absent from SBT females like PBT females. Although this hypothesis will need to be tested by sequencing the entire SBT female genome, our observations suggest that SBT sult1st6y may also be involved in sex development like PBT sult1st6y.

The results presented in this study provide an attractive hypothesis for sult1st6y being a sex determination gene in Thunnus fishes, or at least functioning at an important point in the sex-differentiation cascade. Our hypothesis may be tested by further genetic experiments, although the handling of tuna in tanks is not easy. Recently, genome-editing technology has been applied for the mutagenesis of tuna [48]. For example, it may be possible to check whether the sex ratio of larvae is heavily biased by the knockout of sult1st6y in fertilized eggs. Lastly, it is worth emphasizing that the CNV-based approach used in this study is very useful for identifying a trait-associated region using short reads when it is involved in large structural variants. In particular, when the region contains recently duplicated sequences, the traditional SNV-based approach may not only miss structural variants but also produce artifacts due to cross-mapping to unassociated regions. In other fish species, sex-associated SNVs have often been observed in multiple scaffolds [49–51], some of which might be cases of cross-mapping of short reads, unless the sexes are polygenically determined. We believe that a combination of traditional SNV- and CNV-based approaches is useful for the identification of sex-associated genes in fish species with a genetic sex determination system.

5. Conclusions

In this study, we constructed the highly contiguous genome sequence of male PBT and identified a single region present only in the male individuals. Genome-wide read depth scan suggested that this region was, among all the regions examined, the only one candidate paternally inherited. Therefore, the carrying genes were the only candidates of sex-associated gene in PBT. In particular, a protein-coding gene, namely, sult1st6y, was predicted to encode estrogen sulfotransferase responsible for inactivation of estrogens, strongly suggesting that this gene is involved in PBT’s sex determination or differentiation. In addition, we have found a paralog of sult1st6y, namely, sult1st6a, in a locus unassociated to PBT’s sex in both of males and of females, and hypothesize that this gene may be involved in common estrogen metabolism. Both of sult1st6y and sult1st6a were found also in the SBT transcriptome, where sult1st6y was expressed in male testis, and absent or suppressed in female ovary. Since duplication of sult1st6y and sult1st6a predate the divergence of PBT and SBT, the function of sult1st6y may have been spatiotemporally differentiated from sult1st6a in the proto-paternal line of tuna species. Based on these results, we present a plausible hypothesis that sult1st6y is the sex-determinant gene in tuna or the gene almost close to the determination point. In this study, our results have provided a promising gene target for understanding tuna’s sex determination system, although tunas are still difficult fishes for in vivo experiments. Finally, we address that our CNV-based approach is valid when the sex-associated loci may be derived from structural variants, which might have been missed in traditional SNV-based one. The combination of both approaches will be useful for exploring sex determination genes in a variety of fish species.

Data Availability

The genome data of Pacific bluefin tuna used to support the findings of this study have been deposited in the DNA Data Bank of Japan under accession nos. BOUD01000001-BOUD01000948.

Disclosure

Kentaro Higuchi’s current address is General Planning and Coordination Department, Headquarters, Japan Fisheries Research and Education Agency, 1-1-25 Shin-urashima, Kanagawa-ku, Yokohama, Kanagawa 221-8529, Japan. The preprint version is available at bioRxiv (doi:10.1101/2021.06.15.448492v2).

Conflicts of Interest

The authors declare no conflicts of interest in this work.

Acknowledgments

This work was supported by the Japan Fisheries Research and Education Agency (Grant No. 3BA101). We would like to thank Editage (http://www.editage.jp) for English language editing.

Supplementary Materials

Supplementary 1. Supplementary Table 1: protein-coding genes predicted in scaffold M175. Supplementary Table 2: list of genes used for phylogenetic analysis.

Supplementary 2. Supplementary Figure 1: comparison between PBT male scaffolds and SBT male chromosomal regions. Supplementary Figure 2: comparison between the pseudo-haploid scaffolds obtained from Supernova2 and scaffolds M44/M175. Supplementary Figure 3: mapped read depths around SLRs in 31 resequenced PBT samples. Supplementary Figure 4: genome-wide association with the sex of PBT using male masked reference sequences. Supplementary Figure 5: comparison between male and female scaffolds encoding sult1st6a.

References

[1] T. J. Pandian, *Sex Determination in Fish*, CRC Press, 2011.
[2] R. H. Devlin and Y. Nagahama, "Sex determination and sex differentiation in fish: an overview of genetic, physiological,
and environmental influences,” *Aquaculture*, vol. 208, no. 3–4, pp. 191–364, 2002.

[3] M. W. Pennell, J. E. Mank, and C. L. Peichel, “Transitions in sex determination and sex chromosomes across vertebrate species,” *Molecular Ecology*, vol. 27, no. 19, pp. 3950–3963, 2018.

[4] M. Matsuda, Y. Nagahama, A. Shinomiya et al., “DMY is a Y-specific DM-domain gene required for male development in the medaka fish,” *Nature*, vol. 417, no. 6888, pp. 559–563, 2002.

[5] I. Nanda, M. Kondo, U. Hornung et al., “A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, Oryzias latipes,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 18, pp. 11778–11783, 2002.

[6] Z. Cui, Y. Liu, W. Wang et al., “Genome editing reveals dmrt1 as an essential male sex-determining gene in Chinese tongue sole (Cynoglossus semilaevis),” *Scientific Reports*, vol. 7, no. 1, article 42213, 2017.

[7] T. Kamiya, W. Kai, S. Tasumi et al., “A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, Takifugu rubripes (fugu),” *PLoS Genetics*, vol. 8, no. 7, article e1002798, 2012.

[8] A. Yano, R. Guyomard, B. Nicol et al., “An Immune-Related Gene Evolved into the Master Sex-Determining Gene in Rainbow Trout, Oncorhynchus mykiss,” *Current Biology*, vol. 22, no. 15, pp. 1423–1428, 2012.

[9] T. Koyama, M. Nakamoto, K. Morishima et al., “A SNP in a Steroidogenic Enzyme Is Associated with Phenotypic Sex in Seriola Fishes,” *Current Biology*, vol. 29, no. 11, pp. 1901–1909.e8, 2019.

[10] R. S. Hattori, Y. Murai, M. Oura et al., “A Y-linked anti-Mullerian hormone duplication takes over a critical role in sex determination,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 8, pp. 2955–2959, 2012.

[11] Q. Pan, R. Feron, A. Yano et al., “Identification of the master sex determining gene in northern pike (Esox lucius) reveals restricted sex chromosome differentiation,” *PLoS Genetics*, vol. 15, no. 8, article e1008013, 2019.

[12] Y. Agawa, M. Iwaki, T. Komiya et al., “Identification of male sex-linked DNA sequence of the cultured Pacific bluefin tuna Thunnus orientalis,” *Fisheries Science*, vol. 81, no. 1, pp. 113–121, 2015.

[13] R. Yazawa, Y. Takeuchi, Y. Machida et al., “Production of triploid eastern little tuna, Euthynnus affinis (Cantor, 1849),” *Aquaculture Research*, vol. 50, no. 5, pp. 1422–1430, 2019.

[14] Y. Nakamura, K. Mori, K. Saitoh et al., “Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 27, pp. 11061–11066, 2013.

[15] I. Bar, S. Cummins, and A. Elizur, “Transcriptome analysis reveals differentially expressed genes associated with germ cell and gonad development in the southern bluefin tuna (Thunnus maccoyii),” *BMC Genomics*, vol. 17, no. 1, p. 217, 2016.

[16] A. Suda, I. Nishiki, Y. Iwasaki et al., “Improvement of the Pacific bluefin tuna (Thunnus orientalis) reference genome and development of male-specific DNA markers,” *Scientific Reports*, vol. 9, no. 1, article 14450, 2019.

[17] Y. Nakamura, K. Higuchi, K. Kumon et al., *Prediction of the sex-determination gene in tunas (Thunnus fishes)*, bioRxiv, 2021.

[18] C. S. Chin, D. H. Alexander, P. Marks et al., “Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data,” *Nature Methods*, vol. 10, no. 6, pp. 563–569, 2013.

[19] B. J. Walker, T. Abeel, T. Shea et al., “Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement,” *PLoS One*, vol. 9, no. 11, article e112963, 2014.

[20] S. D. Jackman, L. Coombe, J. Chu et al., “Tigemt: correcting assembly errors using linked reads from large molecules,” *BMC Bioinformatics*, vol. 19, no. 1, p. 393, 2018.

[21] L. Coombe, J. Zhang, B. P. Vandervalk et al., “ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers,” *BMC Bioinformatics*, vol. 19, no. 1, p. 234, 2018.

[22] R. L. Warren, C. Yang, B. P. Vandervalk et al., “LINKS: scalable, alignment-free scaffolding of draft genomes with long reads,” *Gigascience*, vol. 4, no. 1, p. 35, 2015.

[23] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs,” *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, 2015.

[24] B. L. Aken, P. Achuthan, W. Akanni et al., “Ensembl 2017,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D635–D642, 2017.

[25] G. S. Slater and E. Birney, “Automated generation of heuristics for biological sequence comparison,” *BMC Bioinformatics*, vol. 6, no. 1, p. 31, 2005.

[26] G. Marcais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin, “MUMmer4: a fast and versatile genome alignment system,” *PLoS Computational Biology*, vol. 14, no. 1, article e1005944, 2018.

[27] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[28] N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church, and D. B. Jaffe, “Direct determination of diploid genome sequences,” *Genome Research*, vol. 27, no. 5, pp. 757–767, 2017.

[29] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, 2014.

[30] J. Gonzalez-Dominguez and B. Schmidt, “PartDRe: faster parallel duplicated reads removal tool for sequencing studies,” *Bioinformatics*, vol. 32, no. 10, pp. 1562–1564, 2016.

[31] H.-N. Lin and W.-L. Hsu, “MapCaller – an integrated and efficient tool for short-read mapping and variant calling using high-throughput sequenced data,” no. article 783605, 2019bioRxiv, 2019.

[32] J. B. Endelman, “Ridge regression and other kernels for genomic selection with R package rrBLUP,” *The Plant Genome*, vol. 4, no. 3, pp. 250–255, 2011.

[33] H. Li, B. Handsaker, A. Wysoker et al., “The sequence alignment/map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[34] X. Wang, Z. Zheng, Y. Cai et al., “CNVcaller: highly efficient and widely applicable software for detecting copy number variations in large populations,” *Gigascience*, vol. 6, no. 12, pp. 1–12, 2017.
[35] B. J. Haas, A. Papanicolaou, M. Yassour et al., “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis,” Nature Protocols, vol. 8, no. 8, pp. 1494–1512, 2013.

[36] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” Nucleic Acids Research, vol. 25, no. 17, pp. 3389–3402, 1997.

[37] K. Katoh, G. Asimenos, and H. Toh, “Multiple alignment of DNA sequences with MAFFT,” Methods in Molecular Biology, vol. 537, pp. 39–64, 2009.

[38] B. Li and C. N. Dewey, “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome,” BMC Bioinformatics, vol. 12, no. 1, p. 323, 2011.

[39] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, “RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference,” Bioinformatics, vol. 35, no. 21, pp. 4453–4455, 2019.

[40] M. Nowotny, “Retroviral integrase superfamily: the structural perspective,” EMBO Reports, vol. 10, no. 2, pp. 144–151, 2009.

[41] M. Suiko, K. Kurogi, T. Hashiguchi, Y. Sakakibara, and M. C. Liu, “Updated perspectives on the cytosolic sulfotransferases (SULTs) and SULT-mediated sulfation,” Bioscience, Biotechnology, and Biochemistry, vol. 81, no. 1, pp. 63–72, 2017.

[42] S. Yasuda, C. C. Liu, S. Takahashi et al., “Identification of a novel estrogen-sulfating cytosolic SULT from zebrafish: molecular cloning, expression, characterization, and ontogeny study,” Biochemical and Biophysical Research Communications, vol. 330, no. 1, pp. 219–225, 2005.

[43] M. Nakamura, “The mechanism of sex determination in vertebrates—are sex steroids the key-factor?,” Journal of Experimental Zoology Part A, Ecological Genetics and Physiology, vol. 313A, no. 7, pp. 381–398, 2010.

[44] T. Yamamoto, “Sex Differentiation,” in Fish Physiology, W. S. Hoar and D. J. Randall, Eds., pp. 117–175, Academic Press, 1969.

[45] R. Raftogianis, C. Creveling, R. Weinshilboum, and J. Weisz, “Chapter 6: Estrogen metabolism by conjugation,” Journal of the National Cancer Institute Monographs, vol. 2000, no. 27, pp. 113–124, 2000.

[46] T. Sikosek and E. Bornberg-Bauer, “Evolution after and before gene duplication?,” in Evolution after Gene Duplication, pp. 105–131, Wiley-Blackwell, 2010.

[47] M. Matsuda and M. Sakaizumi, “Evolution of the sex-determining gene in the teleostean genus Oryzias,” General and Comparative Endocrinology, vol. 239, pp. 80–88, 2016.

[48] K. Higuchi, Y. Kazeto, Y. Ozaki et al., “Targeted mutagenesis of the ryanodine receptor by Platinum TALENs causes slow swimming behaviour in Pacific bluefin tuna (Thunnus orientalis),” Scientific Reports, vol. 9, no. 1, article 13871, 2019.

[49] C. Dong, P. Jiang, J. Zhang et al., “High-density linkage map and mapping for sex and growth-related traits of largemouth bass (Micropterus salmoides),” Frontiers in Genetics, vol. 10, p. 960, 2019.

[50] B. L. Fowler and V. P. Buonaccorsi, “Genomic characterization of sex-identification markers in Sebastes carnatus and Sebastes chrysomelas rockfishes,” Molecular Ecology, vol. 25, no. 10, pp. 2165–2175, 2016.

[51] B. Star, O. K. Torresen, A. J. Nederbragt, K. S. Jakobsen, C. Pampoulie, and S. Jentoft, “Genomic characterization of the Atlantic cod sex-locus,” Scientific Reports, vol. 6, no. 1, article 31235, 2016.