



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

# Non-synonymous variation and protein structure of candidate genes associated with selection in farm and wild populations of turbot (*Scophthalmus maximus*)

### Citation for published version:

Andersen, Ø, Rubiolo, JA, Pirolli, D, Aramburu, O, Pampín, M, Righino, B, Robledo, D, Bouza, C, De Rosa, MC & Martínez, P 2023, 'Non-synonymous variation and protein structure of candidate genes associated with selection in farm and wild populations of turbot (*Scophthalmus maximus*)', *Scientific Reports*, vol. 13, no. 1, 3019. <https://doi.org/10.1038/s41598-023-29826-z>

### Digital Object Identifier (DOI):

[10.1038/s41598-023-29826-z](https://doi.org/10.1038/s41598-023-29826-z)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

Scientific Reports

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy



The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.





OPEN

## Non-synonymous variation and protein structure of candidate genes associated with selection in farm and wild populations of turbot (*Scophthalmus maximus*)

Øivind Andersen<sup>1,2</sup>, Juan Andrés Rubiolo<sup>3</sup>, Davide Pirolli<sup>4</sup>, Oscar Aramburu<sup>3</sup>, Marina Pampín<sup>3</sup>, Benedetta Righino<sup>4</sup>, Diego Robledo<sup>5</sup>, Carmen Bouza<sup>3</sup>, Maria Cristina De Rosa<sup>4</sup> & Paulino Martínez<sup>3</sup>

Non-synonymous variation (NSV) of protein coding genes represents raw material for selection to improve adaptation to the diverse environmental scenarios in wild and livestock populations. Many aquatic species face variations in temperature, salinity and biological factors throughout their distribution range that is reflected by the presence of allelic clines or local adaptation. The turbot (*Scophthalmus maximus*) is a flatfish of great commercial value with a flourishing aquaculture which has promoted the development of genomic resources. In this study, we developed the first atlas of NSVs in the turbot genome by resequencing 10 individuals from Northeast Atlantic Ocean. More than 50,000 NSVs were detected in the ~21,500 coding genes of the turbot genome, and we selected 18 NSVs to be genotyped using a single Mass ARRAY multiplex on 13 wild populations and three turbot farms. We detected signals of divergent selection on several genes related to growth, circadian rhythms, osmoregulation and oxygen binding in the different scenarios evaluated. Furthermore, we explored the impact of NSVs identified on the 3D structure and functional relationship of the correspondent proteins. In summary, our study provides a strategy to identify NSVs in species with consistently annotated and assembled genomes to ascertain their role in adaptation.

Marine fish species are often distributed across a variety of habitats differing in environmental conditions, particularly water temperature, salinity, dissolved oxygen and light intensity, which in turn affect the distribution of infectious pathogens that, along with predation, compromise their viability. These environmental factors strongly influence somatic growth and reproduction, and furthermore, all of them represent energetically costly metabolic activities engaged by fish. Metabolism and evolution are closely connected<sup>1</sup>, and the many genes underlying these processes are targets of natural selection that may lead to adaptive divergence in organisms inhabiting heterogeneous environments<sup>2-7</sup>. Knowledge about adaptive genetic variation and its spatial structuring is crucial for the sustainable management of wild fish resources, but also for improving production traits by selective breeding of economical important aquaculture species. Genomic sequencing of an increasing number of fish species is contributing to unravelling the broad genetic variation across genomes through identification of thousands of single nucleotide polymorphic sites (SNPs) and their possible association with various traits both in domestic and wild populations.

Turbot (*Scophthalmus maximus*; Scophthalmidae; Pleuronectiformes) is a flatfish widely distributed throughout the European coast in the Northeast Atlantic Ocean from Morocco to the Arctic Circle, including the Baltic Sea, and in the South across the Mediterranean Sea until the Black Sea<sup>8</sup>. The species experiences a diversity of

<sup>1</sup>Nofima, PO Box 5010, 1430 Ås, Norway. <sup>2</sup>Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences (NMBU), Ås, Norway. <sup>3</sup>Department of Zoology, Genetics and Physical Anthropology, Faculty of Veterinary, University of Santiago de Compostela, 27002 Lugo, Spain. <sup>4</sup>Institute of Chemical Sciences and Technologies "Giulio Natta" (SCITEC), CNR, 00168 Rome, Italy. <sup>5</sup>The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, UK. ✉email: oivind.andersen@nofima.no; paulino.martinez@usc.es

physicochemical environments across its range with a north–south temperature cline from  $\approx 7^\circ\text{C}$  up to  $\approx 22^\circ\text{C}$  and with salinities spanning from  $\approx 35$  PSU in the North Atlantic Ocean to  $\approx 2$  PSU in the northern Baltic Sea<sup>9</sup>. Whereas the juveniles and adults are relatively sedentary, the pelagic larvae possess high dispersal potential mediated by oceanic currents and enhanced by the high fecundity of the species<sup>10</sup>. Genetic diversity and population structure of turbot has been investigated with microsatellites and SNPs, mostly in the North Atlantic Ocean<sup>11–14</sup> and to a minor extent in the Southern area<sup>15,16</sup> and adaptive variation, across its full distribution range, was recently assessed using a set of SNPs covering the whole genome<sup>17</sup>. Four main genetic regions: Baltic, Atlantic, Mediterranean and Black Sea, were identified using neutral variation. Consistent signals of divergent selection attributed to salinity and temperature, and stabilizing selection related to salinity, were detected across the turbot genome, including candidate genes at specific regions<sup>18</sup>. Moreover, the same set of markers was used to analyze genetic differentiation between wild and farm populations, as a baseline to evaluate the impact of restocking and farm escapes in the wild<sup>18</sup>. Besides the notable differentiation detected (wild vs farm  $F_{\text{ST}} \sim 0.060$ ), signals of selection mostly attributed to growth and resistance to pathologies were detected at specific genomic regions including candidate genes.

Whereas wild turbot populations have declined over the last decades mainly due to overfishing, this delicious species has become the main flatfish farmed worldwide due to its high commercial value<sup>19</sup>. Intensive farming during six generations has been accompanied by a fast development of genomic resources to identify quantitative trait loci (QTL) and candidate genes associated with: growth<sup>19–23</sup>, temperature tolerance<sup>24</sup>, adaptation to salinity<sup>25</sup>, sex determination<sup>26,27</sup> and resistance to various pathogens<sup>28,29</sup>. Furthermore, runs of homozygosity and genetic diversity across the turbot genome were analyzed to check for selective sweeps in farm and wild populations. This information was integrated with previously reported QTL-associated markers, candidate genes and outlier loci related to natural or artificial selection, and a robust framework on selection signatures across the turbot genome was obtained<sup>30</sup>. Furthermore, functional data on resistance to the main industrial pathogens obtained from the main immune organs have been comparatively assessed and integrated with previous signatures of selection across the turbot genome<sup>31</sup>. The broad information gathered in this species, both in wild and farm populations, make it a suitable candidate for assessing the relevance of the different sources of genetic variation on turbot adaptation to different scenarios.

Studying association between polymorphisms within candidate genes and traits of interest or environmental variables in populations or families with different genomic background is a convenient approach to validate their putative adaptive role on natural or domestic selective pressures<sup>23,32</sup>. The significant association can be taken as evidence that the gene is either directly involved in the control of the trait or in linkage disequilibrium with the responsible variant due to its vicinity. If non-synonymous variation is considered, association could eventually lead to the identification of the causative mutation as reported in various vertebrates, including teleost fish. In Atlantic salmon (*Salmo salar*), non-synonymous SNPs in two strong candidate genes coding for the epithelial cadherin and the NEDD-8 activating enzyme 1 (NAE1)<sup>33,34</sup> have been suggested to be responsible for resistance to infectious pancreatic necrosis virus. Differences in spawning time associated with functionally different protein variants have been documented in Atlantic salmon vestigial-like protein 3 (VGLL3) and in herring (*Clupea harengus*) thyrotropin receptor (TSHR)<sup>35,36</sup>. A hemoglobin polymorphism in turbot was reported to be associated with differences in juvenile growth rates<sup>37,38</sup> and the underlying amino acid substitution was predicted to influence the stability of the oxygen-binding protein<sup>39</sup>.

Advances in modelling three-dimensional (3D) protein structures together with the progressive enrichment on mutation databases are making feasible to approach the interpretation of non-synonymous variation in terms of protein function<sup>40–42</sup>. This information is essential to understand the evolutionary significance of non-synonymous variation associated with environmental variables<sup>43–46</sup>. In silico approaches for predicting the protein 3D structure directly from the sequence information play a key role in filling the gap between the numerous sequences available and the experimentally solved structures<sup>47,48</sup>. In the absence of sequence similarity with other sequences in the protein structure database (PDB), the modelling strategy can rely on threading and ab initio modelling<sup>49,50</sup> or deep learning<sup>42,51</sup> to predict protein structure.

The amount of genomic information on adaptive variation in wild and farmed turbot prompted us to ascertain the putative role of non-synonymous variants (NSV) of candidate genes on selection related to environmental variation in nature or associated with target traits in breeding programs of turbot aquaculture. Specifically, we committed to: (i) call NSV using resequencing data over the recently assembled chromosome-level turbot genome; (ii) filter the most consistent and relevant functional variants among the  $\sim 21,500$  protein coding genes in the turbot genome; (iii) select NSV on candidate genes putatively related to osmoregulation, growth and disease resistance; (iv) identify signals of selection across the whole distribution range of the species and farms; and (v) to validate functional differences of the most consistent variants using 3D structural protein modelling. Our results provide a broad map of NSV across the turbot genome and support the role of several candidate variants on adaptation to osmotic changes or growth in wild and domestic populations.

## Materials and methods

**Calling non-synonymous variation in the turbot genome.** DNA from ten adults (five males and five females) of commercial size (1.5 kg) coming from the breeding program of a turbot company were re-sequenced using 150 bp PE reads on an Illumina NovaSeq 6000 System to  $20\times$  coverage and individually aligned against the turbot reference genome (GCA\_013347765.1)<sup>27</sup> to screen for SNP variation. Individuals were previously checked for parentage using a set of 9 microsatellites to choose unrelated individuals<sup>52</sup> representative of the genetic diversity of the broodstock. The origin of the founders was the NE Atlantic Ocean, and this population has been selectively bred for five generations with the support of the microsatellite tool mentioned above to avoid inbreeding while retaining as much genetic diversity as possible. Quality filtering and removal of residual

adaptor sequences was conducted on read pairs using Fastp v.0.20.0<sup>53</sup>; then, filtered reads were mapped with the Burrows-Wheeler aligner v.0.7.8 BWA-MEM algorithm<sup>54</sup> against the turbot genome and SNPs and indels were called using bcftools v1.<sup>55</sup>, discarding those aligned reads with a mapping quality (MAPQ) < 30 and those SNPs with a Phred quality score < 30. Variants were annotated using SNPeff v5.1<sup>56</sup> taking as reference the updated turbot chromosome-level genome assembly (GCA\_013347765.1<sup>27</sup>).

**Filtering of NSV: reliability and functional information.** The thousands of NSV detected were filtered following functional, technical and population genetics criteria to obtain a map of the most consistent NSV across the turbot genome following previous filtering pipelines reported for the species<sup>27,57</sup>. Functional criteria included: (i) dismiss putative pseudogenes using a conservative criterion, to say, those genes with 3 or more NSVs were discarded; (ii) remove non-sense variants producing truncated proteins; and (iii) discard genes with low-quality annotation. Technical criteria included: (i) availability of  $\pm 100$  bp without additional variation which could compromise primer annealing and PCR amplification for further genotyping; (ii) compatibility of the adjacent regions selected for designing multiplex primer panels for genotyping; (iii) validation of the in silico detected allelic variants with the MassARRAY technology<sup>58</sup>. Population genetics criteria included: (i) discard SNPs deviated from Hardy–Weinberg proportions ( $P < 0.01$ ); and (ii) remove tri-allelic SNPs. From this broad NSVs map, we performed additional filtering to focus on the main traits putatively associated with selection in wild or farm populations where previous information was available to choose a final manageable set of SNPs for validation: (i) select the most relevant candidate genes related to growth, osmoregulation and resistance to pathologies crossing previous literature, mostly on fishes, with previous QTL and functional (differentially expressed genes, DEG) data in turbot<sup>28,29,31</sup>; (ii) identifying suggestive genes close to markers associated with signatures of selection (< 500 kb)<sup>13,14,30</sup>; (iii) discarding deleterious variants from previous information in other species for the same genes available in public repositories (PROVEAN software<sup>59</sup>); (iv) selecting the most diverse SNP per locus (higher MAF: minimum allele frequency). The conservation of the substituted residues in the 18 selected turbot protein variants was examined by blasting against the corresponding proteins in other teleost species available at NCBI (<https://www.ncbi.nlm.nih.gov/>).

**Population genetics of selected NSVs across the turbot distribution range.** *Sampling.* In our screening, we analyzed 13 wild populations including the main genetic regions reported across the turbot distribution range<sup>17</sup>, also representative of the wide variety in temperature and salinity, the main drivers for selection in turbot<sup>17</sup>, but very likely also influencing pathogen distribution<sup>60,61</sup>. We also included samples from the broodstock of the three turbot companies carrying out breeding programs for comparison with wild samples to detect signals of selection related to the main target traits. The broodstock of the three main turbot companies, located in NW Spain and France, were founded with individuals collected from NE Atlantic Ocean<sup>18</sup>, where non-significant genetic differentiation was reported with neutral markers<sup>17</sup>. A total of 355 individuals were analyzed from 16 sampling locations, mostly exceeding 20 individuals/sample (AQUATRACE project; Fig. 1, Table 1). Wild samples included the four main regions of the turbot distribution: Baltic Sea (BAS), Atlantic Ocean (ATL), Mediterranean Sea (MED) and Black Sea (BLS)<sup>17</sup>. The Atlantic Ocean region was overrepresented because of the



**Figure 1.** Sampling of wild turbot (*Scophthalmus maximus*) across its European distribution range.

Sample location	Pop code	Sample size	Genetic region	Coordinates (lat-long)
Baltic Sea—North	BAS-N	25	Baltic Sea	60.2/19.7
Skagerak	SK	25	Atlantic	57.4/9.2
Norway Sea	NOR	19	Atlantic	62.0/4.0
North Sea—South	NS-S	24	Atlantic	51.8/2.0
Ireland- West	IR-W	25	Atlantic	59.3/−4.5
Ireland—East	IR-E	25	Atlantic	53.5/−4.8
English Channel	ECH	18	Atlantic	50.7/8.4
Biscay bay—France	BB-FR	23	Atlantic	46.2/−2.2
Biscay bay—Southeast	BB-SE	25	Atlantic	43.4/−3.8
Spain coast – West	SP-W	26	Atlantic	42.6/−8.9
Adriatic Sea	AD	25	Mediterranean	45.2/12.3
Black Sea North	BLS-N	17	Black Sea	44.6/33.4
Black Sea South	BLS-S	28	Black Sea	41.1/31.1
Domestic	Farm 1	14	na	na
Domestic	Farm 2	11	na	na
Domestic	Farm 3	25	na	na

**Table 1.** Sampling information of turbot (*Scophthalmus maximus*). (na, not applicable).

higher abundance of the species. Farm samples included a representative sample of the broodstock of the three European turbot companies with ongoing breeding programs<sup>18</sup>.

**SNP genotyping.** To genotype and validate in silico allelic variants of the SNPs finally selected we used the MassARRAY technology. Briefly, the protocol consists of a two-step reaction: i) PCR amplification of an amplicon of ~ 150 bp including the selected SNP; and ii) mini-sequencing reaction using an internal primer adjacent to the SNP which extends the primer with a dideoxy nucleotide complementary to the SNP variant<sup>58</sup>. Flanking regions of ± 100 nucleotides of the selected SNPs were obtained from the turbot reference genome (GCA\_013347765.1). Design of primer multiplexes and MassARRAY genotyping was done at the UCIM-Universitat de Valencia Genomics Platform.

**Genetic diversity and differentiation.** Mean number of alleles per locus ( $N_a$ ) and expected ( $H_E$ ) and observed ( $H_o$ ) heterozygosities were estimated to assess genetic diversity per locus. Departure from Hardy–Weinberg equilibrium (HWE) and intrapopulation fixation index ( $F_{IS}$ ) were tested for each locus and population. Global  $F_{ST}$  across loci was estimated considering all samples, but also wild sample and farm sample groups separately. Analyses were performed using GENEPOP v4.0<sup>62</sup>.

**Detection of outlier loci.** We followed two different statistical approaches to detect outlier loci showing signals of divergent or balancing selection implemented in BAYESCAN v2.1<sup>63</sup> and ARLEQUIN v3.5<sup>64</sup>, respectively. Outliers were investigated in: (i) all samples, (ii) wild samples, and (iii) wild vs farm samples; additionally, a hierarchical approach was also explored considering two hierarchical groups (wild vs farmed); in all cases we used as background the neutral datasets previously reported for the same comparisons by do Prado et al.<sup>17,18</sup>. The following BAYESCAN parameters were used: 100,000 burn-in length, prior odds of 10 and 20 pilot runs, to identify outliers using a  $q$  value < 0.05. The FDI<sub>ST</sub> method implemented in ARLEQUIN was used to investigate loss of heterozygosity after selective sweeps regarding  $F_{ST}$ . For this program we used the following parameters: 50,000 simulations, 100 demes per group and 20 groups when a hierarchical model was applied. In all ARLEQUIN analyses, outliers were identified considering a  $P$ -value < 0.01, considering it is prone to a higher number of false positives<sup>65</sup>. The hierarchical scenario could only be implemented with ARLEQUIN, because this option is not available in BAYESCAN.

**Protein 3D structure modelling of non-synonymous variants.** To find potential template structures for homology modelling, a specific PSI-BLAST sequence search in the Protein Data Bank (PDB) was performed (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)<sup>66</sup>. Identified template structures showed large unresolved regions which encompassed point mutations analyzed in the present study. Two different strategies for modelling were therefore undertaken: I-TASSER<sup>48</sup> and RoseTTAFold<sup>51</sup>. I-TASSER is a metasever that automatically employs ten threading algorithms in combination with ab initio modelling to build the tertiary structure of a protein as well as replica-exchange Monte Carlo dynamics simulations for the atomic-level refinement. For comparison an algorithm led by artificial intelligence, RoseTTAFold (<https://robetta.bakerlab.org>) was also used. The presence of intrinsically disordered regions in the proteins was investigated by the following disorder predictors: PONDR<sup>67</sup>, DISOPRED<sup>68</sup>, IUPRED3<sup>69</sup> and PrDOS<sup>70</sup>.

Homology modelling was used to generate the 3D model structures of the polymorphic turbot HbaD (see Results) together with the turbot Hbβ1 subunit (AWP17400.1) in the deoxy form (T-state). The structure of deoxyhemoglobin of the Antarctic icefish *Pagothenia bernacchii* (PDB code: 1HBH)<sup>71</sup> was selected as the most

appropriate template to generate the tetramer model (sequence identities of 82.3 and 76% for  $\alpha$  and  $\beta$  chains, respectively). Twenty models of each Hb variant were built using MODELLER<sup>72</sup> as implemented in Biovia Discovery Studio. The model with the lowest MODELLER objective function was selected for analysis.

## Results

**Non-synonymous variants and filtering.** Among the ~3.3 M SNPs detected in the ten 20×-resequenced turbot samples, 55,176 represented NSVs after quality control (MAPQ < 30; PHRED < 30; Supplementary Table S1). Among the filtering steps used to select a consistent and manageable set of NSVs for validation, genes with  $\geq 3$  NSVs (82.9% drop over the previous step) and the functional criterion of selecting genes previously identified associated with growth, osmoregulation and resistance to pathogens (87.4% drop), were the most decisive (Fig. 2; Supplementary Table S1). In the last step, information on candidate genes related to growth and osmoregulation either in turbot or in other fish species (see Introduction for citations), but also for resistance to the main turbot pathogens, *Aeromonas salmonicida* (AS, furunculosis), *Philasterides dicentrarchi* (PD, scuticociliatosis) and *Enteromyxum scolephthalmi* (ES, enteromixosis) was used to retain 1179 SNPs in 876 genes. The SNP with highest MAF for each gene was retained. A total of 84 turbot NSVs were detected in other species using PROVEAN database, and among them, eight were categorized as deleterious and thus discarded for further analyses (Supplementary Table S2). The number of transitions was very similar to that of transversions in the 876 listed NSVs: 432 transitions (A/G = 227; C/T = 205) vs 444 transversions (A/C = 129; AT = 81; C/G = 120; G/T = 115).

**Selection for genotyping and population screening.** Our intention was to select a final set of ~25 NSVs from the consistent list of 868 candidates to be genotyped in a single multiplex using the MassARRAY technology to validate the reliability of our pipeline and to search for signals of natural or artificial selection in turbot populations across its distribution range. Furthermore, to add functional support, 3D protein structure was evaluated specially on those genes showing significant signals of selection. Accordingly, we focused on genes previously associated with signals of selection in the wild or farm populations related to growth, osmoregulation



**Figure 2.** Filtering steps of non-synonymous variants identified in turbot using technical, population genetics and functional criteria.

and resistance to pathogens, detected either by functional assays (DEG: differentially expressed genes) or QTL associations in turbot, but also in other fish species (Table 2). Most of the genes included in the list matched to more than one selection criteria, except for *hbaD* (hemoglobin subunit alpha-D) of particular interest regarding metabolism and growth<sup>39</sup>. The final list included 22 genes associated with growth (13 genes); resistance to ES (13), AS (7) and PD (9); osmoregulation (3); and signals of natural (7) or artificial (3) divergent selection, mostly in turbot, but also from other fish species (10) (Table 2).

**Multiplex design and genotyping on a MassARRAY platform.** Among the 22 preselected SNPs, 18 could be included in a single multiplex for MassARRAY genotyping using primers designed from the  $\pm 100$  bp flanking regions retrieved from the turbot genome (Supplementary Table S3 and S4). In all cases, the allelic variants detected with MassARRAY genotyping matched with the in silico SNP calling from the re-sequencing turbot data and thus they were validated for further research. Genotypes for the 355 individuals from wild and farm origin were very consistent and only one missing data was detected among the 6390 genotypes (Supplementary Table S5).

**Genetic diversity and differentiation across loci, populations and groups.** Global genetic diversity in the wild for the set of 18 SNPs was significantly higher than previously reported using an anonymous SNP panel across the whole genome ( $N_a$ : 1.77 vs 1.49;  $H_E$ : 0.223 vs 0.090, respectively<sup>17</sup>), which can be explained by the filtering criterion followed for detecting NSVs in this study (at least two variants in the 10 individuals analyzed (20 alleles per locus); minimum allele frequency (MAF)=0.1). Also, genetic diversity was higher on average in farm than in wild samples ( $H_E$ =0.261 vs 0.214) even for the Atlantic region ( $H_E$ =0.227) suggesting a good management of genetic diversity after five generations of selection. Average genetic diversity per locus ranged from *aqp8b* (aquaporin 8b) ( $N_a$ =1.19;  $H_E$ =0.0199) to *vipr1b* (vasoactive intestinal peptide receptor 1b) ( $N_a$ =2;  $H_E$ =0.4664), but other loci, such as *eya3* (eyes absent 3), *hamp* (hepcidin antimicrobial peptide), *fga-like* (fibrinogen-alpha chain-like), *ciart* (circadian-associated transcription repressor) and *tshr* (thyroid stimulating hormone receptor), also showed high genetic diversity figures (Table 3). The remaining loci were polymorphic in most populations (MAF > 0.01). No deviation from Hardy–Weinberg proportions were detected either per locus across populations or per population across loci, excluding Skagerak (SK), which showed a significant excess of heterozygotes for most of the polymorphic loci analyzed ( $P < 0.0023$ ). Interestingly, this population is located in

Gene name	Chrom	Start	REF	ALT	aa substitution	Selection criteria
<i>hamp</i>	1	14423573	A	T	N81Y	ES, AS, PD
<i>fga-like</i>	2	1113231	G	T	R537Q; A574S	Sma-E137 (GR, OUT) gene; 13831_88 (OUT farm) < 500 kb; PD; AS
<i>arhgap42</i>	3	7539292	C	A	T632K; T793N	1916_69 (OUT) gene; GR-OST
<i>vtna</i>	3	15402438	G	T	D185E	Sma-USC214 (GR, PD) < 500 kb; PD
<i>paxbp1</i>	4	10752137	G	A	P47L	ES; GR-OFS
<i>cmtm3</i>	5	10463315	T	C	K83R	5986_20 (OUT farm) < 500 kb; ES
<i>igf1rb</i>	5	22614990	A	G	Y980H	Sma-USC7 (GR) gene; GR -OFS
<i>hmox</i>	8	13184347	A	G	T81A	7560_71 (OUT farm) and 7235_80 (OUT farm) < 500 kb; PD; AS; ES
<i>ciart</i>	10	12731534	A	G	N271S	SmaUSC-E29 (OUT) gene; AS; ES; GR-OFS
<i>slc12a3</i>	10	24208615	G	C	D38N; C938S	ES; OR-OFS
<i>frs2</i>	10	25869534	G	A	A124T; A290P	SmaUSC-E7 (GR, OUT) gene;
<i>igfbp2</i>	14	4200650	G	A	P264S	PD; GR-OFS;
<i>ccnb1</i>	16	2499469	C	T	A390V	Sma-USC146 (OUT) < 500 kb; ES
<i>fgfr3</i>	17	2378412	G	C	P45R	Sma-USC30 (GR, PD) < 500 kb; ES
<i>hbaD</i>	18	9550815	G	A	A44T; V78I	GR-OFS
<i>aqp8b</i>	19	11066153	G	T	Q36H	AS, ES; OR-OFS;
<i>hgs</i>	19	13543170	C	A	P726T	SmaSNP_298 (GR, PD) < 500 kb; ES
<i>sstr3</i>	19	19851767	C	T	S414L	SmaSNP_192 (GR) < 500 kb; ES; GR-OFS;
<i>tshr</i>	20	5830531	T	A	L339Q	Sma-USC273 (GR, PD) < 500 kb; AS
<i>myb</i>	20	10176750	C	T	C4Y	Sma-USC38 (OUT, PD) < 500 kb; AS
<i>vipr1b</i>	21	2341297	T	C	N2D	Sma-E112 (OUT) < 500 kb; Sma-USC91 (GR; VHSV) < 500 kb; ES
<i>eya3</i>	22	5576482	G	T	R22L; S230G	ES; GR-OFS;

**Table 2.** Selection of non-synonymous variants in turbot (*Scophthalmus maximus*) candidate genes following functional criteria. REF/ALT: reference and alternative alleles; selection criteria: differentially expressed genes for resistance to *E. scophthalmi* (ES), *A. salmonicida* (AS) and *P. dicentrarchi* (PD); genetic markers associated with QTL for growth (GR) or resistance to the same pathogens (ES, AS, PD), and to outliers (OUT) for natural or artificial (farm) selection within the gene (gene) or at less than 500 kb from the gene (< 500 kb); other fish-studies (OST) including osmoregulation (OR) and growth (GR).

Gene	Na	H <sub>E</sub>	F <sub>ST</sub> (all)	P-value	F <sub>ST</sub> (wild)	P-value	F <sub>ST</sub> (farm vs wild)	P-value
<i>aqp8b</i>	1.19	0.0199	0.0273	0.1223	-0.0019	0.3827	<b>0.1090</b>	0.0281
<i>cmtm3</i>	1.94	0.1667	0.0065	0.0583	0.0013	0.0505	0.0025	0.0561
<i>eya3</i>	2	0.4259	<b>0.1124</b>	0.0124	<b>0.1105</b>	0.0136	<b>0.1501</b>	0.0239
<i>hamp</i>	2	0.3944	0.0306	0.2804	0.0274	0.2836	0.0327	0.3380
<i>hgs</i>	1.63	0.0439	-0.0094	0.1538	-0.0072	0.1980	-0.0157	0.0859
<i>igf1rb</i>	1.75	0.1352	0.0467	0.4232	0.0338	0.4606	0.1057	0.1005
<i>fga-like</i>	2	0.4238	<b>0.0690*</b>	0.1809	0.0741	0.1228	0.0943	0.1086
<i>arhbap42</i>	1.75	0.0764	0.0065	0.1038	0.0045	0.1143	0.0175	0.2533
<i>hmox</i>	2	0.2272	0.0249	0.2195	0.0246	0.2735	0.0170	0.1605
<i>ciart</i>	2	0.4572	0.0656	0.2193	0.0539	0.3240	0.0937	0.1270
<i>igfbp2</i>	1.19	0.0248	<b>0.0988</b>	0.0169	-0.0019	0.3827	<b>0.3184</b>	0.0047
<i>myb</i>	1.69	0.0667	0.0124	0.1580	0.0157	0.2383	0.0061	0.1268
<i>paxbp1</i>	2	0.3464	<b>-0.0061</b>	0.0111	<b>-0.0091</b>	0.0099	0.0177	0.1639
<i>slc12a3</i>	2	0.2656	0.0621	0.2561	0.0623	0.2125	0.0535	0.3941
<i>sstr3</i>	1.06	0.0025	-0.0048	0.3020	-0.0019	0.3827	-0.0116	0.1674
<i>tshr</i>	2	0.4004	<b>0.1399</b>	0.0015	<b>0.1386</b>	0.0016	<b>0.1376</b>	0.0310
<i>vipr1b</i>	2	0.4664	0.0233	0.1831	0.0220	0.2015	0.0146	0.1326
<i>hbaD</i>	1.63	0.0715	<b>0.1399</b>	0.0249	-0.0084	0.1423	<b>0.3821</b>	0.0042

**Table 3.** Genetic diversity and differentiation of non-synonymous allelic variants in wild and farm turbot (*Scophthalmus maximus*) populations. In bold are shown outlier loci due to divergent (high F<sub>ST</sub>) or stabilizing (low F<sub>ST</sub>) selection with ARLEQUIN at P < 0.01 (consistent) or P < 0.05 (italics, suggestive) with respect to neutrality, and with BAYESCAN (p < 0.05, highlighted with \*).

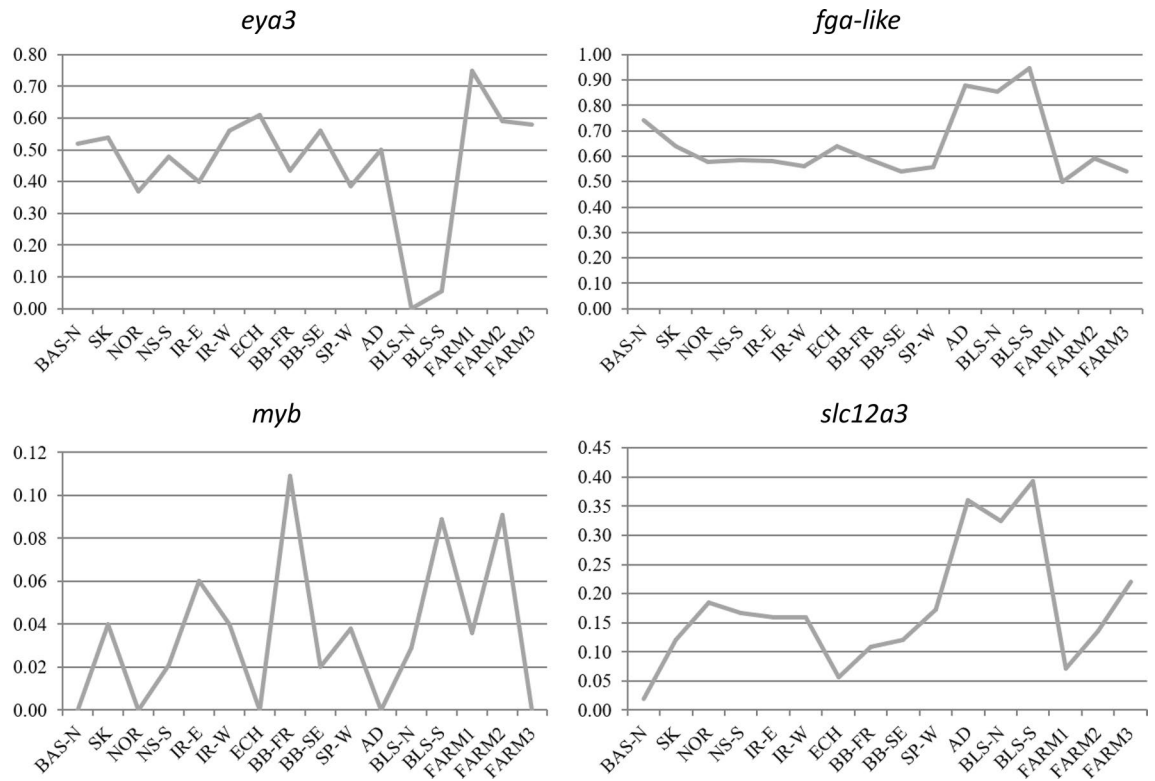
the transition between Baltic Sea and North Sea, where a contact between two highly divergent salinity environments occurs, depicting a rather complex hybridization area<sup>12</sup>.

Seven loci showed MAF < 0.1, among which *aqp8b* and *sstr3* (somatostatin receptor 3) showed rare allelic variants (MAF < 0.01). In fact, the *sstr3* locus was nearly fixed for one allelic variant across most populations, while the *igfbp2* (insulin-like growth factor binding protein 2b) and *aqp8* loci were polymorphic at MAF > 0.1 only in one population (Supplementary Table S5). At the other end, *eya3*, *vipr1b* (vasoactive intestinal peptide receptor 1b), *fga-like* and *ciart* were highly polymorphic (MAF > 0.3). Abrupt changes in allele frequencies at some genetic regions or related to the origin of samples (farm, wild) were observed. For instance, it was remarkable the polymorphism decay in the Black Sea of *eya3*, or the increasing/decreasing polymorphism in the southern populations for *slc12a3* (solute carrier family 12 member 3) and *hmox* (heme oxygenase), respectively (Fig. 3). Also, saw peaks showing the effects of genetic drift or sampling variance were observed in the least polymorphic loci, such as *igf1rb* (insulin-like growth factor 1b receptor), *myb* (v-myb avian myeloblastosis viral oncogene homolog) and *cmtm3* (CKLF-like MARVEL transmembrane domain containing 3). Finally, striking variation was also displayed when comparing farm samples between them or to the wild ones.

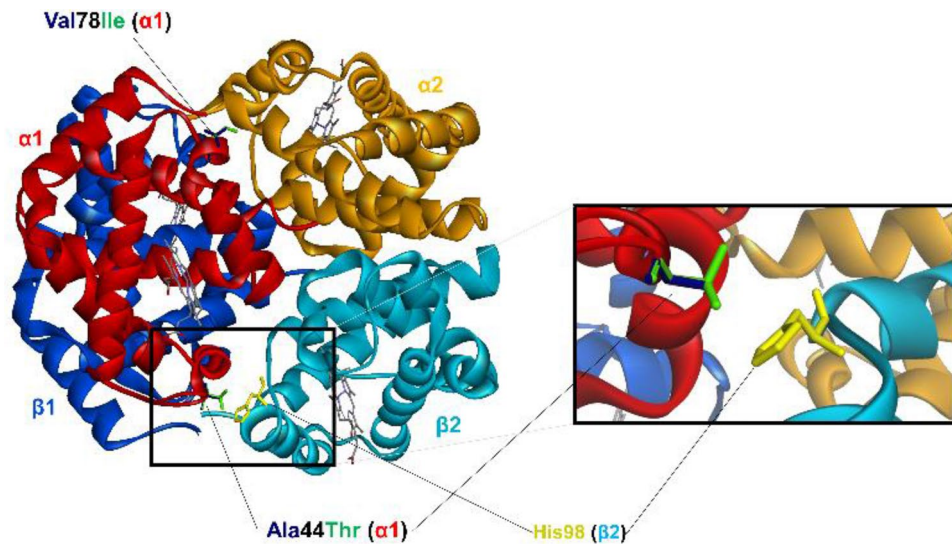
**Genetic differentiation and signals of selection.** We searched for signals of selection on the selected set of NSVs under three different scenarios: i) the 13 WILD populations; ii) ALL the 16 populations (13 wild and 3 farm); iii) comparing wild vs farm populations using a hierarchical approach (HIER). In all cases the set of neutral loci reported by do Prado et al.<sup>17</sup>, when analyzing wild populations, and by do Prado et al.<sup>18</sup>, when comparing wild vs farm populations, were used as the neutral background. A single locus, *fga-like*, which showed a significant decrease of genetic variation in the southern populations, was significant with BAYESCAN (Supplementary Fig. S1), but not with ARLEQUIN. Using the latter software, two loci showed signals of divergent selection, either consistent or suggestive (P < 0.01 and 0.05, respectively), in the three comparisons performed: *eya3* was nearly monomorphic in the Black Sea while at intermediate frequencies in the remaining populations; and *tshr* showed a progressive decrease in genetic diversity from the Baltic to the Black Sea, with an abrupt change in the Adriatic Sea, the only Mediterranean population studied. Another locus, *paxbp1* (PAX3 and PAX7 binding protein 1), showed signals of stabilizing selection in two scenarios (WILD, ALL), and close to significance in the third one (HIER). The comparison of wild and farm populations (HIER, ALL) unveiled significant signals of divergent selection for *igfbp2*, *aqp8b* and *hbaD*, which showed a rather similar pattern of differentiation, being monomorphic in nearly all wild populations while the alternative allele increased in two of the farms analyzed. Finally, locus *igf1rb*, although not significant, showed a notable differentiation (F<sub>ST</sub> = 0.1057) between wild and farm samples (HIER), reaching the highest frequencies of the alternative allele in the same two farms as *igfbp2*, *aqp8b* and *hbaD*.

**3D structural analysis of the non-synonymous variants with signals of selection.** The tetrameric deoxyhemoglobin structure of the polymorphic turbot HbaD and Hbβ1 subunits revealed that the Ala44αThr replacement occurs at the α<sub>1</sub>β<sub>2</sub> interface, which is involved in the allosteric transition of the protein (Fig. 4). The Ala44α variant shows a hydrophobic interaction with His98β that may stabilize the Hb tetramer and so deter-





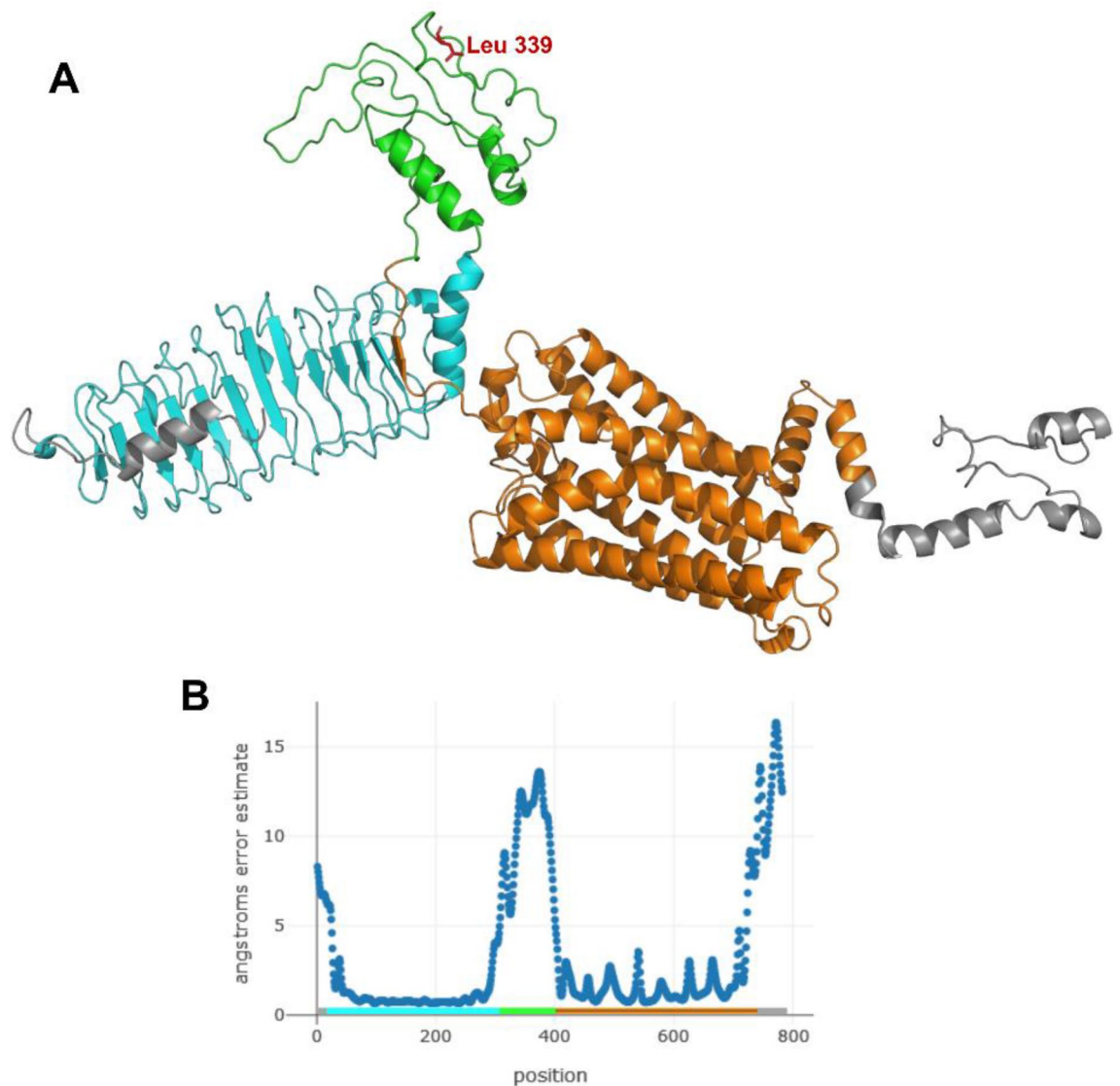
**Figure 3.** Allele frequency variation for some representative loci including non-synonymous variants evaluated across the whole distribution range and the main farm broodstock of turbot (*Scophthalmus maximus*).



**Figure 4.** Ribbon representation of the superimposed Ala44Thr and Val78Ile variants of turbot (*Scophthalmus maximus*) HbaD together with the Hb $\beta$ 1 subunit. Amino acid residues at positions 44 $\alpha$  and 78 $\alpha$  and the interacting His98 $\beta$  residue are shown in stick. The hydrophobic interaction between Ala44 $\alpha$ 1 and His98 $\beta$ 2 is shown in the enlarged section. The heme groups are shown color-coded by atom type.

mine a lower oxygen affinity, whereas the interaction is lost upon replacement of Ala with Thr. The conservative Val78Ile substitution does not affect the protein interfaces.

The absence of a suitable template in the PDB for homology modelling of TSHR, PAXBP1, EYA3 and IGFBP2 led us to generate 3D structures using I-TASSER and RoseTTAfold (Supplementary Table S6). The RoseTTAfold TSHR model showed a confidence score of 0.6 and the per-residue error estimate suggests that the Leu339Glu substitution is positioned in an unstructured region from position 298 to 404 (Fig. 5), corresponding to the hinge between the extracellular leucine-rich repeats and the seven-helix transmembrane domain. RoseTTAfold



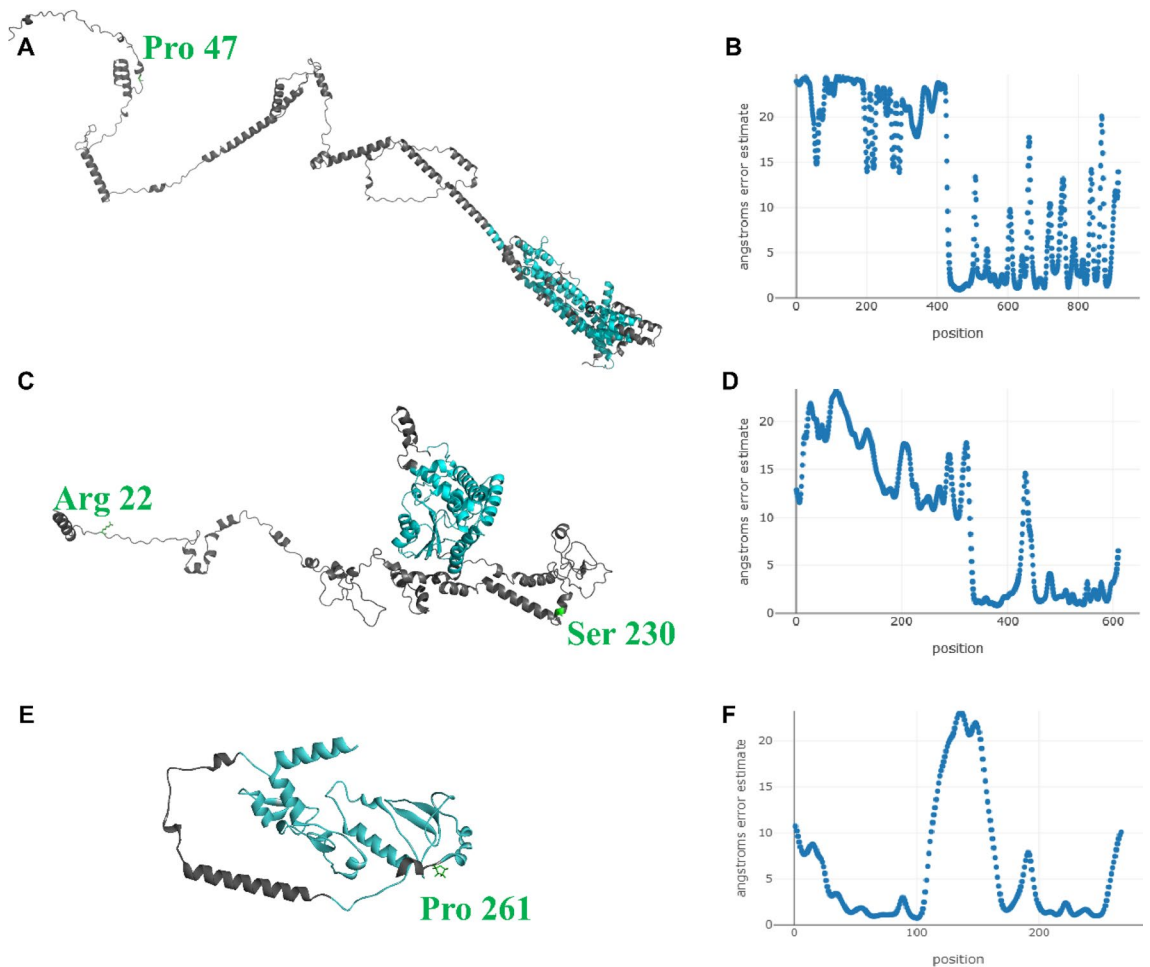
**Figure 5.** Structural model of turbot (*Scophthalmus maximus*) TSHR. (A) Cartoon representation of the hinge region (green) between the seven-parallel-helices domain (orange) and the leucine-reach-repeat domain (light blue). (B) Local quality of the model expressed in per-residue error estimate.

models of PAXBP1 and EYA3 were of low confidence (0.39 and 0.42 scores, respectively), while the IGFBP2 model showed a good confidence score of 0.66, but the C-terminal region containing the Pro261Ser mutation was of low-quality. Modeled structures and corresponding per-residue error estimate are shown in Fig. 6.

## Discussion

Non-synonymous variation plays an important role in evolution and local adaptation to the diverse environment experienced by species with broad distribution ranges<sup>73,74</sup> and has been profusely screened in humans, *Drosophila* and other model species<sup>75–78</sup>. The increasing genomic resources due to the lowering sequencing costs make it feasible to catch in a quick and cheap way a picture of existing NSVs to be further used to investigate its adaptive role<sup>79–81</sup>. Other sources of variation such as structural variants have been associated with adaptation of fish species in the wild<sup>82</sup>, even in flatfish<sup>83,84</sup>, but the relative importance of NSV and structural on adaptation is still a matter of debate and further studies are needed<sup>85</sup>.

Here, we report the first genome-wide collection of NSVs in the turbot, a flatfish species distributed all around the European coasts, where it experiences gradual and abrupt changes in temperature and salinity<sup>17</sup>. Our study is based on genome resequencing of 10 farm fishes (5 males and 5 females) originated after five generations of selection from breeders of Northeast Atlantic Ocean, the most important region of turbot distribution, thus representing a preliminary picture of NSV in the genome of the species. However, it should be noted that expected heterozygosity was higher in farm than in wild samples, even from the Atlantic region, which shows a good management of genetic diversity in the breeding program. Since 10 diploid genomes were sequenced, our capacity to disclose low frequent and rare variants is limited, especially because filtering included a step for reliability related to MAF = 0.1 (at least two variants in the sample). Nonetheless, we could identify more than



**Figure 6.** Structural models of turbot (*Scophthalmus maximus*) PAXBP1 (A), EYA3 (C) and IGFBP2 (E) as calculated by RoseTTAfold. Cartoon representation colored by local model quality: low-quality and high-quality in grey and light blue, respectively. The local quality of the three models, expressed in per-residue error estimate, is shown in panel (B) (PAXBP1), (D) (EYA3) and (F) (IGFBP2).

50,000 NSVs across the ~21,500 protein coding genes annotated in the turbot genome, which will likely increase when a broader sample including the four main genetic regions identified across its distribution range<sup>17</sup> is explored. However, since most genetic diversity in the turbot is contained within populations (global NE Atlantic  $F_{ST} = 0.002^{ns}$ ; global distribution  $F_{ST} \sim 0.090^{17}$ ), our small sample from Northeast Atlantic Ocean would include a significant representation of NSV of the species. After removing those NSVs from putative pseudogenes and those representing non-sense mutations, a set of ~10,000 NSV was retained constituting the most reliable set in our study. Considering the expected frequency of NSVs in our small sample ( $MAF = 0.1$ ) and the high turbot effective population size (usually  $N_e > 10,000$  in the Atlantic Ocean region<sup>17</sup>), it could be assumed that most of this variation is not strongly detrimental and in fact, a very minor proportion of variants were homologous to deleterious mutations in other species. Previous studies on allozyme variation in the turbot supported much lower variation for this fraction of protein coding genes than in other flatfishes (~fivefold lower), unlike the very similar diversity observed with microsatellites, which was interpreted as an ancient bottleneck in this species<sup>11</sup>. If this observation could be extrapolated to all protein coding genes, this would mean that a much higher NSV would occur in other flatfish, which is supported by the ~10 million SNPs detected in Senegalese sole vs ~3 million SNPs in turbot obtained from the recent whole genome resequencing of 12 sole individuals<sup>86</sup>.

The broad NSV collection identified in the turbot was filtered using technical, population genetics and functional information to obtain a consistent database that could be further validated and eventually used with practical purposes on breeding programs and management of wild fisheries. We were very conservative to retain NSVs on functional genes, and those genes with  $\geq 3$  NSVs were dismissed, which dramatically dropped NSVs to ~10,000. We are aware that this filtering is likely very strict and a significant quantity of genes with  $\geq 3$  NSVs could be functional, so the whole  $\geq 50,000$  should be considered as a suggestive repository for future studies. Pseudogene identification in the turbot genome using the vast functional information coming from the AQUA-FAANG project<sup>87</sup> will improve our ability to discriminate pseudogenes, resulting in a more refined list of NSVs. The second most important drop (from ~10,000 to ~1200) was related to previous functional information (differentially expressed genes in response to pathogen challenges or growth) or association (close to QTL for growth and resistance to pathologies) studies in farm populations<sup>28,31</sup> or with signals of selection related to environmental

variables (temperature, salinity) in the wild across its distribution range<sup>30</sup>. The broad genomic information in turbot facilitated the targeting of this subset of NSVs on candidate genes under selection. However, the greater relevance of resistance to pathologies and growth for industry determined a bias in the final selection. Our list includes other interesting genes potentially related to adaptation in the wild (i.e. eight opsin genes very relevant for adaptation to the sea bottom<sup>88</sup>) to be explored in future studies. From this collection, a small subset of NSVs was validated using the MassARRAY genotyping technology on representative wild and farm samples trying to obtain some clues on their relevance for adaptation across its distribution range or in breeding programs. All the 18 SNPs finally genotyped in a single multiplex matched with the in silico predictions supporting the confidence of our pipeline and showed a very robust genotyping with hardly missing data, which makes feasible its further application as a cost-effective molecular tool.

We intended to identify signals of selection in this NSV set, either divergent or stabilizing, in the different scenarios studied using wild and farm populations covering the whole population range of the species and broodstock from companies with breeding programs, respectively. The joint analysis of loci under selection would blur/mask potential population structuration considering the different evolutionary forces<sup>89</sup>, which include clinal, patchiness or local variation patterns involving balanced or divergent selection models. However, locus-specific patterns of spatial variation were observed in the wild, as expected given the environmental variation (both biotic and abiotic) across the turbot distribution range. The most consistent patterns of turbot spatial structure were related to differentiation of the Southern populations at several loci (*fga-like*, *slc12a3*) or specifically in the Black Sea (*eya3*, *hamp*) or the Adriatic Sea (*tshr*), but gradual changes in the Atlantic from the Baltic Sea east and southwards (*ciart*, *LOC118312496*) and very particular local patterns such as *virp1*, were also observed. At the other end, *paxbp1* showed a great constancy across the whole distribution area. Interestingly, some of these gene markers have been associated with strong genetic differentiation at spatial scale in other fish species, like *slc12a3* and *tshr* related to osmoregulation and variation in spawning time, respectively, between Atlantic and Baltic herring<sup>90,91</sup>. Signals of selection for some of these genes has also been reported in other fish species across geographical ranges, such as *paxbp1* linked to myogenesis and thermogenesis<sup>92</sup>, or *virp1* associated with local adaptations to extreme environments<sup>93</sup>.

In addition to *eya3* and *tshr*, outlined before, three other loci, *aqp8b*, *igfbp2* and *hbaD*, showed consistent or suggestive signals of selection when comparing wild vs farm populations. Of note, *igfbp2* and *hbaD* showed a very strong differentiation when comparing wild vs farm populations ( $F_{ST} > 0.3$ ) due to the increase of a rare allelic variant in the wild in both farms. This fact was not observed in farm 2, which could suggest different selective pressures, or alternatively, a founder effect in the farm 2 broodstock. Interestingly, farms 1 and 3 appeared to be genetically closer (average  $F_{ST} = 0.015$ ) with regard to farm 2 ( $F_{ST(1\ vs\ 2)} = 0.030$ ;  $F_{ST(2\ vs\ 3)} = 0.036$ ), either by historical connection or because similar management protocols or targets of selection are followed.

We looked for additional support to the signals of selection detected by analyzing the consequences of the NSVs detected on the 3D protein structure that could refine their function according to environmental variation. For this, the complementary approaches using protein models of related species and de novo models supported by artificial intelligence tools provided information on the putative action of selection on growth, circadian rhythm and osmoregulation related genes. Furthermore, we also explored functional changes of other NSVs regarding previous information in turbot or in other species to ascertain their putative role on adaptation not evidenced in our population genomics analyses.

IGF-I and IGF-II are important regulators of vertebrate growth and development, and their respective coding turbot genes display distinct expression patterns during metamorphosis<sup>94</sup>. The present turbot study revealed polymorphisms in both the IGF binding protein IGFBP2 and the receptor IGF1R. The binding proteins have a higher affinity for IGF than the receptors and can inhibit and/or enhance IGF actions depending on the physiological context<sup>95</sup>. Teleost fish possess multiple *igfbp* genes of which *igfbp2* encodes a growth inhibitory protein<sup>96</sup>. A polymorphism in the chicken *igfbp2* has been found to be associated with growth and body composition<sup>97</sup>. The Pro264Ser polymorphism of turbot IGFBP2 is positioned in the C-terminal domain, which in human IGFBP2 contributes to IGF-1 binding<sup>98</sup>. The C-terminus including Pro264 is highly conserved in teleost IGFBP2 and was monomorphic in all wild turbot populations examined, except for the Spanish west coast population and farm1 and farm3 that displayed the rare Ser264 variant. Most of the current turbot broodstock have originally been recruited from Spanish and French coasts<sup>18,99</sup>, which could explain the presence of the rare IGFBP2b variant in farmed turbot, but its presence could also be connected to selection for growth considering that this is the main target of breeding programs. Similarly, *igf1rb* showed the highest polymorphism in farm1 and farm3, while the alternative allele was missing in the Baltic Sea, Black Sea and Adriatic Sea. An *igf1rb* polymorphism was reported to be associated with growth traits in the freshwater goby *Odontobutis potamophila*<sup>100</sup>, and divergence and polymorphism analysis of *igf1ra* and *igf1rb* in the orange-spotted grouper (*Epinephelus coioides*) suggested their importance in growth regulation and breeding of this species<sup>101</sup>. Moreover, the involvement of *igf1rb* in growth during hypoxia was recently reported in a genome-wide association analysis of adaptation to oxygen stress in farmed Nile tilapia (*Oreochromis niloticus*)<sup>102</sup>. Our study revealed a very strong differentiation of the polymorphic HbaD subunit when comparing wild vs farm populations. We predict that the Thr44 variant identified in farm1 and farm3 increases the oxygen binding affinity similar to the human hemoglobin Kawachi (Pro44α → Arg) variant<sup>103</sup> of importance during hypoxic conditions.

PAXBP1 is involved in skeletal muscle formation by linking the transcription factors PAX3 and PAX7 on chromatin to regulate the muscle progenitor cells proliferation. The pathogenic human variant Arg538Cys underlies syndrome of global developmental delay and myopathic hypotonia<sup>104</sup>, while the significant of the Pro47Leu substitution in turbot PAXBP1 is unknown.

Turbot is an active visual predator and shows circadian cycles of locomotor and food anticipatory activities together with rhythmic expression of core circadian clock genes<sup>105</sup>. Among the polymorphic turbot genes displaying high allelic diversity, we identified *tshr*, *eya3* and *ciart*, which are involved in the regulation of circadian

and seasonal rhythms. TSHR plays an important role in seasonal reproduction through the conserved EYA3-TSH pathway<sup>106,107</sup>. Polymorphisms in herring TSHR were shown to contribute to the regulation of spring or autumn spawning<sup>36</sup>, while the Leu339Glu polymorphism in turbot TSHR is positioned in a flexible region. Such intrinsically disordered regions are common in eukaryotic proteins and important biological functions have been associated with them, such as flexible linker, cellular signal transduction, protein phosphorylation<sup>108,109</sup>. It has been observed that function can arise directly from the disordered state whereas in other cases their function originates from binding-induced folding promoted by other proteins or RNA, DNA molecules<sup>110</sup>. Evidence of EYA3 as an integrator of photoperiodic cues and nutritional regulation was recently found in Atlantic cod (*Gadus morhua*)<sup>111</sup>. The Ser230Gly substitution in turbot EYA3 is positioned in the PST (Pro-Ser-Thr)-rich domain necessary for transcriptional activity of *Drosophila* EYA<sup>112,113</sup>, while both Ser and Gly were identified at the corresponding site in various teleost. The circadian-associated transcriptional repressor CIART is involved in the eye regression of cave molly (*Poecilia mexicana*)<sup>114</sup>, whereas turbot *ciart* proved to be differentially expressed in freshwater- versus seawater-acclimated fish<sup>115</sup>. A missense polymorphism in pig *ciart* was reported to be associated with backfat thickness<sup>116</sup>. Both the Asn and Ser residues in the polymorphic position 271 of turbot CIART are found in other teleost.

The important role played by the kidney in the osmoregulatory response of turbot to low salinity has been examined by transcriptome analysis<sup>25,115</sup>. SLC12A3, or the Na + Cl-cotransporter NCC1 paralog, is highly expressed in the kidney of fish acclimated to freshwater and is crucial for the ion reabsorption in the collecting duct<sup>117</sup>. Turbot *slc12a3* showed highest polymorphic diversity in the Black Sea and Adriatic Sea, in contrast to the Baltic Sea. We noted that the Cys residue at position 938 in turbot *slc2a3* is novel among marine fish, except for Antarctic fish. *aqp8b* is highly expressed in fish kidney tubuli serving as important pathways for reabsorbed water<sup>118</sup>. Turbot *aqp8* was only polymorphic in the Spanish west coast population and in farm1 and farm3 as outlined before for *igfbp2*. The acidic Gln residue at position 36 is invariable in teleost AQP8, and the basic His replacement together with the novel Cys938 variant of Slc2a3 await further studies.

Turbot *vipr1b* showed high polymorphic diversity in both wild populations and farms examined, except in the Baltic Sea and Spanish west coast. A conserved role of the VIP neuropeptide in the immune system and inflammatory processes in olive flounder (*Paralichthys olivaceus*) was suggested by the significant changes in *vip* mRNA levels in spleen and head kidney when exposed to an artificial bacterial challenge by *Edwardsiella tarda*<sup>119</sup>. VIP binds to the N-terminal end of the receptor, which in turbot contains an Asn2Gln polymorphism. VIPR1 polymorphism has been linked to gastrointestinal dysmotility disorders in man<sup>120</sup>, but associated with reproductive traits in birds<sup>121</sup>. Two polymorphic hepcidins have been identified in turbot<sup>122</sup> of which *hep1* was highly polymorphic in all populations and farms examined, particularly in the Black Sea and farm 3. The Asn-81Tyr substitution is positioned in the mature peptide, but it does not seem to affect the conserved Cys residues as shown by the polymorphic *hep2*<sup>122</sup>. Both *hep1* and *hep2* possess antimicrobial activity and promote resistance against bacterial and viral infection, but the antimicrobial activities of *hep2* were significantly stronger than those of *hep1* in vitro and in vivo<sup>123</sup>. However, only *hep1* was upregulated after iron overloading that is consistent with the presence of a hypothetical iron regulatory sequence, which is lacking in *hep2*<sup>123</sup>.

## Conclusions

We constructed the first atlas of NSVs in the turbot genome and designed a conservative pipeline to define a robust dataset that could be further validated for their implication on adaptation in the wild or farm conditions using population genomics or 3D functional approaches. This strategy enabled the identification of consistent or suggestive signals of selection related to growth, osmoregulation, hypoxia or immunity that might be further applied for functional and association studies using a robust and cost-effective genotyping methodology. Our study does not only provides a suitable strategy for turbot, but it could be expanded to other fish species considering the increasing genomic resources available in public databases.

## Data availability

Resequencing data of five males and five females are available at NCBI databases BioProject PRJNA649485 (<https://www.ncbi.nlm.nih.gov/bioproject/649485>), accession number SRX8843737. Genotyping data used in this study is provided in Table S5 and the primer sets for SNP genotyping included in Table S4.

Received: 2 November 2022; Accepted: 10 February 2023

Published online: 21 February 2023

## References

- Ilker, E. & Hinczewski, M. Modeling the growth of organisms validates a general relation between metabolic costs and natural selection. *Phys. Rev. Lett.* **122**, 238101 (2019).
- Boltaña, S. *et al.* Influences of thermal environment on fish growth. *Ecol. Evol.* **7**, 6814–6825 (2017).
- Rosenfeld, J., Richards, J., Allen, D., Van Leeuwen, T. & Monnet, G. Adaptive trade-offs in fish energetics and physiology: Insights from adaptive differentiation among juvenile salmonids. *Can. J. Fish. Aquat. Sci.* **77**, 1243–1255 (2020).
- Robertson, D. R. & Collin, R. Inter- and intra-specific variation in egg size among reef fishes across the isthmus of Panama. *Front. Ecol. Evol.* **2**, 84 (2015).
- Zueva, K. J., Lumme, J., Veselov, A. E., Kent, M. P. & Primmer, C. R. Genomic signatures of parasite-driven natural selection in north European Atlantic salmon (*Salmo salar*). *Mar. Genom.* **39**, 26–38 (2018).
- Rajkov, J., El Taher, A., Böhne, A., Salzburger, W. & Egger, B. Gene expression remodelling and immune response during adaptive divergence in an African cichlid fish. *Mol. Ecol.* **30**, 274–296 (2021).
- Verhille, C. E. *et al.* Inter-population differences in salinity tolerance and osmoregulation of juvenile wild and hatchery-born Sacramento splittail. *Conserv. Physiol.* **4**, 1–12 (2016).

8. Froese, R. & Pauly, D. FishBase (version Feb 2018). In: Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist (Roskov Y. et al.). (2018). [www.catalogueoflife.org/annual-checklist/2019](http://www.catalogueoflife.org/annual-checklist/2019). ISSN 2405–884X.
9. Karås, P. & Klingsheim, V. Effects of temperature and salinity on embryonic development of turbot (*Scophthalmus maximus* L.) from the North Sea, and comparisons with Baltic populations. *Helgolander Meeresuntersuchungen* **51**, 241–247 (1997).
10. Barbut, L. et al. How larval traits of six flatfish species impact connectivity. *Limnol. Oceanogr.* **64**, 1150–1171 (2019).
11. Bouza, C., Presa, P., Castro, J., Sánchez, L. & Martínez, P. Allozyme and microsatellite diversity in natural and domestic populations of turbot (*Scophthalmus maximus*) in comparison with other Pleuronectiformes. *Can. J. Fish. Aquat. Sci.* **59**, 1460–1473 (2002).
12. Nielsen, E. E., Nielsen, P. H., Meldrup, D. & Hansen, M. M. Genetic population structure of turbot (*Scophthalmus maximus* L.) supports the presence of multiple hybrid zones for marine fishes in the transition zone between the Baltic Sea and the North Sea. *Mol. Ecol.* **13**, 585–595 (2004).
13. Vandamme, S. G. et al. Regional environmental pressure influences population differentiation in turbot (*Scophthalmus maximus*). *Mol. Ecol.* **23**, 618–636 (2014).
14. Vilas, R. et al. A genome scan for candidate genes involved in the adaptation of turbot (*Scophthalmus maximus*). *Mar. Genom.* **23**, 77–86 (2015).
15. Turan, C. et al. Genetics structure analysis of turbot (*Scophthalmus maximus*, Linnaeus, 1758) in the Black and Mediterranean Seas for application of innovative Management Strategies. *Front. Mar. Sci.* **6**, 740 (2019).
16. Ivanova, P. et al. Genetic diversity and morphological characterisation of three turbot (*Scophthalmus maximus* L., 1758) populations along the Bulgarian Black Sea coast. *Nat. Conserv.* **43**, 123–146 (2021).
17. do Prado, F. D. et al. Parallel evolution and adaptation to environmental factors in a marine flatfish: Implications for fisheries and aquaculture management of the turbot (*Scophthalmus maximus*). *Evol. Appl.* **11**, 1322–1341 (2018).
18. do Prado, F. D. et al. Tracing the genetic impact of farmed turbot *Scophthalmus maximus* on wild populations. *Aquac. Environ. Interact.* **10**, 447–463 (2018).
19. Robledo, D. et al. Integrating genomic resources of flatfish (Pleuronectiformes) to boost aquaculture production. *Comp. Biochem. Physiol. Part D Genom. Proteom.* **21**, 41–55 (2017).
20. Sánchez-Molano, E. et al. Detection of growth-related QTL in turbot (*Scophthalmus maximus*). *BMC Genomics* **12**, 473 (2011).
21. Rodríguez-Ramilo, S. T. et al. QTL detection for *Aeromonas salmonicida* resistance related traits in turbot (*Scophthalmus maximus*). *BMC Genom.* **12**, 541 (2011).
22. Robledo, D. et al. Integrative transcriptome, genome and quantitative trait loci resources identify single nucleotide polymorphisms in candidate genes for growth traits in turbot. *Int. J. Mol. Sci.* **17**, 243 (2016).
23. Sciarra, A. A. et al. Validation of growth-related quantitative trait loci markers in turbot (*Scophthalmus maximus*) families as a step toward marker assisted selection. *Aquaculture* **495**, 602–610 (2018).
24. Ma, A., Huang, Z., Wang, X. & Xu, Y. & Guo, X., Identification of quantitative trait loci associated with upper temperature tolerance in turbot, *Scophthalmus maximus*. *Sci. Rep.* **11**, 1–12 (2021).
25. Cui, W. et al. Comparative transcriptomic analysis reveals mechanisms of divergence in osmotic regulation of the turbot *Scophthalmus maximus*. *Fish Physiol. Biochem.* **46**, 1519–1536 (2020).
26. Martínez, P. et al. Identification of the major sex-determining region of turbot (*Scophthalmus maximus*). *Genetics* **183**, 1443–1452 (2009).
27. Martínez, P. et al. A genome-wide association study, supported by a new chromosome-level genome assembly, suggests sox2 as a main driver of the undifferentiated ZZ/ZW sex determination of turbot (*Scophthalmus maximus*). *Genomics* **113**, 1705–1718 (2021).
28. Martínez, P. et al. Turbot (*Scophthalmus maximus*) genomic resources: application for boosting aquaculture production. *Genomics in Aquaculture* (Elsevier Inc., 2016). <https://doi.org/10.1016/B978-0-12-801418-9.00006-8>.
29. Saura, M. et al. Disentangling genetic variation for resistance and endurance to scuticociliatosis in turbot using pedigree and genomic information. *Front. Genet.* **10**, 539 (2019).
30. Aramburu, O. et al. Genomic signatures after five generations of intensive selective breeding: Runs of homozygosity and genetic diversity in representative domestic and wild populations of turbot (*Scophthalmus maximus*). *Front. Genet.* **11**, 1–14 (2020).
31. Aramburu, O., Blanco, A., Bouza, C. & Martínez, P. Integration of host-pathogen functional genomics data into the chromosome-level genome assembly of turbot (*Scophthalmus maximus*). *Aquaculture* **564**, 739067 (2023).
32. Saul, M. C., Philip, V. M., Reinholdt, L. G. & Chesler, E. J. High-diversity mouse populations for complex traits. *Trends Genet.* **35**, 501–514 (2019).
33. Moen, T. et al. Epithelial cadherin determines resistance to infectious pancreatic necrosis virus in Atlantic salmon. *Genetics* **200**, 1313–1326 (2015).
34. Pavelin, J. et al. The nedd-8 activating enzyme gene underlies genetic resistance to infectious pancreatic necrosis virus in Atlantic salmon. *Genomics* **113**, 3842–3850 (2021).
35. Barson, N. J. et al. Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature* **528**, 405–408 (2015).
36. Chen, J. et al. Functional differences between TSHR alleles associate with variation in spawning season in Atlantic herring. *Commun. Biol.* **4**, 795 (2021).
37. Imsland, A. K., Brix, O., Nævdal, G. & Samuelsen, E. N. Hemoglobin genotypes in turbot (*Scophthalmus maximus* Rafinesque), their oxygen affinity properties and relation with growth. *Comp. Biochem. Physiol. A Physiol.* **116**, 157–165 (1997).
38. Imsland, A. K., Foss, A., Stefánsson, S. O. & Nævdal, G. Hemoglobin genotypes of turbot (*Scophthalmus maximus*): Consequences for growth and variations in optimal temperature for growth. *Fish Physiol. Biochem.* **23**, 75–81 (2000).
39. Andersen, Ø., Rubiolo, J. A., De Rosa, M. C. & Martínez, P. The hemoglobin Gly16β1 Asp polymorphism in turbot (*Scophthalmus maximus*) is differentially distributed across European populations. *Fish Physiol. Biochem.* **46**, 2367–2376 (2020).
40. Torrisi, M., Pollastri, G. & Le, Q. Deep learning methods in protein structure prediction. *Comput. Struct. Biotechnol. J.* **18**, 1301–1310 (2020).
41. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
42. AlQuraishi, M. Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).
43. Powder, K. E., Cousin, H., McLinden, G. P. & Craig Albertson, R. A nonsynonymous mutation in the transcriptional regulator *lbh* is associated with cichlid craniofacial adaptation and neural crest cell development. *Mol. Biol. Evol.* **31**, 3113–3124 (2014).
44. Lamichhaney, S. et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).
45. Gupta, A. M., Chakrabarti, J. & Mandal, S. Non-synonymous mutations of SARS-CoV-2 leads epitope loss and segregates its variants. *Microbes Infect.* **22**, 598–607 (2020).
46. Verde, C. et al. Structure, function and molecular adaptations of haemoglobins of the polar cartilaginous fish *Bathyraja eatonii* and *Raja hyperborea*. *Biochem. J.* **389**, 297–306 (2005).
47. Pearce, R. & Zhang, Y. Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* **297**, 100870 (2021).
48. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinf.* **9**, 40 (2008).
49. Pirolli, D. et al. Insights from molecular dynamics simulations: Structural basis for the V567D mutation-induced instability of zebrafish alpha-dystroglycan and comparison with the murine model. *PLoS ONE* **9**, e103866 (2014).

50. Lee, J., Freddolino, P. L. & Zhang, Y. From Protein Structure to Function with Bioinformatics. In *From Protein Structure to Function with Bioinformatics: Second Edition* (ed. Rigden, D. J.) (2017). <https://doi.org/10.1007/978-94-024-1069-3>
51. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a 3-track neural network. *Science* **373**, 871–876 (2021).
52. Castro, J. *et al.* Potential sources of error in parentage assessment of turbot (*Scophthalmus maximus*) using microsatellite loci. *Aquaculture* **242**, 119–135 (2004).
53. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv ID 1303.3997v2 **00**, 1–3 (2013).
55. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
56. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
57. Vera, M. *et al.* Development and validation of single nucleotide polymorphisms (SNPs) markers from two transcriptome 454-runs of turbot (*Scophthalmus maximus*) using high-throughput genotyping. *Int. J. Mol. Sci.* **14**, 5694–5711 (2013).
58. Ellis, J. A. & Ong, B. *The MassARRAY<sup>®</sup> system for targeted SNP genotyping*. *Methods in molecular biology* vol. 1492 (2017).
59. Choi, Y. & Chan, A. P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
60. Costello, M. J. Ecology of sea lice parasitic on farmed and wild fish. *Trends Parasitol.* **22**, 475–483 (2006).
61. Blanchet, S., Rey, O. & Loot, G. Evidence for host variation in parasite tolerance in a wild fish population. *Evol. Ecol.* **24**, 1129–1139 (2010).
62. Rousset, F. GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–106 (2008).
63. Foll, M. & Gaggiotti, O. A Genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics* **993**, 977–993 (2008).
64. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567 (2010).
65. Narum, S. R. & Hess, J. E. Comparison of  $F_{ST}$  outlier tests for SNP loci under selection. *Mol. Ecol. Resour.* **11**, 184–194 (2011).
66. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402 (1997).
67. Romero, P. *et al.* Sequence complexity of disordered protein. *Prot. Struct. Funct. Genet.* **42**, 38–48 (2001).
68. Jones, D. T. & Cozzetto, D. DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).
69. Mészáros, B., Erdős, G. & Dosztányi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucl. Acids Res.* **46**, W329–W337 (2018).
70. Ishida, T. & Kinoshita, K. PrDOS: Prediction of disordered protein regions from amino acid sequence. *Nucl. Acids Res.* **35**, W460–464 (2007).
71. Ito, N., Komiyama, N. H. & Fermi, G. Structure of deoxyhaemoglobin of the Anctartic fish *Pagothenia bernacchi* and structural basis of the root effect. *J. Mol. Biol.* <https://doi.org/10.2210/pdb1hbh/pdb> (1995).
72. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
73. Gou, X. *et al.* Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. *Genome Res.* **24**, 1308–1315 (2014).
74. Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713 (2013).
75. Macpherson, J. M., Sella, G., Davis, J. C. & Petrov, D. A. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics* **177**, 2083–2099 (2007).
76. Howe, D. G. *et al.* ZFIN, the Zebrafish model organism database: Increased support for mutants and transgenics. *Nucl. Acids Res.* **41**, 854–860 (2013).
77. Huber, C. D., Kim, B. Y., Marsden, C. D. & Lohmueller, K. E. Determining the factors driving selective effects of new nonsynonymous mutations. *Proc. Natl. Acad. Sci. USA* **114**, 4465–4470 (2017).
78. Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD<sup>®</sup>): Optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* **139**, 1197–1207 (2020).
79. Naruse, K., Hori, H., Shimizu, N., Kohara, Y. & Takeda, H. Medaka genomics: A bridge between mutant phenotype and gene function. *Mech. Dev.* **121**, 619–628 (2004).
80. Chintalapati, M. & Moorjani, P. Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev.* **62**, 58–64 (2020).
81. Rodin, R. E. *et al.* The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. *Nat. Neurosci.* **24**, 176–185 (2021).
82. Cayuela, H. *et al.* Thermal adaptation rather than demographic history drives genetic structure inferred by copy number variants in a marine fish. *Mol. Ecol.* **30**, 1624–1641 (2021).
83. Kess, T. *et al.* A putative structural variant and environmental variation associated with genomic divergence across the Northwest Atlantic in Atlantic Halibut. *ICES J. Mar. Sci.* **78**, 2371–2384 (2021).
84. Le Moan, A., Bekkevold, D. & Hemmer-Hansen, J. Evolution at two time frames: ancient structural variants involved in post-glacial divergence of the European plaice (*Pleuronectes platessa*). *Heredity (Edinb.)* **126**, 668–683 (2021).
85. Ruigrok, M. *et al.* The relative power of structural genomic variation versus SNPs in explaining the quantitative trait growth in the marine teleost *Chrysophrys auratus*. *Genes (Basel)* **13**, 1129 (2022).
86. De la Herran, R. *et al.* A chromosome-level genome assembly enables the identification of the follicle stimulating hormone receptor as the master sex determining gene in *Solea senegalensis*. *Mol. Ecol. Resour.* **00**, 1–19 (2023).
87. Harrison, P. W. *et al.* The FAANG data portal: Global, open-access, “FAIR”, and richly validated genotype to phenotype data for high-quality functional annotation of animal genomes. *Front. Genet.* **12**, 639238 (2021).
88. Figueras, A. *et al.* Whole genome sequencing of turbot (*Scophthalmus maximus*; Pleuronectiformes): A fish adapted to demersal life. *DNA Res.* **23**, 181–192 (2016).
89. Moore, J. S. *et al.* Conservation genomics of anadromous Atlantic salmon across its North American range: Outlier loci identify the same patterns of population structure as neutral loci. *Mol. Ecol.* **23**, 5680–5697 (2014).
90. Barrio, A. M. *et al.* The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *Elife* **5**, e12081 (2016).
91. Pettersson, M. E. *et al.* A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Res.* **29**, 1919–1928 (2019).
92. Bo, J. *et al.* Opah (*Lampris megalopsis*) genome sheds light on the evolution of aquatic endothermy. *Zool. Res.* **43**, 26–29 (2022).
93. Wang, S. *et al.* Resequencing and SNP discovery of Amur ide (*Leuciscus waleckii*) provides insights into local adaptations to extreme environments. *Sci. Rep.* **11**, 5064 (2021).
94. Meng, Z., Hu, P., Lei, J. & Jia, Y. Expression of insulin-like growth factors at mRNA levels during the metamorphic development of turbot (*Scophthalmus maximus*). *Gen. Comp. Endocrinol.* **235**, 11–17 (2016).

95. Duan, C., Ren, H. & Gao, S. Insulin-like growth factors (IGFs), IGF receptors, and IGF-binding proteins: Roles in skeletal muscle growth and differentiation. *Gen. Comp. Endocrinol.* **167**, 344–351 (2010).
96. Duan, C., Ding, J., Li, Q., Tsai, W. & Pozios, K. Insulin-like growth factor binding protein 2 is a growth inhibitory protein conserved in zebrafish. *Proc. Natl. Acad. Sci. USA* **96**, 15274–15279 (1999).
97. Furqon, A., Gunawan, A., Ulupi, N., Suryati, T. & Sumantri, C. A Polymorphism of Insulin-like growth factor binding protein 2 gene associated with growth and body composition traits in Kampung Chickens. *J. Vet.* **19**, 183 (2018).
98. Kibbey, M. M., Jameson, M. J., Eaton, E. M. & Rosenzweig, S. A. Insulin-like growth factor binding protein-2: Contributions of the C-terminal domain to insulin-like growth factor-1 binding. *Mol. Pharmacol.* **69**, 833–845 (2006).
99. Coughlan, J. P. *et al.* Microsatellite DNA variation in wild populations and farmed strains of turbot from Ireland and Norway: A preliminary study. *J. Fish Biol.* **52**, 916–922 (1998).
100. Zhang, H. *et al.* Characterization and Identification of Single Nucleotide Polymorphism within the IGF-1R gene associated with growth traits of *Odontobutis potamophila*. *J. World Aquac. Soc.* **49**, 366–379 (2018).
101. Guo, L., Yang, S., Li, M. M., Meng, Z. N. & Lin, H. R. (2016) Divergence and polymorphism analysis of IGF1Ra and IGF1Rb from orange-spotted grouper, *Epinephelus coioides* (Hamilton). *Genet. Mol. Res.* **15**, 1. <https://doi.org/10.4238/gmr15048768> (2016).
102. Yu, X. *et al.* Genome-wide association analysis of adaptation to oxygen stress in Nile tilapia (*Oreochromis niloticus*). *BMC Genomics* **22**, 426 (2021).
103. Harano, T. *et al.* Hemoglobin Kawachi [ $\alpha$ 44 (CE2) Pro  $\rightarrow$  Arg]: A new hemoglobin variant of high oxygen affinity with amino acid substitution at  $\alpha$ 1 $\beta$ 2 contact. *Hemoglobin* **6**, 43–49 (1982).
104. Alharby, E. *et al.* A homozygous potentially pathogenic variant in the PAXBP1 gene in a large family with global developmental delay and myopathic hypotonia. *Clin. Genet.* **92**, 579–586 (2017).
105. Ceinos, R. M. *et al.* Differential circadian and light-driven rhythmicity of clock gene expression and behaviour in the turbot, *Scophthalmus maximus*. *PLoS ONE* **14**, e0219153 (2019).
106. Nishiwaki-Ohkawa, T. & Yoshimura, T. Molecular basis for regulating seasonal reproduction in vertebrates. *J. Endocrinol.* **229**, R117–R127 (2016).
107. Wood, S. H. *et al.* Circadian clock mechanism driving mammalian photoperiodism. *Nat. Commun.* **11**, 4291 (2020).
108. Piovesan, D. *et al.* DisProt 7.0: A major update of the database of disordered proteins. *Nucl. Acids Res.* **45**, 219–227 (2017).
109. Pajkos, M. & Dosztányi, Z. Chapter Two - Functions of intrinsically disordered proteins through evolutionary lenses. in *Dancing Protein Clouds: Intrinsically Disordered Proteins in the Norm and Pathology. Part C* (ed. Uversky, V. N. B. T.-P. in M. B. and T. S.) vol. 183 45–74 (Academic Press, 2021).
110. Malagrino, F. *et al.* Understanding the binding induced folding of intrinsically disordered proteins by protein engineering: Caveats and pitfalls. *Int. J. Mol. Sci.* **21**, 3484 (2020).
111. Doyle, A., Cowan, M. E., Migaud, H., Wright, P. J. & Davie, A. Neuroendocrine regulation of reproduction in Atlantic cod (*Gadus morhua*): Evidence of Eya3 as an integrator of photoperiodic cues and nutritional regulation to initiate sexual maturation. *Comput. Biochem. Physiol. -Part A Mol. Integr. Physiol.* **260**, 111000 (2021).
112. Silver, S. J., Davies, E. L., Doyon, L. & Rebay, I. Functional dissection of eyes absent reveals new modes of regulation within the retinal determination gene network. *Mol. Cell. Biol.* **23**, 5989–5999 (2003).
113. Jin, M. & Mardon, G. Distinct biochemical activities of eyes absent during drosophila eye development. *Sci. Rep.* **6**, 23228 (2016).
114. McGowan, K. L., Passow, C. N., Arias-Rodriguez, L., Tobler, M. & Kelley, J. L. Expression analyses of cave mollies (*Poecilia mexicana*) reveal key genes involved in the early evolution of eye regression. *Biol. Lett.* **15**, 20190554 (2019).
115. Cui, W. *et al.* Transcriptomic analysis reveals putative osmoregulation mechanisms in the kidney of euryhaline turbot *Scophthalmus maximus* responded to hypo-saline seawater. *J. Oceanol. Limnol.* **38**, 467–479 (2020).
116. Mármol-Sánchez, E., Quintanilla, R., Cardoso, T. F., Jordana Vidal, J. & Amills, M. Polymorphisms of the cryptochrome 2 and mitoguardin 2 genes are associated with the variation of lipid-related traits in Duroc pigs. *Sci. Rep.* **9**, 9025 (2019).
117. Takvam, M., Wood, C. M., Kryvi, H. & Nilsen, T. O. Ion transporters and osmoregulation in the kidney of teleost fishes as a function of salinity. *Front. Physiol.* **12**, 664588 (2021).
118. Englund, M. B. & Madsen, S. S. The role of aquaporins in the kidney of euryhaline teleosts. *Front. Physiol.* **2**, 51 (2011).
119. Nam, B. H. *et al.* Identification and characterization of the prepro-vasoactive intestinal peptide gene from the teleost *Paralichthys olivaceus*. *Vet. Immunol. Immunopathol.* **127**, 249–258 (2009).
120. Paladini, F. *et al.* Age-dependent association of idiopathic achalasia with vasoactive intestinal peptide receptor 1 gene. *Neurogastroenterol. Motil.* **21**, 597–602 (2009).
121. Hosseinpour, L., Nikbin, S., Hedayat-Evrigh, N. & Elyasi-Zarringhabaie, G. Association of polymorphisms of vasoactive intestinal peptide and its receptor with reproductive traits of turkey hens. *South Afr. J. Anim. Sci.* **50**, 345–352 (2020).
122. Pereiro, P., Figueras, A. & Novoa, B. A novel hepcidin-like in turbot (*Scophthalmus maximus* L.) highly expressed after pathogen challenge but not after iron overload. *Fish Shellfish Immunol.* **32**, 879–889 (2012).
123. Zhang, J., Yu, L., Ping, L., Fei, M. & Sun, L. Turbot (*Scophthalmus maximus*) hepcidin-1 and hepcidin-2 possess antimicrobial activity and promote resistance against bacterial and viral infection. *Fish Shellfish Immunol.* **38**, 127–134 (2014).

## Acknowledgements

This study was supported by Consellería de Educación, Universidade e Formación Profesional from Xunta de Galicia (Grant No. ED481A2020/119), which additionally supported Oscar Aramburu PhD Thesis with a fellowship (Grant No.: ED481A-2020/119). The authors wish to thank the provision of DNA samples and population information used in this study of the EU AQUATRACE (No. 311920) project and to the Flanders Research Institute for Agriculture, Fisheries and Food (ILVO, Belgium).

## Author contributions

P.M. and Ø.A. designed the research. D.R., O.A., C.B. and J.A.R. performed the genomic analyses to construct N.S.V. atlas. M.P. and P.M. carried out the population genomics analyses. M.C.D.R., D.P. and B.R. performed protein modelling. P.M., Ø.A. and M.C.D.R. wrote the paper. All authors have revised and approved the submitted version.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-29826-z>.



**Correspondence** and requests for materials should be addressed to Ø.A. or P.M.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023