Original Article

A genome-wide association study, supported by a new chromosome-level genome assembly, suggests *sox2* as a main driver of the undifferentiatiated ZZ/ZW sex determination of turbot (*Scophthalmus maximus*)

Paulino Martínez [a,1], Diego Robledo [b,1], Xoana Taboada [c], Andrés Blanco [a], Michel Moser [d], Francesco Maroso [e], Miguel Hermida [a], Antonio Gómez-Tato [f], Blanca Álvarez-Blázquez [g], Santiago Cabaleiro [h], Francesc Piferrer [i], Carmen Bouza [a], Sigbjørn Lien [d], Ana M. Viñas [c,*]

[a] *Departament of Zoology, Genetics and Physical Anthropology, Faculty of Veterinary, Universidade de Santiago de Compostela, Campus de Lugo, 27002 Lugo, Spain*
[b] *The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, UK*
[c] *Departament of Zoology, Genetics and Physical Anthropology, Faculty of Biology, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[d] *Centre for Integrative Genetics, Department of Animal and Aquacultural Sciences, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway*
[e] *Department of Life Science and Biotechnology, University of Ferrara, 44121 Ferrara, Italy*
[f] *Departament of Mathematics, Faculty of Mathematics, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[g] *Instituto Español de Oceanografía (IEO), Centro Oceanográfico de Vigo, Cabo Estay-Canido, 36280 Vigo, Spain*
[h] *Cluster de Acuicultura de Galicia (Punta do Couso), Aguiño-Ribeira, 15695 A Coruña, Spain*
[i] *Institut de Ciències del Mar, Consejo Superior de Investigaciones Científicas (CSIC), 08003 Barcelona, Spain*

ARTICLE INFO

ABSTRACT

Background: Understanding sex determination (SD) across taxa is a major challenge for evolutionary biology. The new genomic tools are paving the way to identify genomic features underlying SD in fish, a group frequently showing limited sex chromosome differentiation and high SD evolutionary turnover. Turbot (*Scophthalmus maximus*) is a commercially important flatfish with an undifferentiated ZW/ZZ SD system and remarkable sexual dimorphism. Here we describe a new long-read turbot genome assembly used to disentangle the genetic architecture of turbot SD by combining genomics and classical genetics approaches.
Results: The new turbot genome assembly consists of 145 contigs (N50 = 22.9 Mb), 27 of them representing >95% of its estimated genome size. A genome wide association study (GWAS) identified a ~ 6.8 Mb region on chromosome 12 associated with sex in 69.4% of the 36 families analyzed. The highest associated markers flanked *sox2*, the only gene in the region showing differential expression between sexes before gonad differentiation. A single SNP showed consistent differences between Z and W chromosomes. The analysis of a broad sample of families suggested the presence of additional genetic and/or environmental factors on turbot SD.
Conclusions: The new chromosome-level turbot genome assembly, one of the most contiguous fish assemblies to date, facilitated the identification of *sox2* as a consistent candidate gene putatively driving SD in this species. This chromosome SD system barely showed any signs of differentiation, and other factors beyond the main QTL seem to control SD in a certain proportion of families.

1. Background

Sex is one of the most intriguing topics from evolutionary and developmental perspectives. Sex determination (SD) refers to the mechanism underlying the fate of the gonadal primordium at the initial stages of development responsible for the sex of a mature individual.

Theories on the evolution and architecture of SD have been largely influenced by studies on mammals, birds and *Drosophila*, characterized by conspicuous sex chromosome heteromorphisms associated with highly conserved SD genes. However, the increasing data from ectothermic vertebrates show a sharply different picture [1–3].

Fish display highly diverse chromosome SD systems, including models analogous to the XX/XY and ZW/ZZ systems of mammals and birds, but also others involving multiple sex chromosomes; however, in general chromosome heteromorphisms are rare in this vertebrate group [4]. Further, polygenic SD systems, mixed models involving both genetic and environmental cues, and an extensive within-species variation have been reported [2]. As in mammals and birds, major SD genes (SDg) driving sex determination have been described in fish, but unlike homeotherms, many diverse SD genes underlie fish SD: from classical transcription factors such as *dmy* [5] or *sox3* [6], to transforming growth factor ß-related genes such as *gsdf* [7] or *amh* [8,9] and its receptor *amhr2* [10], or genes related to the steroidogenic pathway such as *bcar1* [11] or *hsd17b1* [12]. Unexpected SD genes have also been documented, such as the salmonid interferon-related *sdY* [13]. Moreover, association studies in other fish have identified SD loci in non-orthologous genomic regions to other studied species, thus the current list is expected to be expanded in the near future [14–17]. In contrast to the highly canalized hierarchical systems and conserved developmental pathways of homeothermic vertebrates, the multiple options to recruit new SD genes suggest a more flexible hierarchy in fish [2].

The environment adds another layer of complexity and fish SD has been described as a continuum from strictly environmental to purely genetic, involving minor and major genetic factors [18]. It is difficult to encompass the broad diversity of fish SD within a conceptual genetic framework, but probably the closest would be a complex threshold trait usually involving major genes, where a quantitative genetics approach within a genomic framework would be recommended to disentangle its architecture [2,3]. To identify major genes and understand within-species variation, the new genomic and bioinformatic approaches that render long and reliable whole chromosome assemblies represent a powerful tool [19,20], essential to identify the tiny variation that can underlie SD in fish [2].

The turbot (*Scophthalmus maximus*) is an important farmed flatfish in Europe (~10,000 tons) and China (~60,000 tons, [21]). This species shows one of the highest growth rate sex dimorphisms among marine fish, females largely outgrowing males [22], and therefore obtaining all female populations is one of the main goals of the turbot industry. Turbot SD is compatible with a ZW/ZZ system driven by a major SD locus on LG5 [23,24], but no sex chromosome heteromorphisms have been detected [25,26]. Previous data support the small differentiation of the SD region in this species and the presence of a non-orthologous XX/XY SD system to that of turbot LG5 in a congeneric species (*S. rhombus*) [27]. A major SD region on LG5 and three additional suggestive sex-related quantitative trait loci (QTL) (LG6, LG8 and LG21) were identified through a genome scan with a moderate density genetic map [24,28]. Mapping and mining around those QTL using the turbot genome [29] led to the identification of candidate genes at the major and minor QTL, but none of them showed marker-association at the species-level [27]. Furthermore, temperature influence has also been demonstrated in turbot [23,30]. Despite this complexity, a microsatellite sex associated marker (SmaUSC-E30) at the major SD region is being used by industry to obtain all-female populations with acceptable outcomes [2].

Here, we constructed a new highly contiguous genome assembly for turbot, which was taken as a reference for disentangling the complex genetic architecture of SD in this species. We combined functional genomics, genome wide association studies (GWAS) and classical segregation analyses in a large sample of ~1200 turbot from 36 families with this new genome assembly to: i) identify the candidate SD gene and the potential causative mutation differentiating males and females; ii) characterize the ZW differential region and its genetic properties within

an evolutionary framework; iii) refine gonad development information around the critical window where gonad fate is established; and iv) assess interfamily variation underlying SD. We identified *sox2* as the most likely candidate for SD in turbot and explored to what extent other genetic or environmental factors influence SD in a species with little differentiation between the Z and W chromosomes.

2. Results

2.1. The new turbot genome assembly

Long-read Nanopore sequencing of the ZZ male rendered 1,219,548 reads with an average length of 19.060 bp (longest read: 429,199 bp; N50 = 32,402 bp) totalling 23.2 Gbp (40.8× coverage). Long reads were assembled into 145 contigs with N50 = 22.9 Mb and a total length of 560 Mb, very close to the total estimated turbot genome size (568 Mb, 98.6%; [29]). The correspondence between this new assembly and the previously published highly dense genetic map [31] was very high (99.3% of the 17,971 mapped markers were concordant between linkage groups and contigs; Supplemental File 1). A total of 24 contigs >10 Mb spanned 91.2% of the estimated 568 Mb genome length. Two pairs of these contigs constituted the two arms of the two metacentric chromosomes (C) of the turbot karyotype (C1 and C2; [31]), so these 24 contigs would represent the main body of the 22 turbot chromosomes [25]. The turbot genetic map allowed anchoring seven additional contigs (> 3 Mb), resulting in more than 95% of the turbot genome reliably assembled into 22 highly contiguous chromosomes: 9 chromosomes constituted by two contigs and 13 chromosomes by a single contig.

The contiguity of the new assembly largely improved that of the previously reported turbot genomes [29,32], as well as that of the assemblies of the two other flatfish species with publicly available genomes [33,34] (Table 1). The new assembly showed a very high correspondence with the last version of turbot genome assembly [29] using a highly dense genetic map [31] (Fig. 1), albeit small discrepancies were detected at the telomeres and centromeres of the chromosomes; the new assembly performed better at regions enriched in repetitive DNA and with lower recombination rate, where mapping accuracy is poorer. The completeness of the new genome assembly was assessed against the BUSCO Actinopterygii database; out of 4584 BUSCO groups, 4441 complete BUSCOs were reported (96.8%; 94.3% single-copy and 2.5% duplicated).

Orthology between the new turbot assembly and the two other flatfish with available reference genomes, Japanese flounder (*P. olivaceous*; [33]) and half-smooth tongue sole (*C. semilaevis*; [34]), was performed using the 22,751 turbot gene sequences annotated in the turbot genome [29] (Fig. 2). The results improved our previous comparison [31] and more refined details were obtained at some chromosomes where new minor syntenies were identified (Supplemental File 2). Interestingly, the Z and W chromosomes of tongue sole, which were previously matched to two different turbot chromosomes [31], here were orthologous to a single turbot chromosome (C9), as expected considering the limited genetic differentiation between them [34]. Also, as reported by Robledo et al. [21], the suggested SD chromosome of flounder (C24) matched to the turbot SD-bearing C12.

2.2. The turbot sex determining region

To locate and narrow the SD region of turbot we performed an association analysis in a farmed turbot population consisting of 1135 fish distributed across 36 full-sib families. The gonad of each animal was analyzed by histology to determine their phenotypic sex. The sex ratio was close to 1:1 (563 females and 592 males; χ^2 test, $P = 0.394$) and only two families showed unbalanced sex ratios (Table 2). All animals were genotyped for 18,214 SNPs using RAD sequencing. A GWAS revealed a major SD QTL at C12 (Fig. 3A; Supplemental File 3). The genome-wide significant markers spanned over a region of ~6.8 Mb of the new

Table 1
Comparative statistics of the new turbot (*S. maximus*) genome assembly.

	Scophthalmus maximus				*Cynoglossus semilaevis*[c]	*Paralichthys olivaceous*[d]
	This study	Figueras et al. (2016)[a]	Xu et al. (2020)[b]			
			Female	Male		
GC_content (%)	43.3	43.3	43.4	43.7	41.2	43.0
Contig/scaffold N50 (Mb)	22.9	4.27	6.2	5.9	0.9	3.9
Largest contig/scaffold (Mb)	31.7	19.0	19.9	19.5	4.7	13.3
Total length (Mb)	560	544	569	585	477	546
Contig/scaffold number	145	16,463	28,256	9724	31,181	7202

[a] [29].
[b] [32].
[c] [34].
[d] [33].

assembly (henceforth "the main SD region"), with the two most significant SNP markers placed ~420 kb away (Fig. 3B; Supplemental File 3). These two markers flanked *sox2*, and were in the vicinity of *dnajc19* and *fxr1*, the three candidate genes at the SD region reported by Taboada et al. (2014; Fig. 3C), encoding for a DNA-binding transcription factor, a heat shock protein and a RNA-binding protein, respectively.

2.2.1. Narrowing the SD region

To narrow the main SD region a segregation analysis was performed using the 13 most informative full-sib families (F7, F9, F10, F18, F19, F32, F42, F47, F52, F54, F60 and FX; Supplemental Files 4 and 5) among those showing the highest association between markers in C12 and sex (Table 2). Within each family, markers located within the main SD region heterozygous in the dam (ZW) and homozygous in the sire (ZZ) were used to track the Z and W chromosomes of the dam in offspring. Results were compatible with the ZW/ZZ system reported in turbot C12 (LG5, [24]). Recombinants between the Z and W chromosomes were identified in each family by the change in the genotypic linkage phase (Supplemental File 4) and each crossing-over was narrowed by the two adjacent informative markers (Supplemental File 4). Recombinants from right and left sides in each family delimited the minimum region where the sex determining gene (SDg) should be located in the W chromosome of the dam (Supplemental File 5). By integrating the information of the 13 selected families, the region harbouring the major SD locus was narrowed to 961 kb of the new turbot assembly (Fig. 4; Supplemental File 5).

2.2.2. Recombination across the SD region

The 121 more informative recombinants (those where the crossing-over point could be delimited in a region <300 kb) were used to study the recombination pattern between the maternal W and Z chromosomes (Supplemental File 6). Crossing-over was strongly impeded at the proximal region of C12 (0 to 2 Mb), supporting the positional effect of centromeric heterochromatin on the recombination of the adjacent genomic region (Fig. 5). Moreover, crossovers were distributed throughout the rest of the region evaluated, but the region where the suggested SD candidate genes (*sox2*, *dnajc19* and *fxr1*) are located showed evidence of recombination blockage, especially where *sox2* is located.

2.3. Genetic differences between Z and W chromosomes at the SD region

To identify differences between the Z and W chromosomes, we performed Nanopore sequencing of the DNA of a WW turbot female, which rendered 848,617 reads with an average length of 6.3 kb (longest read: 84.0 kb; N50 = 8529 bp) totalling 5.3 Gbp and representing 9.4× coverage of the turbot genome. The Nanopore-based assembly of the chromosome W SD region was polished using Illumina sequencing of WW females (see below). The SD region of the W assembly was split into

18 contigs with 60.3 kb on average and an N50 of 75.3 kb (maximum contig length: 139.9 kb), totalling 1085.1 kb. No gene duplications or structural reorganizations were detected when compared to the Z chromosome of the new turbot assembly.

Additionally, five WW females and five ZZ males were re-sequenced using Illumina 150 bp PE reads and separately aligned against both the W and Z chromosome assemblies. Reads covered the full length of the contiguous Z chromosome for comparison between WW and ZZ individuals to look for diagnostic SNPs differentiating both chromosomes. A large SNP dataset was identified and genotyped using SAMtools and the relative component of genetic differentiation (F_{ST}) between ZZ and WW individuals estimated. Those loci with highest F_{ST} were manually inspected in the alignment to confirm diagnostic differences between the Z and W chromosomes, i.e. showing fixed alternative allelic variants. One of the WW females was suspicion to be a genotypic male because the genotype pattern of those more divergent SNPs was nearly identical to ZZ males and therefore, was discarded for further analyses. Although a variety of SNPs, indels and duplications were detected in the region (Supplemental File 7), only two SNPs showed consistent differences between W and Z chromosomes (SNP.1 and SNP.2). These SNPs, fixed for alternative allelic variants in the Z and W chromosomes, were located between the genes *dnajc19* and *fxr1* (Supplemental File 8), surrounding the SmaUSC-E30 microsatellite used for sexing turbot [2].

A SNaPshot assay was designed to evaluate the consistency of this association at species level in a sample of 92 individuals, 46 ZZ males and 46 WW females, belonging to 11 unrelated families of a turbot breeding program. Most ZZ males showed an AAGG genotype, while most WW females were GGAA supporting an AG haplotype for the Z chromosome and a GA haplotype for the W chromosome (Supplemental File 9; χ^2, $P = 0$), consistent with the Illumina sequencing results. However, 14.1% individuals showed genotype and phenotype discordances, 7.6% of them showing a ZW constitution, potentially occurring by recombination between informative markers and the SDg when obtaining the WW individuals (it takes 3 generations to produce WW fish; [2]). All in all, SNP.1 showed the best association, since for SNP.2 five WW females of a single family showed a GGAG genotype instead of GGAA, not congruent with their sex.

2.4. Functional screening of the SD region: sox2 as the putative main driver of turbot SD

2.4.1. Identification of transcripts at the SD region expressed during the initial stages of sex differentiation

To identify the genes located in the main SD region of turbot (965 kb) that are expressed during the initial steps of sex differentiation, ribo-depleted RNA libraries obtained from a superpool of larvae (3 pools per time, 5 to 30 dph), a pool of male gonads and a pool of female gonads (3 individuals per sex and time, 35 to 90 dph) were sequenced. A total of 20 genes were expressed along that period in the delimited 961.4 kb region (Fig. 6A; Supplemental File 10). Kininogen-1 (*kng-1*),
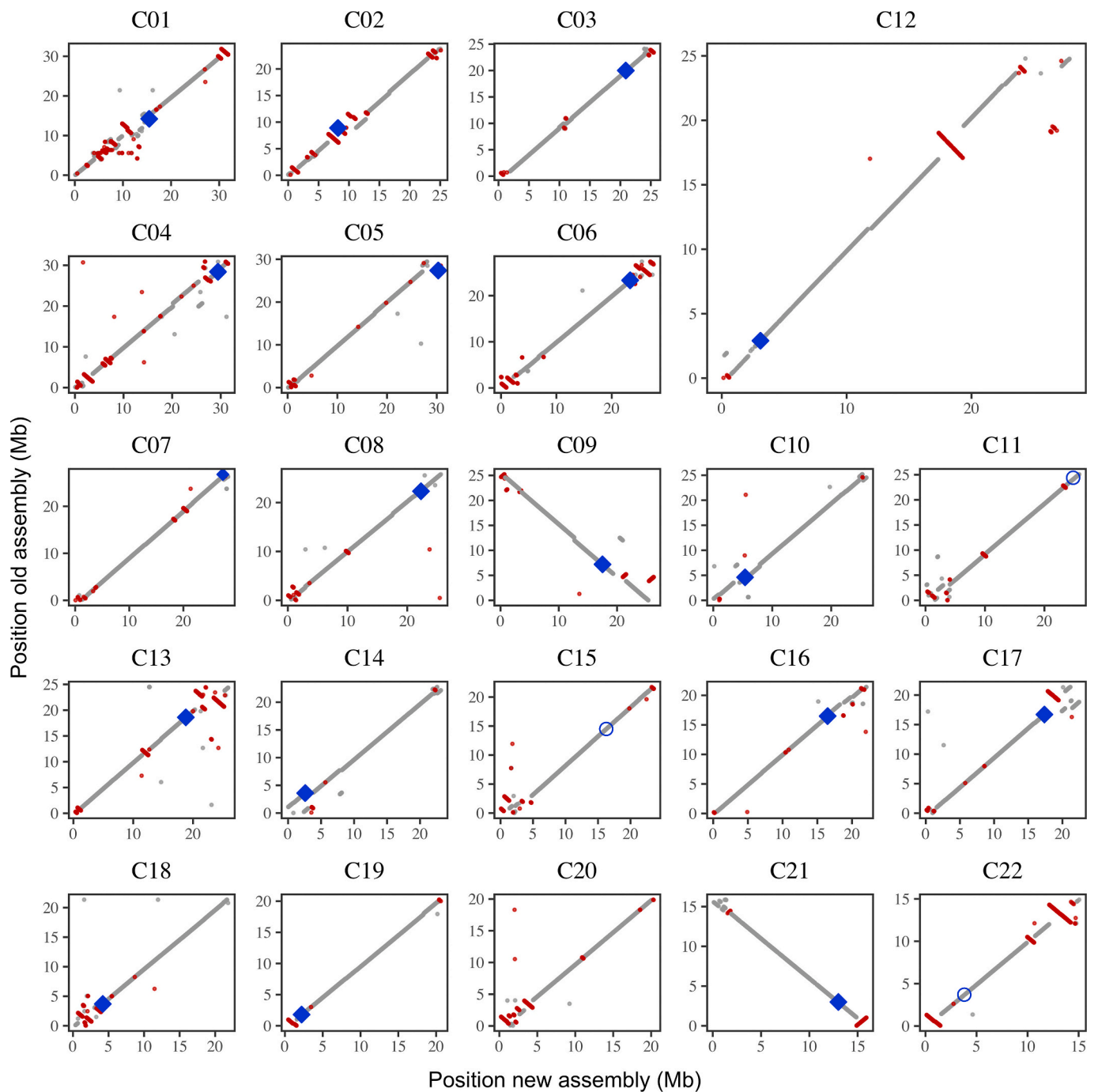
Fig. 1. Alignment of the new turbot (*S. maximus*) genome assembly with that by Maroso et al. (2018) using LASTZ. Blue squares show the markers close to centromeres (< 5 cM); blue empty circles show markers very far away from centromeres (> 40 cM) and red lines missalignments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

complement factor H (*cfh*) and Fragile X mental retardation syndrome-related protein 1 (*fxr1*) were the highest expressed genes, while the lowest one was *sox2ot* (see below). None of the genes in the region showed signs of differential expression between the male and female pools obtained from samples encompassing the whole critical period of sex determination (between 35 and 90 dph; Supplemental File 10), which is not surprising if, as in other species, the critical development window where differential expression between sexes occurs is narrow. Among the three reported candidate genes in the region related to gonad differentiation or reproduction, *dnajc19, fxr1* and *sox2* [27], only the two latter were expressed. Additionally, a long non-coding RNA (lncRNA), annotated as *sox2* overlapping transcript (*sox2ot*) and

encompassing 78,309 bp, was found in the region where *sox2* is located. This gene, conserved across vertebrates [35], includes *sox2* in one of its introns. Within this lncRNA, a splicing variant corresponding to a short non-coding RNA (ncRNA) transcript showed a faint expression along the inspected period both in males and females. Interestingly, a ~ 250 bp region of this ncRNA was complementary to a region of the *sox2* mRNA 3' UTR, thus suggesting a potential role as negative regulator of *sox2*.

2.4.2. qPCR of candidate genes along the critical period of gonad development

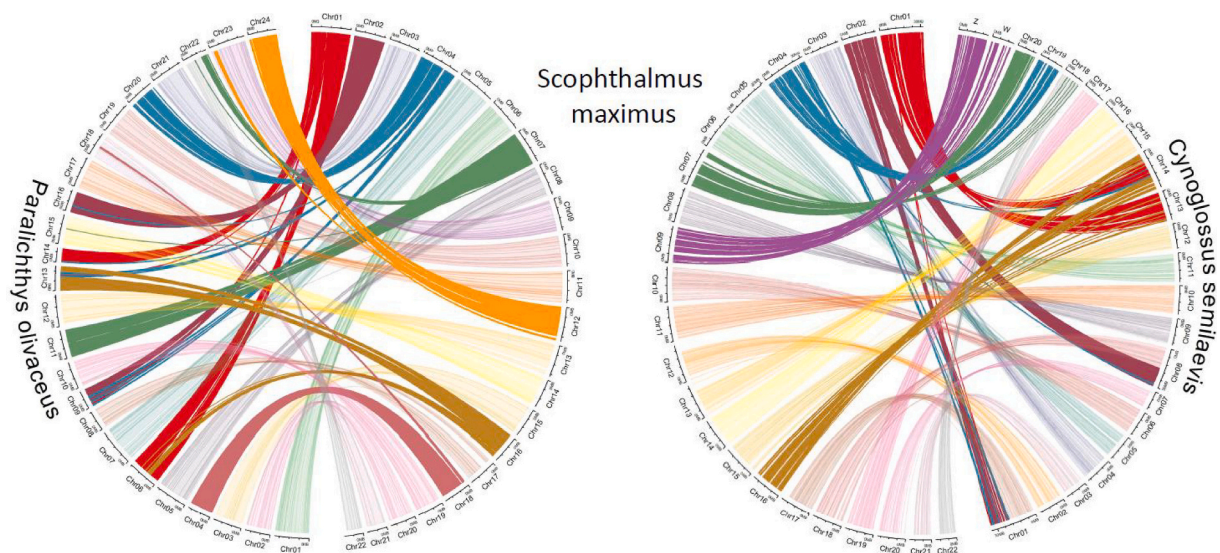Gene expression of candidate genes *fxr1, dnajc19, sox2* and *sox2ot*

Fig. 2. Orthology between turbot (*S. maximus*) vs. Japanese flounder (*P. olivaceous*) and half-smooth tongue sole (*C. semilaevis*) using the turbot genome assembly. Highlighted in darker colors those orthologies splitted in different chromosomes.
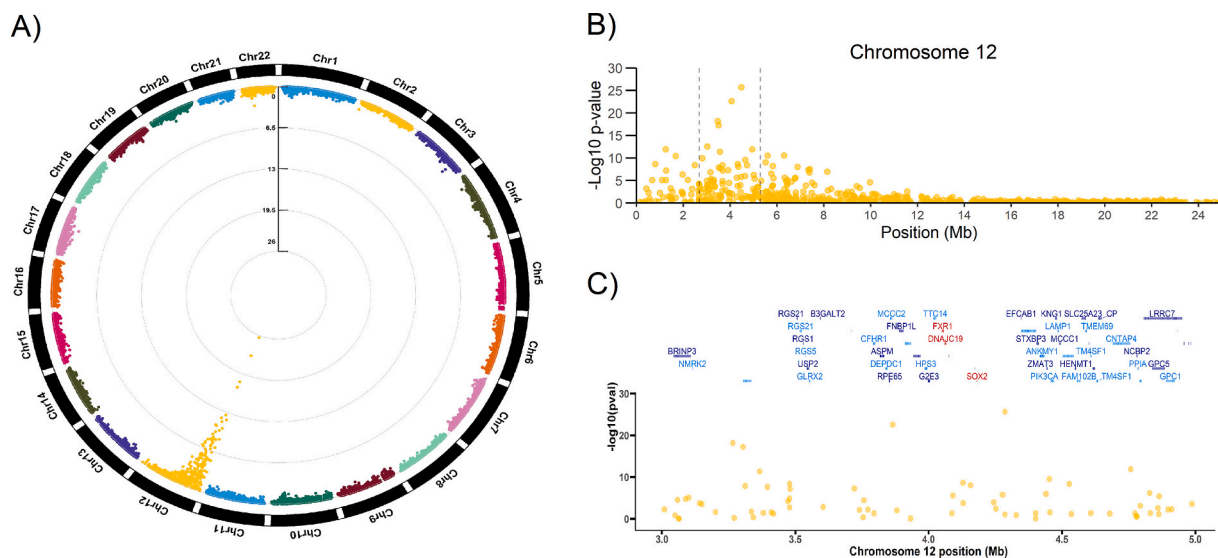

Fig. 3. Genome wide association study (GWAS) to look for association in turbot (*S. maximus*) between 18,271 SNPs and sex. A progressive zoom is shown from the whole genome (A) to chromosome 12 (B), and specifically, to a narrowed SD region on C12 (C) where SD candidate genes are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was individually evaluated by qPCR in males and females during the early stages of gonad development, specifically between 30 and 90 dph, i.e. from weaning until the time where gonad differentiation is established at genetic level (up-regulation of *amh* and *cyp19a1a* in males and females, respectively; Supplemental File 11; [30]). This is the period where gonads can be dissected with confidence for gene expression evaluation, and likely when SD occurs in turbot, after metamorphosis but before gonadal differentiation. No expression differences were detected for *fxr1*, *dnajc19* and *sox2ot* between males and females, but a significant increase of *sox2* expression was detected in females at 55 dph using two different sets of primers (*t*-test: $P = 0.011$ and 0.018 for 5′ and 3' UTR amplicons respectively), and even before at 50 dpf, although less consistent (only in the 3' UTR amplicon; $P < 0.05$). Significance was even higher when the 50–55 dph window information was merged ($P = 0.002$ and 0.001 for 5′ and 3' amplicons, respectively), considering the observed trend and the narrow development window evaluated (Fig. 6B; Supplemental File 12).

2.5. Secondary factors influencing turbot sex determination

Our analyses, as previously reported [24,28], suggest the existence of secondary factors influencing sex determination in turbot. A GWAS was performed separately in each of the 36 families analyzed (Fig. 7) and two families showed suggestive signals of association at C22 (F19 and F36). This is consistent with the association at chromosome-wide level of one SNP in the same chromosome in the whole dataset ($P < 0.0000865$; Supplemental File 3). Other families also showed the highest associated marker in other chromosomes (Supplemental File 13), although none of them significant after multiple test correction. These associations at other chromosomes involved several SNPs with P-values above the background, which considering the low sample size managed (23–38 individuals per family) suggests that other genetic factors might be influencing SD. Further, other families such as F12, F14 and F56 did not show any signal of association suggesting that environmental or stochastic factors may exist. Moreover, a highly significant trend (χ^2 test;
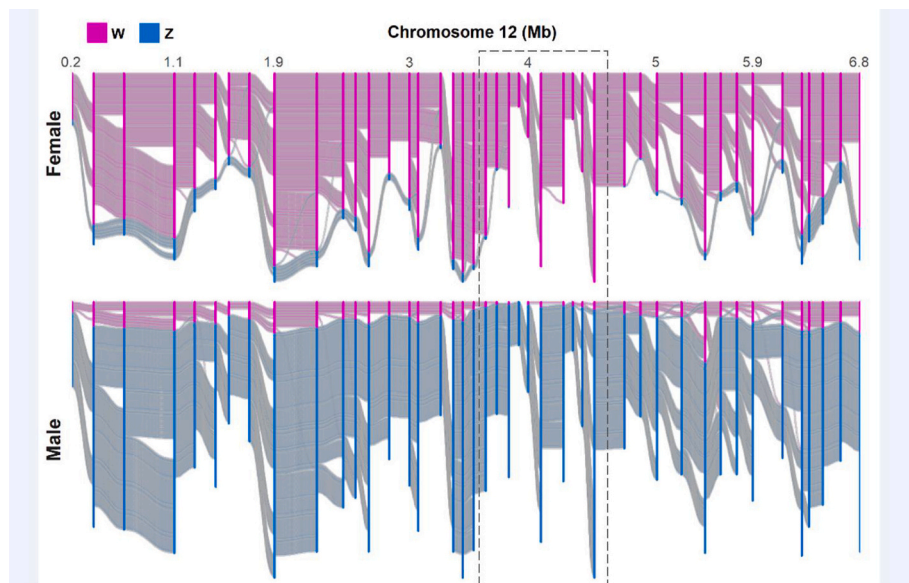
Fig. 4. Segregation of Z and W maternal chromosomes along the turbot (*S. maximus*) SD region. The transmission of the maternal Z and W chromosomes to the offspring in the 13 selected full-sib families is shown along 6.8 Mb of the turbot main SD region. Offspring were split according to their phenotypic sex in "Female" and "Male". A total of 238 markers located in the SD region heterozygous in the dam (ZW) and homozygous in the sire (ZZ) were used to track the Z and W chromosomes from the dam to the male and female offspring, respectively, and to identify recombinants between them (Suppl. file 6). Nearby SNP markers informative in different families were collapsed to 44 single physical positions represented as vertical bars in the figure to facilitate visual representation. The colors of the bars refer to the W chromosome (magenta) of female offspring or the Z chromosome (blue) of the male offspring inherited from the dam. The haplotype of each offspring can be tracked using the horizontal lines and recombination events are represented by changes from "W" genotype to "Z", or viceversa. Some gaps are observed for particular individuals due to non-informative SNPs at some bars. The dotted box represents a conservative estimation where the SDg should be located based on the consistency between genotypes and phenotypes (see Suppl. file 5). Note how in this region, where *sox2* is located (~4 Mb bar), the W and Z chromosome constitution is mostly retained in females and males, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
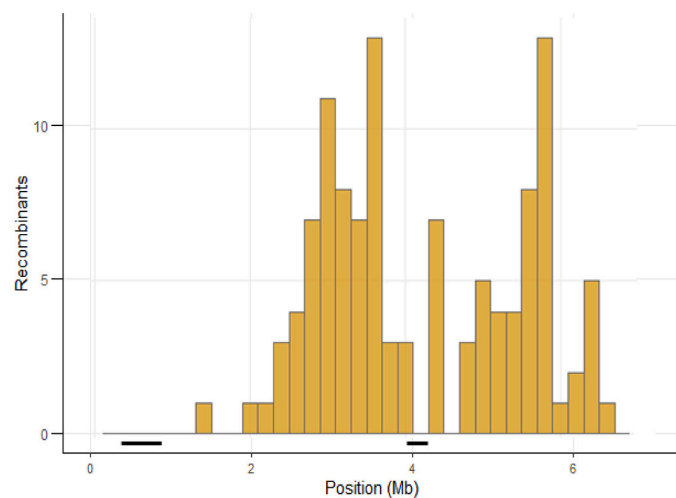
Fig. 5. Frequency histogram of crossing-overs between Z and W chromosomes across the proximal región of turbot (*S. maximus*) chromosome 12. The bars at the bottom show from left to right the estimated position of the centromere (Martínez et al., 2008) and of the small region where the SD candidate genes (*fxr1*, *dnajc19* and *sox2*) are located.

$P < 10^{-5}$) of genetic females developing as phenotypic males was detected in those families where sex was significantly associated with C12 (59 ♀ → ♂ vs 13 ♂ → ♀; Table 2).

3. Discussion

Fish exhibit a remarkable gonad development plasticity and a high evolutionary SD turnover, which poses important challenges for disciplines such as evolutionary and development biology. In this study, we constructed a chromosome-level turbot genome assembly to disentangle the architecture of sex determination in turbot. Combining this new reference genome with genomics and classical genetics approaches in a broad sample of individuals and families painted a comprehensive picture of SD in this species, which contributes to understand the genetic basis and evolution of SD in fish [36].

3.1. A new reliable turbot genome assembly

The new long-read based assembly of the turbot genome is among the best fish genomes assembled to date [37], and represents the first highly contiguous flatfish genome with 31 contigs constituting the 22 chromosomes and representing >95% of the total genome size. This new assembly showed a very good correspondence with our previous turbot assembly [31], but with much lower gapping and improved assembly at those regions enriched in repetitive elements, such as centromeres and telomeres. Comparative mapping with other flatfish improved the orthologous relationships previously reported [21,31,32,38,39], allowing the robust integration of genomic data of the three most important aquaculture flatfish, pertaining to three different families within pleuronectiformes. Furthermore, our study enabled us to refine the orthology relationships for the Z and W chromosomes of *C. semilaevis*, now orthologous to a single turbot autosome (C9). The new turbot assembly represents a very useful tool for the integration of all previous genomic information, which will contribute to improve worldwide production of this important aquaculture species. This assembly also constitutes a high-quality reference genome for functional annotation and comparative genomics in fish and vertebrates, including evolutionary studies within Pleuronectiformes to ascertain the specific adaptations to a demersal lifestyle [29,33].

3.2. The main sex determining gene of turbot

The extensive genome-wide screening in combination with a classical segregation analysis confirmed that turbot SD is controlled mainly by a small region of less than 1 Mb at the proximal end of C12 that follows a ZW/ZZ pattern of inheritance [23,24,40]. Evidence suggests that *sox2* might be the main driver of sex determination in this species: i) this gene is surrounded by the two most significant SNPs detected by GWAS and located within the genomic region delimited by the

Table 2

Sex proportions and its association with markers at the C12 SD region of turbot (*S. maximus*).

Family	N	gen/phen discord. (%)	P-value association	RF (%)	% females	♀ → ♂	♂ → ♀
F7	36	8.6	9.07E-07	27.8	50.0	1	2
F8	32	19.4	0.000235	31.3	31.3	6	0
F9	28	0.0	2.03E-07	17.9	42.2	0	0
F10	37	5.7	1.87E-07	29.7	59.5	1	1
F11	33	19.4	0.000235	21.2	33.3	6	0
F13	29	17.9	0.000241	20.7	34.5	5	0
F15	30	6.7	0.000003	16.7	46.7	2	0
F16	32	10.0	0.000011	25.0	46.9	2	1
F18	33	0.0	1.54E-08	12.1	36.4	0	0
F19	38	2.7	2.36E-08	18.4	42.1	1	0
F20	34	6.5	0.000012	20.6	41.2	2	1
F32	38	5.5	1.20E-07	21.1	29.0	2	0
F39	36	21.2	0.000485	25.0	39.9	6	1
F42	32	6.3	0.000001	25.0	34.4	2	0
F43	27	12.0	0.000095	37.0	51.9	3	0
F45	31	16.7	0.000108	16.1	67.8	0	5
F47	30	3.7	0.000002	35.7	32.2	1	0
F48	31	6.7	0.000002	32.3	41.9	2	0
F52	23	0.0	0.000002	26.1	52.2	0	0
F54	30	7.1	0.000005	16.7	43.3	2	0
F55	30	17.9	0.000465	20.0	60.0	4	1
F58	34	11.8	0.000010	11.8	35.3	3	1
F60	35	0.0	3.30E-08	31.4	51.4	0	0
F63	30	23.3	0.000465	23.3	33.3	7	0
FX	28	3.7	0.000001	17.9	42.9	1	0
F12	36	–	ns	19.4	94.4*		
F14	29	–	ns	31.0	58.6		
F22	36	–	ns	2.8	47.2		
F38	24	–	ns	20.8	45.8		
F56	30	–	ns	20.0	60.0		
F28	33	–	ns	18.2	69.7		
F30	34	–	ns	23.5	58.8		
F33	37	–	ns	21.6	83.8		
F36	29	–	ns	17.2	38.7		
F53	27	–	ns	25.9	14.8*		
F59	23	–	ns	20.0	73.1		
mean/total	1135	9.3		22.3	47.9	59	13
range	23–38	0–23.3		2.8–37.0	14.8–94.4	0–6	0–5

The analysis was carried out across 36 turbot (*S. maximus*) families. **N**: offspring per family; **gen/phen discord.**: % of discordances between genetic and phenotypic sex at the SD region (see Supplemental File 4); **P-value** of the gen-phen association (after Bonferroni correction ($P < 0.0014$); **RF**: recombination frequency between the Z and W chromosomes; ♀ → ♂: number of genetic females that are phenotypic males; ♂ → ♀ number of genetic males that are phenotypic females. * unbalanced sex-ratio families ($P < 0.05$).

segmentation analysis; and ii) it is the only gene within the delimited SD region that shows differential expression between males and females between 5 and 90 dph (50–55 dph) and, importantly, prior to the first signals of germ cell proliferation (up-regulation of *gsdf1* and *vasa* at 60–65 dph; [30]). This gene had been previously discarded as the SD gene of turbot because no SNPs showing alternative allelic variants fixed in Z and W chromosomes could be identified in its coding region in a large sample of wild males and females [27]. Although this still holds, data suggest that the mutation responsible for differential expression of *sox2* between sexes might be located on a regulatory element. We identified a putative diagnostic variant between Z and W chromosomes in our study (SNP.1). Although it is located far from *sox2* (149 kb) it cannot be discarded as part of a regulatory element, since *sox2* enhancers have been identified within a 200 kb region around this gene [41]. Further, this marker lies very close to SmaUSC-E30 (~ 1.5 kb) the microsatellite marker previously used to identify genetic sex in families. However, its diagnostic condition or functional implication on SD is difficult to establish considering the implication of other genetic and environmental factors on SD and the putative recombination between SmaUSC-E30 and the SDg when obtaining the WW females. Another relevant disclosed element is the lncRNA *sox2ot*, a gene regulating *sox2* [35,39,42], which spans across most of the SD region and includes *sox2* in one of its introns. Interestingly, a splicing variant of *sox2ot* (~250 bp) showed full complementarity with a region of the 3' UTR of *sox2*, suggesting a potential negative post-transcriptional mechanism controlling its activity. Although the functional annotation of lncRNAs is still poor,

they have been involved in the regulation of genes related to reproduction and particularly to SD in different species [43], such as *Drosophila* [44] and the crustacean *Daphnia magna* [45]. The role of this splicing variant of *sox2ot* should deserve further investigation. Recently, *sox2* has been related to gonad differentiation in the Zhikong scallop *Chlamys farreri* [46] and suggested as the candidate SD gene, the black tiger shrimp *Penaeus monodon* [47], two important aquaculture species. All in all, if, as our data support, *sox2* is the main SD driver in turbot, this would be the first case, to our knowledge, where an ovary triggering gene from a default testis development has been reported in fish. Further work should be done to fully understand the regulation of *sox2* and specifically the putative role of *sox2ot*, and for achieving that goal, knocking genes through CRISPR methodologies, despite their difficulties on marine species because of their hard chorion, would be an invaluable tool.

3.3. Turbot SD as a complex trait

Our data support the presence of additional secondary factors influencing turbot SD in accordance with previous information [23,38]. Despite the major SD region showed a significant association between sex and markers in nearly 70% of the families, other genetic and/or environmental factors seem to interact with the main locus to determine the sex of each fish. In particular an association at C22 was significant at chromosome-level with the whole dataset and showed P-values close to significance in two families ($P < 0.001$) despite the low within-family

Fig. 7. Genome wide association study of sex at the 36 turbot (*S. maximus*) families.

sample size used. Other sex-related QTL had been previously reported in turbot, always as secondary factors and only in a single family [24,28]. None of the previously reported secondary QTL were identified in the broad family sample analyzed in this study, and those here detected seemed to influence sex in a few families. These observations suggest either the presence of rare allelic variants affecting sex in secondary QTL or, more likely, complex genetic-environment interactions altering the canonical SD hierarchy. We identified eleven families where no signals of genomic association were detected, suggesting other factors driving sex. Environmental factors influencing SD have been described in fish, where high rearing temperatures tend to induce masculinization [48]. In the case of turbot, the influence of temperature does not seem to work in a simple way [23], although Robledo et al. [30] showed a significant trend for female biased ratios at low temperatures. The influence of other environmental cues on SD has not been studied in turbot, so we cannot rule out that other factors might be also playing a role on turbot

sex. Anyway, it should be noted that in our study, individuals from different families cohabited in the same tank following a design thought to randomize environmental effects among families [49,50]. thus suggesting, some kind of interaction between genetic and environmental factors underlying sex determination in turbot. Globally, turbot SD resembles a complex trait, where environmental and major and minor genetic factors and their interactions control its sex.

3.4. SD in other flatfish and teleosts: some insights from turbot

Despite the increasing number of genomes sequenced and assembled, the sequences of the Y and W sex chromosomes are unknown in most species; the presence of repetitive elements represents a handicap for their assembly and they usually require at least twice the depth of a homogametic genome [51]. In fish, the early evolutionary stage of most sex chromosomes and the viability of WW or YY individual [1] facilitate
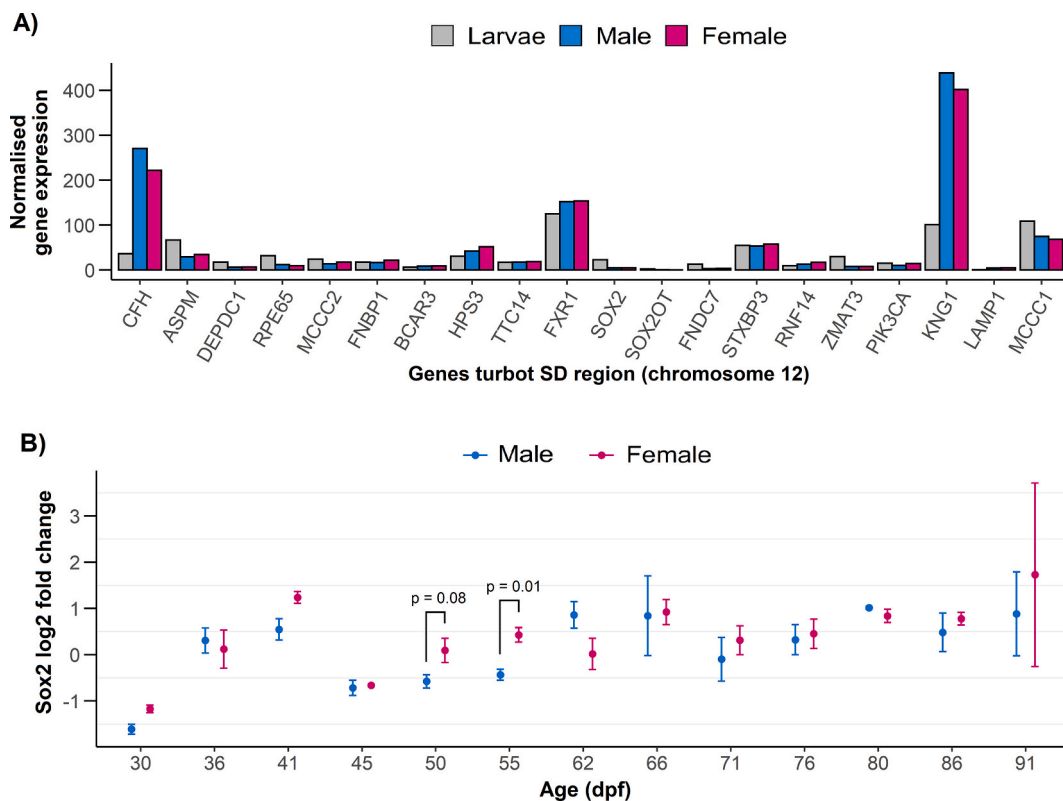
Fig. 6. Normalized expression of transcripts detected in the SD region of turbot (*S. maximus*) across the critical gonad differentiation period. A) transcripts identified in pools of larvae, females and males sexed from 5 to 90 dpf; B) *Sox2* qPCR in males and females from 30 to 90 dpf.

this task, but at the same time the small differences between sex chromosomes make extremely difficult to find diagnostic differences between them. Finding the causative variants underlying major SD QTL becomes even more complicated in turbot due to the incomplete penetrance of the SDg. Sequencing the turbot W and Z chromosomes allowed us to confirm the low differentiation between sex chromosomes in this species, as previously suggested by Taboada et al. [27]. These authors proposed a young evolutionary origin of the turbot SD system considering that the brill, a congeneric species (*S. rhombus*), showed sex-association with *sox9*, a key gene related to testis differentiation located at LG21, a non-orthologous chromosome to turbot C12. Although a putative diagnostic SNP between Z and W chromosomes was detected in our study, as outlined above, no conclusive genetic differences supporting a SD causal variant were identified. In this regard, the turbot's sex chromosome pairs would be at an initial differentiation stage close to that of the pufferfish (*Takifugu rubripes*; [10]), where the only difference between X and Y chromosomes was a single differential SNP at the promoter of the *amh* receptor. Very recently, a small differentiation between sex chromosomes has been reported in *Esox lucius*, where a small male-specific insertion containing the SD gene has been documented [9]. However, in this study the original duplication of *amh* was estimated around 40 MYA, which suggests caution when considering a young SD system only based on the differentiation between sex chromosomes. Although examples of highly differentiated SD systems such as of the genus *Characidium* and *Eigenmannia* exist in fish [4], the differential region in most species where the main SD gene has been identified spans less than 1 Mb [2]. Theory postulates that the evolution from a homomorphic to a heteromorphic sex chromosome pair begins with the suppression of recombination to maintain favorable haplotype combinations avoiding sexual conflict [52–54]. Our data suggests that recombination is suppressed around *sox2* and the lncRNA *sox2ot*. Taboada et al. [27] also suggested a transition from an ancestral XY to a new ZW system in turbot, highlighting an incomplete dominance of the

W chromosome as evidenced by the higher amount of ZW males, an observation also supported with much more data in this study.

The increasing genomic resources within flatfish enabled us to carry out a comparative genomic approach on the genetic basis and diversification of SD in this interesting fish group taking the new turbot assembly as reference (see Supplemental File 14). As indicated above, the SD region and candidate genes of turbot (*sox2*, C12) and smooth tongue sole (*dmrt1*, Z chromosome [34]) are not orthologous. Despite both species share a ZZ/ZW system, the Z chromosome of smooth tongue sole is orthologous to turbot C9, not related to sex determination. However, smooth tongue sole and Pacific halibut (*Hippoglossus stenolepis*) might share the same SD system since the LG7 of Pacific halibut, where two sex-linked markers are located [38], is syntenic to the smooth tongue sole Z chromosome, and one of these sex-linked markers (Hs10183) is located 3 Mbp away of *dmrt1* in both assemblies. On the other hand, the SD-bearing C12 chromosomes of turbot (C12) and that of Japanese flounder (*P. olivaceous*) (chromosome 24) are orthologous (see Fig. 2); however, the information from Japanese flounder should be taken with caution, since it proceeds from a single microsatellite sequence deposited in a patent for sexing in this species [21] (Supplemental File 14). Anyway, the SD region of Japanese flounder would be located far away from that of turbot (> 14 Mb), in the distal region of C12 (19,6 Mb) close to the *amh* gene in the new turbot genome assembly. Finally, two strongly SD-associated SNPs found at LG13 of Atlantic halibut (*H. hippoglossus*), a flatfish with an XX/XY SD system [39], are syntenic to turbot C16. All these data support the heterogeneity of SD systems in flatfish, including non-orthologus ZW/ZZ and XX/XY systems, as previously documented by Luckemback et al. (2011), but also highlight the novelty of *sox2* as a new putative SDg in a context dominated by the recurrent recruitment of some particular genes such those pertaining to *dmrt* or *amh* related gene families.

3.5. Applications for turbot farming

One of the goals of this work was to find a reliable sex marker for the turbot industry. In turbot no morphological dimorphism exists between sexes prior to sexual maturity (24 months), and therefore sex cannot be identified at juvenile stages [55]. The marker USC-E30, a microsatellite used as sexing tool under a Spanish patent (Ref. number: 2354343), although useful and applied by the industry to obtain all female populations, has the drawback of the high and recurrent mutation rate of microsatellites, and therefore requires establishing associations within each family. In this sense, the SNP.1 marker associated to Z and W chromosomes in our study will provide a more confident and straightforward tool for precocious sexing in turbot breeding programs.

4. Potential implications

The study of turbot sex determination (SD) has added new information to the evolutionary theory on this intriguing topic: i) it is among those showing the least differentiation between Z and W chromosomes reported in fish, and only a suggestive differential SNP could be identified between Z and W chromosomes; ii) recombination between Z and W chromosomes seems to be suppressed at a short stretch involving sox2 and the *sox2ot* long non-coding RNA; iii) *sox2*, a gene of the *sox* family of transcription factors, is the most consistent candidate for SD in turbot and could be added, if confirmed, to the growing list of fish SD genes; iv) the lncRNA sox2ot may be involved in the regulation of *sox2*, thus adding new data to the implication of lncRNA on SD; v) other genomic regions seem to be involved in SD in turbot, although these secondary factors have a limited and sporadic influence; vi) environmental factors seem to also play a role on SD in a number of families, as previously suggested. Nonetheless, the major SD region at C12 seems to prevail at the population level, showing association with sex in over 70% of the families studied. Globally, SD in turbot is similar to a complex trait, albeit with a major gene explaining most of the variance. Intra-specific variation has also been observed in many species possessing a master SD gene, but the genetic and environmental influence does not seem so remarkable, which could be related to a recent evolutionary origin. The identification of a differential SNP between Z and W chromosomes will facilitate obtaining all-female populations in turbot aquaculture.

5. Materials and methods

5.1. Sampling

Different samples were used for the different methodological approaches used to build a new turbot genome assembly and study SD in this species: i) One mature turbot male coming from a turbot farm was used for long-read Oxford Nanopore sequencing (40×) to assemble the turbot genome; one WW mature female was also sequenced with Nanopore technology at 10× coverage to investigate differences between the Z and W chromosomes at the SD region. Next-seq Illumina sequencing was used to polish the Nanopore assemblies and explore the differences between Z and W chromosomes in more detail and in a broader sample; for this, five ZZ males and five WW females were sequenced using 150 bp paired-end (PE) reads at 20× coverage; WW females were obtained from unrelated families in the breeding program of a turbot company using mainly SmaUSC-E30 microsatellite marker to track the W chromosome across generations as described by Martínez et al. [2]. ii) 1160 fishes (between 133 and 217 days post hatching, dph) pertaining to 36 full-sib families and their 47 parents (total 1207 fishes) were confidently sexed by histology and genotyped for 18,214 SNPs using 2bRAD-Seq to look for association between sex and markers at population and family level through genome wide association studies (GWAS). This biological material was also used for a segregation analysis to narrow the SD region and establish a crossing-over map over its entire length. These 1181 fishes came from an experiment for resistance

to scuticociliatosis performed at the CETGA (Centro Tecnológico Gallego de Acuicultura) hatchery facilities within the FISHBOOST project (KBBE-2013-7-613,611); iii) Larvae/fry from 5 to 90 dph (18 sampling points) were used for RNA-seq analysis (six pools of larvae/six individual fry per sampling point) to identify expressed transcripts in the SD region from hatching until the gonadal fate is established (90 dph; [30]). These samples were obtained at the IEO (Instituto Español de Oceanografía) of Vigo using 10 families of unrelated parents; qPCR gene expression analysis of candidate SD genes along the critical period of turbot gonad differentiation was performed individually in 13 of those sampling points (from 35 to 90 dph), using three males and three females per point (sexed with the SmaUSC-E30 marker [2]), when gonads could be confidently dissected after metamorphosis.

5.2. Sequencing and assembly of the turbot genome

5.2.1. Nanopore sequencing

DNA from one ZZ male and one WW female was extracted from fresh blood using the MagAttract HMW DNA Kit to isolate genomic DNA >150 kb. Following sample lysis, the DNA was bound to the surface of magnetic beads; during the washing steps, contaminants and PCR inhibitors were removed and pure high-molecular-weight DNA was eluted in Buffer AE.

Prior to long-read sequencing, high-molecular-weight DNA from the ZZ male was treated with the Circulomics Short Read Eliminator kit (Circulomics) to remove unwanted short fragments. Then, PCR adapters (20 µl) from Oxford Nanopore Technology (ONT) (SQK-LSK109 Ligation Sequencing Kit 1D) were ligated. Prepared library was loaded onto two PromethION flow cells (FLO-PRO002) and sequenced on a PromethION Beta machine. Flow cells were run for a total of 8 and 13.6 h. Flowcells were treated with DNAse I, reloaded with library DNA and run again multiple times to maximize sequencing output.

Nanopore reads were base-called using Guppy, version 3.0.5 (Oxford Nanopore, MinKNOW v.19.05.1) using "High Accuracy" flip-flop model. Reads were trimmed for 50 bp from the start and quality-filtered for average PHRED quality >7 and minimum length of 4000 bp with fastp ([56] v 0.19.5), options "–disable trim_poly_g –disable_adapter_trimming -q 7 -l 4000 -f 50 ".

Long read runs from the WW female purified DNA were generated at CNAG (Centre Nacional de Análisis Genómico, Barcelona) following the protocol with the Rapid Sequencing kit SQK-RAD002 (ONT, Oxford, UK). Briefly, approximately 200 ng of purified DNA was tagmented for 1 min at 75 °C with the Fragmentation Mix (ONT, Oxford, UK). The Rapid Adapters (ONT, Oxford, UK) were added along with Blunt/TA Ligase Master Mix (NEB, Beverly, MA) and incubated for 30 min at room temperature. The resulting library was combined with Running Buffer with Fuel (ONT, Oxford, UK) and Library Loading Beads (ONT, Oxford, UK) and loaded onto a primed R9.4 SpotOn Flow cell (FLO-MIN106). Sequencing and initial base calling was performed with a MinION Mk1B MinKNOW v1.7.10 software package running for 48 h.

5.2.2. Illumina sequencing

Genomic DNA of five ZZ males and five WW females was extracted from fin clips using SSTNE buffer (a TNE buffer modified by adding spermidine and spermine) and a standard NaCl isopropanol precipitation [57]. Barcoded libraries of 150 bp PE were constructed for each individual and subsequently sequenced in a Next-seq Illumina machine at an estimated coverage of 20× per individual. The quality of the sequencing was assessed using FastQC v.0.11.5 (http://www.bioin formatics. babraham.ac.uk/projects/fastqc/). Quality filtering and removal of residual adaptor sequences was conducted on read pairs using Trimmomatic v.0.36 [58]. Specifically, Illumina adaptors were clipped from the reads, leading and trailing bases with a Phred score < 20 were removed, and the read trimmed if the sliding window average Phred score over four bases was <20. Only reads where both pairs were longer than 36 bp post-filtering were retained. Filtered reads were

mapped to the new turbot assembly using Burrows-Wheeler aligner v.0.7.8 BWA-MEM algorithm [59]. Pileup files describing the base-pair information at each genomic position were generated from the alignment files using the mpileup function of samtools v1.6 [60], discarding those aligned reads with a mapping quality <30 and those bases with a Phred score < 30.

5.2.3. ZZ male genome assembly

Reads of the ZZ male sample were assembled with the Flye progam ([61] v2.5.0) using different overlap sizes (−m) of 10,000, 12,000 and 20,000. The three genome assemblies were merged with quickmerge ([62] v0.3) using default settings. The resulting quickmerge assembly was polished in two stages using nanopore reads and Illumina reads. First stage used MarginPolish-HELEN pipeline [63] with nanopore read alignments. The output was polished in a second stage with pilon [64] using aligned Illumina reads as input. Using genetic marker information from Maroso et al. [31], the polished sequences were scaffolded with CHROMONOMER [65] into 22 pseudochromosomes, resembling the *n* = 22 karyotype of *S. maximus*.

5.2.4. WW female genome assembly

A de novo assembly was done in the WW female using those reads matching at the delimited SD region (see segregation analysis) of chromosome 12 (C12) where the major SD region is located (LG5; [24,31]). For selection of reads, both the new ZZ assembly from this work and the previous version of the turbot genome [31] were taken as reference. Assembly was performed with the selected reads following the instructions provided by Oxford Nanopore (https://nanoporetech.com/resource-centre/hybrid-assembly-pipeline-released-using-canu-racon-and-pilon); briefly, assembly of Nanopore reads was done using Flye assembler ([61] v2.5.0). Resulting contigs of the WW genome was polished with the module Racon [66], and finally, assemblies were curated using the Pilon tool [64] by adding the Illumina reads.

Gene annotation.

Gene models from genome GCA_003186165.1 were transferred to the new ZZ male genome assembly with gmap ([67] gmap-2020-03-12) using parameters -f gff3_gene -O -n 0.

5.2.5. Confidence of the new turbot assembly: integrating previous genomic information

To track gene space completeness, BUSCO ([68] v3.0.2) was run for the final genome using actinopterygii_odb9 dataset and "-sp zebrafish" in the "genome" mode. The new turbot genome assembly was compared with those previously reported in this species [29,32] to check for confidence. The assembly by Figueras et al. [29] was recently improved by Maroso et al. [31], who used a highly dense genetic map for anchoring more than 95% of its scaffolds in the 22 chromosomes of the turbot karyotype [25]. Basic assembly statistics of the contigs here obtained vs scaffolds of the previous genomes were compared including N50 size, largest contig/scaffold, number of contigs/scaffolds, total genome length and % GC.

The highly dense genetic map by Maroso et al. [31] was used to integrate the large contigs obtained in this study into single chromosome sequences. The sequences of the RAD-tags corresponding to the 18,214 markers ordered through the turbot consensus map [31] were used to look for homology in the contigs for their integration in the 22 linkage groups/chromosomes of the turbot map using the CHROMONOMER program [65]. Finally, LASTZ (using options '–notransition – step=23000 – nogapped – format=rdotplot') was used to compare the reported turbot assemblies [31,32] with the new turbot genome to check for its confidence in future genomic studies, and in particular here, to be used as a reference to study SD in this species.

5.2.6. Comparative genomics with other flatfish species: the SD regions

The improved ZZ male turbot genome was further exploited to complete and refine previous information about genomic reorganizations and chromosome evolution of flatfish, and specifically to check the putative orthology of the SD regions. For this, we compared sequences of the new turbot genome with genomes of Japanese flounder (*Paralichthys olivaceous*; GCA_001904815.2 [33]) and half-smooth tongue sole (*Cynoglossus semilaevis*; GCF_000523025.1 [34]) using LASTZ (same options as outlined before). The sequences of the 22,751 turbot genes were used to identify syntenic patterns with the genomes of both flatfish species with MCscan ([69] JCVI v0.9.14). For this, a threshold *E*-value < e-20 was used for significant homology and the best match was retrieved for comparison. Orthologous relationships between species were represented by plotting syntenic blocks generated by MCscan in circle diagrams using circlize package in R ([70] v0.4.9).

5.3. High-throughput SNP genotyping

The 2b-RAD method [71] was applied to genotype 18,214 SNPs in the 1207 fish (47 parents and 1160 offspring) [31] which were then used for GWAS and segregation analysis on turbot SD. Briefly, SNPs were filtered out according to SNP call-rate < 0.8, individual call-rate < 0.8, identity-by-state >0.95 (both individuals removed) and minor allele frequency (MAF) < 0.01. The filtering done enabled us to obtain a consistent set of SNPs showing a genotyping error < 0.4% estimated using family data [31].

5.4. Sex verification

The phenotypic sex of each fish was verified by histology. The age of the fish was ~150 days post hatching (dph), an age at which turbot gonads are differentiated [30]. For histological analysis, samples were fixed in 4% paraformaldehyde buffer overnight, rinsed in phosphate buffer saline the next day and stored in 70% ethanol. Samples were then dehydrated in a series of alcohols and embedded in paraffin, sectioned at 7 μm thick, stained with hematoxylin-eosin and observed under the microscope. Those animals that could not be sexed unequivocally due to poor quality of the sample or the preparation were discarded (238 out of the original 1445 fish [31]).

5.5. Pinpointing the turbot SD gene: genome-wide association studies (GWAS) and segregation analysis

GWAS was performed using GenABEL R package [72] by applying the mmscore function [73], which accounts for the relatedness between individuals using a genomic kinship matrix. Significance thresholds were calculated using a genome-wide Bonferroni correction, where significance was obtained dividing 0.05 by the number of informative SNPs [74]. GWAS was performed using the whole population data as well as within each family.

A segregation analysis, which complemented GWAS, was done to: i) identify the minimum genomic region where the SDg is located; ii) to gather additional information on other genetic or environmental factors influencing sex; and iii) to establish a crossing-over map at the SD region. For narrowing the SD region, only those informative families showing a highly significant association between sex and markers at the SD region were used. For this, both the association of sex with each marker in each family was obtained through a Fisher's exact test and a global estimation for the major SD region was evaluated through 2×2 contingency tests using only families with a *P*-value <0.00001. The availability of a high density SNP genotype dataset (1 SNP/28 kb) in parents and offspring enabled us to select only informative backcrosses (heterozygous dam x homozygous sire) to track changes in the linkage-phase in offspring within each family (Supplemental File 4). Thus, crossing-overs between the maternal Z and W chromosomes could be identified to check the fraction of the Z or W chromosome inherited from the dam regarding the sex of offspring and to narrow the smaller region were the SDg should be located. Finally, information of all selected families was combined to delimit the region where the SD gene should

be located in the turbot population analyzed. Additionally, this information was used to identify those individuals that, despite having inherited the W or Z chromosomes from the mother, showed an unexpected phenotype, to say, they were a ZW male or a ZZ female, suggesting other genetic or environmental factors underlying. Finally, a crossing-over map of the SD region was constructed using the aforementioned information to evaluate the distribution of crossing-overs across the main SD region and to check for a putative recombination blockage between the W and Z chromosomes.

5.6. Searching for the SD causal variant using the new genome assembly

We looked for genetic differences between the W and the Z chromosomes, including putative reorganizations or duplications, which could be related to SD in turbot. Z and W assemblies were aligned at the narrowed SD region using Bowtie 2 [75]. SAMtools were used to identify and genotype SNPs across the narrowed SD region in the five ZZ males and five WW females to obtain allele frequencies in both sex samples. Then, the relative component of genetic differentiation between both datasets (F_{ST}) was estimated for each locus using GENEPOP 4.7 (http://kimura.univ-montp2.fr/~rousset/Genepop.htm), and finally the differences between the Z and W assemblies at the SD region were manually inspected using the graphic viewer IGV 2.4.10 [76].

The small sample size (10 individuals) and the limited coverage of this re-sequencing evaluation suggested a further validation of the associations detected at species level on a larger sample of males and females. For this, a sample of 46 ZZ males and 45 WW females coming from 10 unrelated families were genotyped using either a PCR on agarose gel for large indels or a SNaPshot protocol for SNPs. Primers for PCR amplification were designed with Primer 3 (http://bioinfo.ut.ee/primer3/) and amplifications were checked on 1% agarose gels. Additionally, an internal primer, either forward or reverse, was designed for the SNPs mini-sequencing reaction required for genotyping following the SNaPshot protocol in an ABI3730 automatic sequencer (Applied Biosystems).

5.7. RNA sequencing (RNA-Seq)

RNA extraction was performed using the RNeasy mini kit (Qiagen) with DNase treatment and RNA quality and quantity were evaluated in a Bioanalyzer (Bonsai Technologies) and in a NanoDrop® ND-1000 spectrophotometer (NanoDrop® Technologies Inc), respectively. RNA samples from whole larvae (10 per time) or male and female fry gonads (3 per time) across the gonad formation and differentiation period (1 to 90 dph) were evenly pooled for sequencing 100 bp PE reads on an Illumina HiSeq 2000 at the Wellcome Trust Centre for Human Genetics, Oxford. Male and female gonads were obtained from 35 dph until 90 dph, when gonads could be confidently extracted, and the genetic sex established in each individual using the SmaUSC-E30 marker [2]. The quality of the sequencing was assessed using FastQC v.0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality filtering and removal of residual adaptor sequences was conducted on read pairs using Trimmomatic v.0.36 [59]. Specifically, Illumina adaptors were clipped from the reads, leading and trailing bases with a Phred score < 20 were removed and the read trimmed if the sliding window average Phred score over four bases was <20. Only reads where both pairs were >36 bp post-filtering were retained. Filtered reads were mapped to the most recent turbot genome assembly (ASM318616v1; Genbank accession GCA_003186165.1; [31]) using STAR v.2.5.2b [77], the maximum number of mismatches for each read pair was set to 10% of trimmed read length, and minimum and maximum intron lengths were set to 20 bases and 1 Mb respectively. STAR alignment files were used to reconstruct the turbot transcriptome using Cufflinks v.2.2.1 [78]. Transcript abundance was quantified and normalized using kallisto v0.44.0 [79].

5.8. Real-Time PCR

The RNA samples across the gonad differentiation period where the gonad could be extracted (35 to 90 dph) were individually analyzed by qPCR to test gene expression on candidate genes in males and females (genotyped using the SmaUSC-E30 marker). Primers for candidate genes at the major SD region identified by their genomic position and their expression during the gonadal differentiation period were designed for qPCR using the Primer 3 software. Reactions were performed using a qPCR Master Mix Plus for SYBR Green I No ROX (Eurogenetec) following the manufacturer instructions, and qPCR was carried out on a MX3005P (Agilent Technologies). Analyses were performed using the MxPro software (Agilent). The ΔΔCT method was used to estimate expression taking the ribosomal protein S4 (RPS4) and ubiquitin (UBQ) as reference genes. These two genes had been previously validated for qPCR in turbot gonads by Robledo et al. [80]. Two technical replicates were included for each sample. *t*-tests were used to determine significant differences between sex.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.04.007.

Ethics approval and consent to participate

Experimental protocols were approved by the Institutional Animal Care and Use Committee of the University of Santiago de Compostela (Spain).

Availability of data and materials

All dataset generated during this study are included (in this article and as supplementary information files) or are available in public repositories. Turbot Genome Sequencing data are at NCBI databases (Bioproject: PRJNA631898) (https://www.ncbi.nlm.nih.gov/genome/?term=turbot) and BioProject es PRJNA649485 (https://www.ncbi.nlm.nih.gov/bioproject/649485; Accessión numbers: SRX8843737, SRX8843739, SRX8843738).

Consent for publication

Not applicable.

Author contributions

AMV, SL and PM conceived and designed the study. PM, DR and AV wrote the manuscript. XT contributed to the segregation analysis along with qPCR of candidate genes. DR performed GWAS and RNA seq analysis. MM, AGT and AB assembled the turbot Z and W genomes. FM and MH integrated the new genome assembly with the turbot genetic map. BAB and SC provided the families with individuals sexed used for the functional and association analyses of the study, respectively. AB and FM worked on the identification and validation of Z/W sex-associated markers. SL, CB and FP discussed the final version of the manuscript. All authors read and approved the final version of the manuscript.

CRediT authorship contribution statement

Paulino Martínez: Funding acquisition, Supervision, Conceptualization, Writing - original draft, Investigation, Formal analysis, Writing - review & editing. **Diego Robledo:** Writing - original draft, Investigation, Formal analysis, Writing - review & editing. **Xoana Taboada:** Investigation, Formal analysis, Writing - review & editing. **Andrés Blanco:** Software, Data curation, Formal analysis, Writing - review & editing. **Michel Moser:** Software, Data curation, Formal analysis, Writing - review & editing. **Francesco Maroso:** Investigation, Validation, Writing - review & editing. **Miguel Hermida:** Investigation, Validation, Writing - review & editing. **Antonio Gómez-Tato:** Software, Data curation, Formal analysis, Writing - review & editing. **Blanca Álvarez-Blázquez:** Resources, Writing - review & editing. **Santiago Cabaleiro:** Resources, Writing - review & editing. **Francesc Piferrer:** Writing - review & editing. **Carmen Bouza:** Investigation, Formal analysis, Writing - review & editing. **Sigbjørn Lien:** Funding acquisition, Supervision, Conceptualization, Writing - review & editing. **Ana M. Viñas:** Funding acquisition, Supervision, Conceptualization, Writing - original draft, Investigation, Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no competing interest.

Acknowledgements

References

[1] R.H. Devlin, Y. Nagahama, Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences, Aquaculture 208 (2002) 191–364.

[2] P. Martínez, A.M. Viñas, L. Sánchez, et al., Genetic architecture of sex determination in fish: applications to sex ratio control in aquaculture, Front. Genet. 5 (2014) 340.

[3] B. Capel, Vertebrate sex determination: evolutionary plasticity of a fundamental switch, Nat. Rev. Genet. 18 (2017) 675–689.

[4] M.B. Cioffi, C.F. Yano, A. Sember, L.A.C. Bertollo, Chromosomal evolution in lower vertebrates: sex chromosomes in neotropical fishes, Genes 8 (2017) 258.

[5] 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 417 (2002) 559–563.

[6] Y. Takehana, M. Matsuda, T. Myosho, et al., Co-option of Sox3 as the male-determining factor on the Y chromosome in the fish *Oryzias dancena*, Nat. Commun. 5 (2014) 4157.

[7] T. Myosho, H. Otake, H. Masuyama, et al., Tracing the emergence of a novel sex-determining gene in medaka, *Oryzias luzonensis*, Genetics 191 (2012) 163–170.

[8] R.S. Hattori, Y. Murai, M. Oura, et al., A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 2955.

[9] 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 Genet. 15 (2019), e1008013.

[10] 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 Genet. 8 (2012), e1002798.

[11] L. Bao, C. Tian, S. Liu, et al., The Y chromosome sequence of the channel catfish suggests novel sex determination mechanisms in teleost fish, BMC Biol. 17 (2019) 6.

[12] T. Koyama, M. Nakamoto, K. Morishima, et al., A SNP in a steroidogenic enzyme is associated with phenotypic sex in *Seriola* fishes, Curr. Biol. 29 (2019) 1901–1909.

[13] A. Yano, B. Nicol, E. Jouanno, et al., The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids, Evol. Appl. 6 (2013) 486–496.

[14] C.L. Peichel, J.A. Ross, C.K. Matson, et al., The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome, Curr. Biol. 14 (2004) 1416–1424.

[15] T.G. Kirubakaran, Ø. Andersen, M.C. De Rosa, et al., Characterization of a male specific region containing a candidate sex determining gene in Atlantic cod, Sci. Rep. 9 (2009) 116.

[16] J.R. Ser, R.B. Roberts, T.D. Kocher, Multiple interacting loci control sex determination in Lake Malawi cichlid fish, Evolution 64 (2010) 486–501.

[17] C. Palaiokostas, M. Bekaert, M.G.Q. Khan, et al., Mapping and validation of the major sex-determining region in Nile Tilapia (*Oreochromis niloticus* L.) using RAD sequencing, PLoS One 8 (2013), e68389.

[18] D.J. Penman, F. Piferrer, Fish gonadogenesis. Part I. Genetic and environmental mechanisms of sex determination, Rev. Fish. Sci. 16 (S1) (2008) 16–34.

[19] M.H. Tan, C.M. Austin, M.P. Hammer, et al., Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (*Amphiprion ocellaris*) genome assembly, GigasSience 7 (2018), gix137.

[20] W. Pan, T. Jiang, S. Lonardi, OMGS: optical map-based genome scaffolding, in: L. Cowen (Ed.), Research in Computational Molecular Biology. RECOMB 2019. Lecture Notes in Computer Science 11467, Springer, Cham, 2019.

[21] D. Robledo, M. Hermida, J.A. Rubiolo, et al., Integrating genomic resources of flatfish (Pleuronectiformes) to boost aquaculture production, Comp. Biochem. Phys. Part D: Genom. Proteom. 21 (2017) 41–55.

[22] F. Piferrer, R.M. Cal, C. Gómez, et al., Induction of gynogenesis in the turbot (*Scophthalmus maximus*): effects of UV irradiation on sperm motility, the Hertwig effect and viability during the first 6 months of age, Aquaculture 238 (2004) 403–419.

[23] P. Haffray, E. Lebègue, S. Jeu, et al., Genetic determination and temperature effects on turbot *Scophthalmus maximus* sex differentiation: an investigation using steroid sex-inverted males and females, Aquaculture 294 (2009) 30–36.

[24] P. Martínez, C. Bouza, M. Hermida, et al., Identification of the major sex-determining region of turbot (*Scophthalmus maximus*), Genetics 183 (2009) 1443–1452.

[25] C. Bouza, L. Sánchez, P. Martínez, Karyotypic characterization of turbot (*Scophthalmus maximus*) with conventional, fluorochrome, and restriction endonuclease banding techniques, Mar. Biol. 120 (1994) 609–613.

[26] N. Cuñado, J. Terrones, L. Sánchez, et al., Sex-dependent synaptic behaviour in triploid turbot, *Scophthalmus maximus* (Pisces, Scophthalmidae), Heredity 89 (2002) 460–464.

[27] X. Taboada, M. Hermida, B.G. Pardo, A. Viñas, et al., Fine mapping and evolution of the major sex determining region in turbot (*Scophthalmus maximus*). G3, Gene Genom. Genet. 10 (2014) 1871–1880.

[28] M. Hermida, C. Bouza, C. Fernández, et al., Compilation of mapping resources in turbot (*Scophthalmus maximus*): a new integrated consensus genetic map, Aquaculture 414–415 (2013) 19–25.

[29] A. Figueras, D. Robledo, A. Corvelo, et al., Whole genome sequencing of turbot (*Scophthalmus maximus*; Pleuronectiformes): a fish adapted to demersal life, DNA Res. 23 (2016) 181–192.

[30] D. Robledo, L. Ribas, R. Cal, et al., Gene expression analysis at the onset of sex differentiation in turbot (*Scophthalmus maximus*), BMC Genomics 16 (2015) 973.

[31] F. Maroso, M. Hermida, A. Millán, et al., Highly dense linkage maps from 31 full-sibling families of turbot (*Scophthalmus maximus*) provide insights into recombination patterns and chromosome rearrangements throughout a newly refined genome assembly, DNA Res. 25 (2018) 439–450.

[32] X. Xu, C. Shao, H. Xu, et al., Draft genomes of female and male turbot *Scophthalmus maximus*, Sci. Data 7 (2020) 90.

[33] C. Shao, B. Bao, Z. Xie, et al., The genome and transcriptome of Japanese flounder provide insights into flatfish asymmetry, Nat. Genet. 49 (2017) 119–124.

[34] S. Chen, G. Zhang, C. Shao, et al., Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle, Nat. Genet. 46 (2014) 253–260.

[35] P.P. Amaral, C. Neyt, S.J. Wilkins, et al., Complex architecture and regulated expression of the *Sox2ot* locus during vertebrate development, RNA 15 (2009) 2013–2027.

[36] D. Charlesworth, Young sex chromosomes in plants and animals, New Phytol. 224 (2019) 1095–1107.

[37] Y. Youa, X. Shanab, Q. Shiab, Research advances in the genomics and applications for molecular breeding of aquaculture animals, Aquaculture 526 (2020) 735357.

[38] D.P. Drinan, T. Loher, L. Hauser, Identification of genomic regions associated with sex in Pacific halibut, J Hered. 109 (3) (2018 Mar 16) 326–332, https://doi.org/10.1093/jhered/esx102. 29136178.

[39] C. Palaiokostas, M. Bekaert, A. Davie, M.E. Cowan, M. Oral, J.B. Taggart, K. Gharbi, B.J. McAndrew, D.J. Penman, H. Migaud, Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing, BMC Genomics 14 (2013) 566.

[40] X. Taboada, D. Robledo, C. Bouza, et al., Reproduction and sex control in turbot, in: H.P. Wang, F. Piferrer, S.L. Chen, Z.G. Shen (Eds.), Sex Control in Aquaculture II, John Wiley & Sons Ltd, 2019, pp. 565–582.

[41] H.Y. Zhou, Y. Katsman, N.K. Dhaliwal, et al., A *Sox2* distal enhancer cluster regulates embryonic stem cell differentiation potential, Genes Dev. 28 (2014) 2699–2711.

[42] T.C. Messemaker, S.M. van Leeuwen, P.R. van den Berg, et al., Allele-specific repression of *Sox2* through the long non-coding RNA *Sox2ot*, Sci. Rep. 8 (2018) 386.

[43] F. Piferrer, Epigenetics of sex determination and gonadogenesis, Dev. Dyn. 242 (2013) 360–370.

[44] B.B. Mulvey, U. Olcese, J.R. Cabrera, et al., An interactive network of long non-coding RNAs facilitates the *Drosophila* sex determination decision, Biochim. Biophys. Acta 2014 (1839) 773–784.

[45] Y. Kato, C.A.G. Perez, N.S. Mohamad Ishak, et al., A 5′ UTR-overlapping lncRNA activates the male-determining gene doublesex1 in the crustacean *Daphnia magna*, Curr. Biol. 28 (2018) 1811–1817.

[46] S. Liang, D. Liu, X. Li, et al., *SOX2* participates in spermatogenesis of Zhikong scallop *Chlamys farreri*, Sci. Rep. 9 (2019) 76.

[47] L. Guo, Y.H. Xu, N. Zhang, et al., A high-density genetic linkage map and QTL mapping for sex in black tiger shrimp (*Penaeus monodon*), Front. Genet. 10 (2019) 326.

[48] N. Ospina-Alvarez, F. Piferrer, Temperature-dependent sex determination in fish revisited: prevalence, a single sex ratio response pattern, and possible effects of climate change, PLoS One 3 (2008), e2837.

[49] O. Anacleto, S. Cabaleiro, B. Villanueva, et al., Genetic differences in host infectivity affect disease spread and survival in epidemics, Sci Rep. 9 (1) (2019) 4924.

[50] M. Saura, M.J. Carabaño, A. Fernández, et al., Disentangling genetic variation for resistance and endurance to scuticociliatosis in turbot using pedigree and genomic information, Front. Genet. 10 (2019) 539.

[51] M. Tomaszkiewicz, P. Medvedev, K.D. Makova, Y and W chromosome assemblies: approaches and discoveries, Trends Genet. 33 (2017) 266–282.

[52] J.J. Bull, Evolution of Sex Determining Mechanisms, Benjamin Cummings Press, Menlo Park, CA, USA, 1983.

[53] B. Charlesworth, The evolution of chromosomal sex determination and dosage compensation, Curr. Biol. 6 (1996) 149–162.

[54] B. Charlesworth, P. Sniegowski, W. Stephan, The evolutionary dynamics of repetitive DNA in eukaryotes, Nature 371 (1994) 215–220.

[55] P. Martínez, D. Robledo, S.T. Rodríguez-Ramilo, et al., Turbot (*Scophthalmus maximus*) genomic resources: Application for boosting aquaculture production, in: S. MacKenzie, S. Jentoft (Eds.), Genomics in Aquaculture, London, Elsevier, 2016, pp. 131–163.

[56] S. Chen, Y. Zhou, Y. Chen, et al., Fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890.

[57] V.P. Cruz, M. Vera, B.G. Pardo, et al., Identification and validation of single nucleotide polymorphisms as tools to detect hybridization and population structure in freshwater stingrays, Mol. Ecol. Resour. 17 (2017) 550–556.

[58] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2004) 2114–2120.

[59] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv 00 (2013) 1–3, 1303.3997 [q-bio.GN].

[60] H. Li, B. Handsaker, A. Wysoker, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[61] M. Kolmogorov, J. Yuan, Y. Lin, et al., Assembly of long, error-prone reads using repeat graphs, Nat. Biotechnol. 37 (2019) 540–546.

[62] M. Chakraborty, J.G. Baldwin-Brown, A.D. Long, et al., Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage, Nucleic Acids Res. 44 (2016), e147.

[63] K. Shafin, T. Pesout, R. Lorig-Roach, et al., Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes, Nat. Biotechnol. (2020), https://doi.org/10.1038/s41587-020-0503-6.

[64] B.J. Walker, T. Abeel, T. Shea, et al., Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS One 9 (2014), e112963.

[65] J. Catchen, A. Amores, S. Bassham, Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny, bioRxiv. (2020), https://doi.org/10.1101/2020.02.04.934711.

[66] R. Vaser, I. Sović, N. Nagarajan, et al., Fast and accurate de novo genome assembly from long uncorrected reads, Genome Res. 27 (2017) 737–746.

[67] T.D. Wu, C.K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, Bioinformatics 21 (2005) 1859–1875.

[68] M. Seppey, M. Manni, E.M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness, in: M. Kollmar (Ed.), Gene Prediction. Methods in Molecular Biology vol. 1962, Humana, New York, NY, 2019, pp. 227–245.

[69] H. Tang, J.E. Bowers, X. Wang, et al., Synteny and collinearity in plant genomes, Science 320 (2008) 486–488.

[70] Z. Gu, L. Gu, R. Eils, et al., *circlize* implements and enhances circular visualization in R, Bioinformatics 30 (2014) 2811–2812.

[71] S. Wang, E. Meyer, J.K. McKay, M.V. Matz, 2b-RAD: a simple and flexible method for genome-wide genotyping, Nat. Methods 9 (2012) 808–810.

[72] Y.S. Aulchenko, S. Ripke, A. Isaacs, et al., GenABEL: an R library for genome-wide association analysis, Bioinformatics 23 (2007) 1294–1296.

[73] Chen WM, Abecasis G.R. Family-based association tests for genome wide association scans. Am. J. Hum. Genet. 2007;81:913–26.

[74] P. Duggal, E.M. Gillanders, T.N. Holmes, et al., Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies, BMC Genomics 9 (2008) 516.

[75] B. Langmead, S. Salzberg, Fast gapped-read alignment with bowtie 2, Nat. Methods 9 (2012) 357–359.

[76] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, et al., Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26.

[77] A. Dobin, C.A. Davis, F. Schlesinger, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21.

[78] C. Trapnell, B.A. Williams, G. Pertea, et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (2010) 511–515.

[79] N.L. Bray, H. Pimentel, P. Melsted, et al., Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (2016) 525–527.

[80] D. Robledo, J. Hernández-Urcera, R.M. Cal, et al., Analysis of qPCR reference gene stability determination methods and a practical approach for efficiency calculation on a turbot (*Scophthalmus maximus*) gonad dataset, BMC Genomics 15 (2014) 684.