PLOS ONE

RESEARCH ARTICLE

A novel transcriptome-derived SNPs array for tench (*Tinca tinca* L.)

Girish Kumar[1☯]*, Jorge Langa[2☯], Iratxe Montes[2], Darrell Conklin[3,4], Martin Kocour[1], Klaus Kohlmann[5], Andone Estonba[2]

1 Research Institute of Fish Culture and Hydrobiology, South Bohemian Research Center of Aquaculture and Biodiversity of Hydrocenoses, Faculty of Fisheries and Protection of Waters, University of South Bohemia in Ceske Budejovice, Czech Republic, **2** Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country UPV/EHU, Leioa-Bilbao, Bizkaia, Spain, **3** Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Gipuzkoa, Spain, **4** IKERBASQUE, Basque Foundation for Science, Bilbao, Bizkaia, Spain, **5** Department of Aquaculture and Ecophysiology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

☯ These authors contributed equally to this work.
* girishkumar.nio@gmail.com

Citation: Kumar G, Langa J, Montes I, Conklin D, Kocour M, Kohlmann K, et al. (2019) A novel transcriptome-derived SNPs array for tench (*Tinca tinca* L.). PLoS ONE 14(3): e0213992. https://doi.org/10.1371/journal.pone.0213992

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: October 2, 2018

Accepted: March 5, 2019

Published: March 19, 2019

Data Availability Statement: Transcriptome has been uploaded to NCBI Transcriptome Shotgun Assembly Sequence Database and it is available at GenBank with accession number GFZX00000000.1. Polymorphic SNPs have been uploaded to EBI's European Variation Archive under the study accession number PRJEB23783.

Funding: This research was supported by projects CENAKVA and Reproductive and genetic approaches for fish biodiversity conservation and aquaculture (CZ02.1.01/0.0/0.0/16_025/0007370) funded by Ministry of Education, Youth and Sports

Abstract

Tench (*Tinca tinca* L.) has great economic potential due to its high rate of fecundity and long-life span. Population genetic studies based on allozymes, microsatellites, PCR-RFLP and sequence analysis of genes and DNA fragments have revealed the presence of Eastern and Western phylogroups. However, the lack of genomic resources for this species has complicated the development of genetic markers. In this study, the tench transcriptome and genome were sequenced by high-throughput sequencing. A total of 60,414 putative SNPs were identified in the tench transcriptome using a computational pipeline. A set of 96 SNPs was selected for validation and a total of 92 SNPs was validated, resulting in the highest conversion and validation rate for a non-model species obtained to date (95.83%). The validated SNPs were used to genotype 140 individuals belonging to two tench breeds (Tabor and Hungarian), showing low (F_{ST} = 0.0450) but significant (<0.0001) genetic differentiation between the two tench breeds. This implies that set of validated SNPs array can be used to distinguish the tench breeds and that it might be useful for studying a range of associations between DNA sequence and traits of importance. These genomic resources created for the tench will provide insight into population genetics, conservation fish stock management, and aquaculture.

Introduction

Tench (*Tinca tinca* L.) is a freshwater fish species within the *Cyprinidae* family that spawns and grows ideally at water temperatures of 20–29˚C [1, 2]. Its native distribution is Eurasia; however, due to human-mediated movement, tench can also be found in temperate and tropic freshwater regions across the globe [3]. Due to its attractive appearance and specific meat flavour, tench has relevant economic importance and is commonly used in aquaculture and

of the Czech Republic, and by the Genomic Resources Research Group from the Basque University System (IT558-10) funded by the Department of Education, Universities and Research of the Basque Government. JL is supported by the pre-doctoral program Education Department of the Basque Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

sport fishing [4]. For example, tench farming is a common aquaculture activity in Europe and has recently expanded to China [5]. All of these facts motivate the increase of its annual global aquaculture production [6] of about 1400 tons [7].

Tench has very interesting features that set the species apart from other members of the *Cyprinidae* and that have popularized tench as an experimental model [8]. These include: an unequivocal body colour, normally green to brown-green, with golden, blue and albinotic phenotypes also existing [9]; small and hardly visible scales deeply embedded into the dermis; obvious sexual dimorphism in pelvic fins [4], specific reproductive biology [1]; low incidence of viral and bacterial diseases but high susceptibility to some chemical compounds [10]; and monophyletic origin (all descendants of a common ancestor) within *Tinca* genus [11]. Genetics studies have also shown that tench is still a diploid species (2n = 48) [12], which is advantageous for some genetic studies, compared to many cyprinids that are polyploid species [13].

Genetic studies on tench have until now been based on allozymes [14, 15], microsatellites [16, 17], PCR-RFLP [18, 19] and sequence polymorphism of genes and DNA fragments [20–22]. These studies have revealed the existence of Western and Eastern phylogroups [6, 19]. Individuals from both phylogroups have undergone natural and human-aided hybridization and this has produced hybrids that appear in natural water bodies as well as in cultured stocks along Europe.

The rapid development and application of sequencing technologies is now permitting researchers to discover thousands of SNPs at relatively low cost compared to the traditional Sanger sequencing method [23]. Transcriptome sequencing is considered a cost-effective strategy for discovering SNPs in non-model species. In fact, as a transcriptome is directly associated with functional regions in a genome, transcriptome-derived SNPs can be informative for adaptive variation [24–26] and they can be used not only for assessing population genetic structure, but also for genomic selection for traits of interest to aquaculture such as growth, sex determination or disease resistance (e.g. [27–29]). Given these advantages, SNPs derived from transcriptomes have been widely discovered and studied in many fish species [29–42].

The aim of this study was: to discover and validate transcriptome-derived SNPs in *T. tinca*, based on the strategy designed by Montes *et al*. and successfully applied in other fish species [38, 43]. The SNPs array was then used to disentangle the population genetic structure of two cultured tench breeds (Tabor and Hungarian), previously identified as stocks representing mixture of haplotypes out of both phylogroups [22].

Materials and methods

Ethics statement

The handling and usage of experimental fish in this study was done in accordance with the Czech Act. No 256/1992 Coll. as amended under supervision of the Institutional Animal Care and Use Committee (IACUC) of the University of South Bohemia (USB), Faculty of Fisheries and Protection of Waters (FFPW) in Vodňany. The USB FFPW has approval of the Ministry of Agriculture of the Czech Republic for handling and usage of experimental animal's ref. no. 16OZ15759/2013-17214. The presented study was included in the planned activities dealing with study of biodiversity, genetic, physiological and reproductive variability and performance of selected freshwater fish species. The experimental stock was reared under the common semi-intensive pond management conditions. The fish sacrificed for the study were euthanized in accordance with the Ordinance no. 419/2012 Coll. as amended. The fish were euthanized by blow into the head using a blunt object and bleeding. One of the co-authors was present during handling and processing the fish owned the certificate (no. 0135/2000-V3) which allows him to conduct and manage experiments involving animals according to the above mentioned act.

Sample collection

In the methodology followed for SNP discovery, two samplings (corresponding to the two sequencing approaches) were performed; one for transcriptome sequencing, and another for genome sequencing.

For transcriptome sequencing, 4 tench individuals (2 males and 2 females) were sampled. The sampled individuals belonged to two metabolic activities (summer season with 20˚C water temperature, and winter season with 4˚C water temperature) and two breeds (Hungarian and Tabor) cultured in Vodňany, Czech Republic since 1990's [44] (present Faculty of Fisheries and Production of Waters, University of South Bohemia in České Budějovice). The Tabor breed was established by collecting fish from ponds of a Czech county, and the Hungarian breed by introducing the tench from Hungary. To increase the homozygosity, inbreeding and gynogenesis within each breed were applied. Both breeds, containing approximately 120 adult individuals, have been maintained to date by intra-linear mating only for 6 generations. Previous studies on these fish have shown that both breeds have gene pools mixed of both Western and Eastern phylogroups [22, 45]. The transcriptome changes according to genes expressed. Expression of various genes depends on many inner and outer factors (e.g. fish age, health status, phase of reproductive cycle, weather, season—growing or wintering etc.). That is why we sampled fish in winter (no-growing season) and summer (growing season) in order to cover different genes expressed in mature 4-year old fish. Each fish was humanely sacrificed and two different tissues were collected- whole brain (without pituitary) and back muscle (approx. 1 g) and immediately frozen in liquid nitrogen, and stored at -80˚C until RNA extraction was performed. We had eight initial tissue samples, though two (brain in both cases) were not suitable for sequencing due to RIN values below 8. The remaining six samples (two of them in duplicate) were used for library construction and Illumina sequencing (S1 Table).

For genome sequencing, a total of ten tench individuals from six different locations were collected (S2 Table) in order to cover maximal available genetic diversity, including phylogroup origin of tench species. Samples were taken from the tench tissue collection of Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany and they represented populations throughout Neighbor-joining trees inferred from studies focused on genetic diversity of the growth hormone (GH) gene [22], microsatellites [17] and mitochondrial DNA [18].

RNA and DNA extraction

Total RNA was isolated using Qiazol lysis reagent (Qiagen). The isolated RNA was quantified with a Nanodrop 2000 (Thermo Scientific) and integrity of RNA (RIN) was determined using an Agilent 2100 Bioanalyzer (Agilent Technologies). Samples with RIN values above 8 were used for RNA sequencing, and used for library construction and Illumina sequencing. According to the RNA quality standards, six samples were sequenced (S1 Table).

Genomic DNA was isolated from muscle, fin or blood samples using the peqGOLD Tissue DNA Mini Kit (Peqlab Biotechnologie) following manufacturer instructions. The quantity and quality of DNA was measured with Qubit 2.0 Fluorometer and 0.8% agarose gel electrophoresis. The DNA samples with concentrations $\geq$ 50 ng/μl, 260/280 ratios of 1.8–2.0 and clear high molecular weight bands on the gel were used for genome sequencing. An equimolar amount of total DNA was then pooled for the library preparation.

Library construction and Illumina sequencing

A multiplex sequencing library was prepared by labeling each sample (six RNA samples, two of them replicated; and two DNA pools) with specific 10-mer barcoding oligonucleotides.

Transcriptomic and genomic libraries were sequenced in a single lane of Illumina HiSeq2000 and HiSeq2500 platforms, respectively. Sequencing reactions were performed separately for transcriptome and genome with paired-end 101 bp and 126 bp reads, respectively. Sequencing was performed at CNAG- Centre Nacional d'Anàlisi Genòmica, Barcelona, Spain. All sequence data have been submitted to NCBI's submission portal under the BioProject accession number PRJNA414567.

Genome size estimation

We estimated the genome size of *Tinca tinca* by means of the frequencies of the kmers in the DNA reads. Reads were processed with Jellyfish 2.2.10 [46] using the *count* subcommand with a kmer size of 25. The frequencies were computed using the *histo* subcommand. Finally, the genomic haploid length, along with the repetitive and unique contents and rate of heterozygosity, was computed using the GenomeScope web service [47].

Transcriptome *de novo* assembly and annotation

Raw RNA reads were processed in a strict four-step procedure in order to obtain a high-quality reference. First, adaptors and low-quality reads were removed with Trimmomatic v0.33 [48] by deleting the first 13 nucleotides of the read. Removal of adapters was done with the ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 parameters by setting a minimum mean PHRED quality value of 10, trailing bases with quality value at least 20, and a minimum read length of 31 bases. Second, contaminants indicated by the UniVec database were removed with SeqClean (https://sourceforge.net/projects/seqclean/). Third, Trimmomatic was run on Single End mode to remove low quality and excessively short reads with the following parameters: minlen:31 avgqual:10 minlen:31 trailing:19 minlen:31 tophred33. Finally, the paired-end structure of the reads was recovered with a custom script written in Python with help of the Biopython package [49].

After the transcriptome reads were trimmed, paired and unpaired high-quality reads (all RNASeq data) were assembled into contigs using Trinity v2.0.6 [50]. The resulting transcriptome was uploaded to NCBI Transcriptome Shotgun Assembly Sequence Database and it is available at GenBank with accession number GFZX00000000.1. Full implementation of assembly procedure is available at https://github.com/GenomicResources/ttin_assembly.

To measure the quality of the assembled transcriptome, we used a two-fold approach. First we backmapped (with Bowtie2) the trimmed reads against the generated reference to measure the fidelity of the assembly with respect to the reads. According to the authors of Trinity, transcriptomes with mapping rates above 80% are considered good assemblies. Second, we used BUSCO v3.0.2 [51, 52] to assess the quality of the assembly by searching for *Actynopterygii* Single Copy Orthologs (SCOs). The program searches the homology between our transcriptome and a set of precomputed proteins that are known to be conserved across the evolution of a large set of species, classifying them as SCOs, conserved but duplicated, fragmented, or missing.

Finally, TransDecoder v2.0.1 (https://transdecoder.github.io/) and Trinotate (http://trinotate.github.io/) were used for transcriptome annotation and generation of a tench proteome. Transdecoder is a pipeline that extracts the possible open reading frames (ORFs) from a raw transcriptome to predict if it has homology with BLAST [53] against a protein reference database like Swiss-Prot [54] (downloaded on August 2015), UniRef90 [55] (accessed on August 2015), or homology via Hidden Markov Models with HMMER [56] (retrieved on August 2015) by querying the Protein Families database (Pfam, [57]. Once ORFs are called and possible homologies to elements in the different databases are hypothesized, a proteome is generated.

The next step in the procedure is the annotation of both the transcriptome and the predicted proteome developed as described above with Trinotate. It consists of homology searches, as done in the TransDecoder step, with help of BLASTX, BLASTP and HMMER, to then make use of a database (downloaded on August 2015) containing annotations from Gene Ontology (The Gene Ontology Consortium, 2000), KEGG [58], and eggnog [59].

Chimeras and duplicated regions were filtered out from the assembled transcriptome with stringent filters. First, contigs were quantified with Kallisto [60] and those with zero counts were removed with help of the Sleuth R package [61]. Additionally, according to the generated proteome, contigs with no coding potential were removed. Finally, genes that produce two or more isoforms were deleted. These procedures were performed using custom scripts in Python, R (R Core Team 2016), SAMtools [62] and Snakemake [63]. Implementation of the annotation procedure is available online at https://github.com/GenomicResources/ttin_trinotate. The resulting filtered transcriptome was used in the following steps of intron-exon boundary (IEB) prediction and SNP discovery.

SNP calling and IEB prediction

Tench SNP calling was performed as described by [38]. Two parallel SNP calling approaches were performed by aligning transcriptome (T2T) and genome (G2T) trimmed reads to the filtered transcriptome. This alignment was performed with Bowtie2 in local mode [64]. In this pipeline, PCR duplicates from both transcriptome and genome reads were removed using the SAMtools *rmdup* command [62]. Subsequently, variants were called with SAMtools *mpileup* command [62]. In order to avoid false SNPs, a maximum contig depth of 200x was set to avoid both repetitive sequences and false positive local alignments; the minimum contig depth allowed for T2T variants was 8x and 20x in the case of G2T variants in order to remove transcripts with low coverage that could bias the SNP calling procedure; the minimum variant count allowed for T2T variants was 2 high quality (HQ) bases (i.e., the alternative base appears at least twice), and 3 HQ bases for G2T variants. This last filtering step requires the SNPs to have higher MAFs when coverage is lower. After applying all of these filters, only common variants present in both T2T and G2T SNP discovery approaches were considered as putative SNPs. The implementation of the transcriptome filtering and SNP calling procedures is available online at https://github.com/GenomicResources/ttin_snps.

It is well known that genotyping procedures (for PCR based technology like fluidigm) will fail if primers are spanning or otherwise close to intron-exon boundaries [65]. Therefore, the filtered transcriptome reference was *in silico* assessed to detect IEBs as described by [66]. This is done by mapping genomic reads to the transcriptome, and computing p-values for *change points*. These are locations in the transcriptome where one or more genomic reads do not map throughout their whole length but rather the mapping is initiated or terminated internally to the read. Locations with low p-values represent a surprising number of change points at that location, hence a likely IEB. Predicted IEBs are annotated and avoided during genotyping primer design.

SNP genotyping and validation

A total of 140 tench samples belonging to two breeds (Tabor, N = 66 and Hungarian, N = 74) were genotyped for selected subset of 96 candidate SNPs. Only one SNP per contig was chosen and selection was not biased to any gene family. As growth-related traits are of main importance in most cultured fish species and growth hormone (GH) gene might be associated with growth [6], the SNP array was within each breed also associated with GH gene genotype distinguishing alleles of Eastern or Western phylogroup haplotype. Assignment of an individual to

Eastern (E) only, Western (W) only or hybrid (H) GH gene genotype was performed using the sequence analysis of GH gene [22]. In the pure Western GH gene genotype the first GH gene fragment including polymorphic side 1 (PS1) and the second GH gene fragment including PS7 were 344 bp and 451 bp long, respectively, while the individuals of pure Eastern GH gene genotype had fragments of 341 bp and 455 bp in length, respectively. In hybrids, haplotypes of both phylogroups were observed (i.e. 341 and 344 for PS1 and 451 and 455 for PS7). Flanking sequences of a subset of SNPs selected for validation were used for primers and probe design according to Fluidigm Genotyping System requirements. After genotyping, SNPs were categorized as *no signal* (unamplified SNPs), *disperse* (call rate < 80%), *monomorphic* (minor allele frequency, MAF < 0.01) and *psv* (paralogous sequence variant; all individuals are heterozygotes). For the conversion rate (proportion of all genotyped SNPs showing polymorphism), *no signal* and *disperse* SNPs were discarded, while only polymorphic SNPs (no *monomorphic*, neither *psv*) were used for the estimating the validation rate (proportion of polymorphic SNPs reliably scored in a sample of individuals). Polymorphic SNPs were uploaded to EBI's European Variation Archive under the study accession number PRJEB23783.

Population genetic structure

For each polymorphic SNP, minor allele frequency, and expected and observed heterozygosities (H_e and H_o, respectively) were estimated using the software package GeneClass2 [67]. Deviations from Hardy-Weinberg equilibrium (HWE) were evaluated for each *locus* using Fisher's exact test implemented in Genepop 4.0 [68] with 10,000 dememorizations, 100 batches and 5,000 iterations per batch.

To determine the genetic structure of tench individuals, genotype data were analyzed with Structure 2.3.4 software [69]. The number of clusters *k* was determined by comparing log-likelihood ratios in 10 runs for values of *k* between 1 and 10. Each run started with a burn-in period of 10,000 steps followed by 100,000 MCMC replicates. The optimal *k* was estimated as proposed by [69] and [70] and bar plots were generated using Pophelper v1.0.7 [71].

Based on this initial structure, the Bayesian likelihood method implemented in Bayescan 2.1 [72] was used to detect loci under natural selection (outlier loci). Bayescan was run with twenty pilot runs of 5,000 iterations, an additional burn-in of 50,000 iterations and prior odds of 10 for neutral model. Critical values were adjusted with a false discovery rate (FDR) procedure (q<0.1) [73]. Results of the outlier test were used to partition the SNP dataset into neutral and outlier loci; i.e., markers presumably under natural selection. Those loci resulting as outlier were removed from prospective analysis, regarding neutral variation, and annotations of the genomic regions including those loci were re-inspected.

Finally, neutral genetic differentiation and inbreeding were assessed. Neutral genetic differentiation was estimated with unbiased F_{ST} (distance matrix: pairwise difference) [74] using Arlequin v3.5 [75]. Inbreeding was estimated with F_{IS} [74] statistic using Fstat software [76]. The statistical significance of F_{ST} and F_{IS} was tested by 1,000 permutations for each pairwise comparison. In all cases with multiple comparisons, error rates were corrected using the sequential Bonferroni procedure [77].

Results

Transcriptome and genome sequencing

In total 32 million paired-end transcriptomic reads, with an average length of 101 bp, were sequenced (S3 Table). In the case of genome, 316 million genomic reads with a read length of 126 bp were generated, encompassing 154 million reads generated for Western pool (19.6 Gbp), and 162 million reads for Eastern pool (20.4 Gbp). GenomeScope estimated that the

Tinca tinca has a maximum genome size of 778,555,248 base pairs, where 599,234,146 base pairs (76.97%) constitute unique regions (S4 Table; S1 Fig). Overall, genome sequences constituted an estimated 51.58x coverage of the tench genome.

Transcriptome *de novo* assembly and annotation

Trimming of raw transcriptome reads did not result in a significant removal of reads, but 16% of nucleotides were discarded (S5 Table). The transcriptome *de novo* assembly consisted of 267,058 contigs (294.7 Mbp), which are the result of potentially 174,378 genes. The length of the assembled contigs ranged from 224 bp to 23,703 bp with an average length of 1,103 bp (S2 Fig).

Given the high number of sequences that Trinity yielded, we assessed the quality of our transcriptome by read mapping and by the contents of Single Copy Orthologs. On the one hand, the backmapping method achieved mapping success rates between 96.54% and 99.38% (S6 Table), suggesting therefore a good reconstruction of the *Tinca tinca* transcriptome. On the other hand, BUSCO reported that the transcriptome contains 85.9% of the Actinopterygii BUSCOs (where 40.4% are single copies), 6.7% are fragmented, and only 7.4% are completely missing (S7 Table). We conclude that given that even though we only sampled two tissues (muscle and brain) of *Tinca tinca*, this assembly is a good representation of the transcriptome.

According to the gene-isoform distribution in S3 Fig the distribution is skewed towards genes composed by one transcript. There are 10,705 genes of that composition (out of 174,378 genes, 86.42%, and out of 267,058 isoforms, 56.43%). The mean of the distribution is 1.53 transcripts per gene. As an extreme value, there is a gene (possibly a gene family) composed of 55 transcripts.

Regarding annotation, 89,832 transcripts were annotated (33.63%) as 126,187 proteins and 32,619 genes. From these, 64,676 transcripts (105,812 proteins and 9,295 genes) had a positive match to the UniRef90 database with *blastp* (S8 Table); similarly, 101,606 contigs (39,169 genes) were positively mapped with *blastx* (S9 Table). In both cases, top reference transcripts belonged to the same species: *Danio rerio*, *Astyanax mexicanus*, *Oncorhynchus mykiss*, *Oreochromis niloticus*, and *Ictalurus punctatus* (S4 and S5 Figs; S10 Table).

Overall, 67,953 contigs (77,626 proteins and 22,996 genes) were positively matched to 5,054 different protein domains according to the Pfam database (S6 Fig). The five most popular domains were: C2H2-type zinc finger (6.19%), Immunoglobulin domain (4.02%), Ankyrin repeat (3.22%), Leucine rich repeat (3.06%), and Zinc finger, C2H2 type (2.58%; S11 and S12 Tables).

According to the EggNOG database, 43,291 contigs (43,366 proteins and 14,714 genes) had a match against 3,338 different elements of the EggNOG database, including Serine Threonine protein kinase (7.63%), repeat-containing protein (3.03%), Zinc finger protein (2.95%), Ankyrin repeat (2.47%) and GTP-binding protein (1.27%) (S7 Fig and S13 Table).

Finally, Gene Ontology (GO) analysis showed 88,031 contigs (89,014 proteins and 30,345 genes). The highest number of GO terms was assigned to biological processes (48.63%) followed by molecular functions (29.66%) while cellular component has the least assigned terms (21.70%; S8 Fig). The three most commonly assigned GO terms in biological process category were genes involved in *Transcription*, *DNA-templated* (2.03%), *Regulation of Transcription*, *DNA-templated* (1.38%) and *Signal Transduction* (0.73%). In the molecular function ontology, *ATP binding* (5.77%), *Metal ion binding* (5.32%), *Zinc ion binding* (4.08%) and *DNA binding* (4.06%) were the most represented terms. The three major assigned GO terms for cellular component were nucleus (10.51%), cytoplasm (10.35%) and integral components of the membrane (7.26%; S9–S12 Figs; S14 and S15 Tables).

Table 1. Descriptive statistics of G2T, T2T and common discovered SNPs.

	G2T	T2T	Common
Contigs with SNPs	15,593	13,721	16,263
Number of contigs in filtered assembly	18,479	18,479	18,479
Transcripts with SNPs (%)	84.38	74.25	88.01
SNPs number	131,188	98,869	169,643
Assembly size (bp)	20,316,163	20,316,163	20,316,164
Mean mutation rate (SNPs/bp)	0.006	0.005	0.008
SNPs per transcript	8.41	0.14	0.10

https://doi.org/10.1371/journal.pone.0213992.t001

SNP discovery and validation

According to kallisto, a total of 262,801 contigs (out of 267,058) had an expression value above zero transcripts per million (TPM). Therefore 98.41% of the original assembly remained valid for further analysis. From those, 89,832 contigs were identified as having no coding potential and were discarded. Finally, contigs representing more than one isoform were also removed. After all these filters, the transcriptome was reduced to 18,479 contigs spanning 20.32 Mbp.

The filtered transcriptome was used as reference for mapping genome (G2T) and transcriptome (T2T) trimmed reads. The trimming process did not significantly decrease the number of transcriptome or genome reads (S16 and S17 Tables). The mapping process resulted in 19.51% of genomic reads and 22.63% of transcriptome reads assigned to the filtered transcriptome (S18 Table). From these mappings, a total of 131,188 G2T SNPs were called in 15,593 transcripts (8.41 G2T SNPs/transcript; Table 1), and 98,869 T2T SNPs were called in 13,721 transcripts (7.21 T2T SNPs/transcript). Together, G2T and T2T called 169,643 SNPs in 16,263 transcripts, but only 60,414 SNPs in 11,769 transcripts (5.13 SNPs/transcript) were common to both sets. These 60,414 SNPs represented the final set of putative SNPs discovered in the tench transcriptome.

Regarding IEB avoidance, 4,091 transcripts out of 18,479 were signaled as not having multi-mapped reads (those that map to more than one transcript); and a total of 2,937 transcripts contained one or more predicted IEB. A total of 16,764 IEBs were predicted (on average 5.70 predicted IEB per transcript). These predicted IEBs were annotated and avoided during genotyping primer design.

A set of 96 SNPs was selected based on IEB prediction analysis for validation and genotyping on Fluidigm Genotyping System. From the 96 SNPs that were genotyped, 4 (4.17%) were categorized as *no signal*, while the remaining 92 SNPs were *polymorphic* with >80% call rate. Therefore, conversion and validation rates of 95.83% were achieved.

Population genetics

Mean H_o and H_e for the Hungarian breed were 0.508 and 0.460, respectively. Similar levels of H_o (0.455) and H_e (0.458) were found in the Tabor breed. Tests of deviation from HWE for each locus revealed no significant departure from HWE after sequential Bonferroni correction. The STRUCTURE analysis evidenced population structure with K = 2 (Evanno method; Fig 1A), and K = 3 (Pritchard method; Fig 1B) being the most likely number of clusters. The average of the mean posterior probability (LnP(D)) estimated from 10 independent runs on K = 2 and K = 3 was -16533.7 and -16176.1, respectively. These clusters clearly indicate the differences between the two breeds, but not between the GH gene genotypes (Fig 1).

A total of six SNPs were detected as being under diversifying selection (positive alpha values); this is, they show extremely different allele frequencies in the two breeds. These outlier

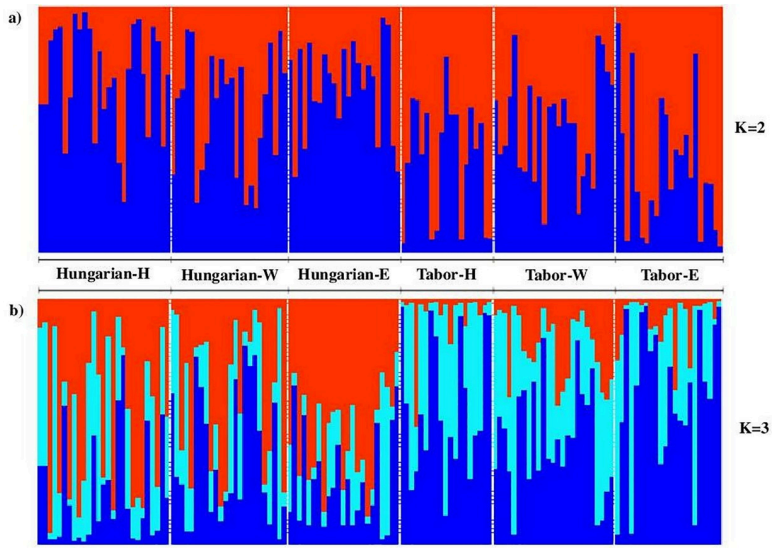

Fig 1. Results from STRUCTURE analysis for K = 2 (a) and K = 3 (b). Individuals corresponding to each breed (Hungarian, Tabor) and GH gene phylogroup genotype (H: Hybrid; W: Western; E: Eastern) are separated with vertical white bars.

https://doi.org/10.1371/journal.pone.0213992.g001

SNPs were located in the following genes: MRPL32 (39S ribosomal protein L32, mitochondrial), CENPF (Centromere Protein F), GRM1 (Glutamate Metabotropic Receptor 1), SPRY4 (Protein sprouty homolog 4), TRIP4 (Thyroid Hormone Receptor Interactor 4) and CN080 (Uncharacterized protein C14orf80 homolog) (Table 2). Of these six SNPs, all were found to be synonymous mutations, except interestingly for the SNP within *Activating signal cointegrator 1* (see Table 2) which encodes for either a Val (hydrophobic amino acid) or a Ser (polar

Table 2. Annotation of selected loci based top BLAST hit and GO ontology.

Locus ID	Genomic BLAST Hit	GO ID	e-value	Gene function
TR107177|c0_g1_i1	Sprouty homolog 4-like	GO:001602 GO:0021594GO:0030097GO:0040037GO:004874 GO:0070373	0.0E0	P: Negative regulation of fibroblast growth factor receptor signaling pathway; P: Rhombomere formation; P: Skeletal muscle fiber development; P: Hemopoiesis; P: Negative regulation of ERK1 and ERK2 cascade; C: Membrane
TR57930|c0_g1_i1	Centromere F	GO:0008134GO:0042803GO:0045502	0.0E0	F: Protein homodimerization activity; F: Transcription factor binding; F: Dynein binding
TR48380|c0_g1_i1	39S ribosomal L32, mitochondrial	GO:0005743GO:0005762GO:0003735GO:0016787GO:0006412	2.2 E-105	F: Structural constituent of ribosome; C: Mitochondrial large ribosomal subunit; C: Mitochondrial inner membrane; F: Hydrolase activity; P: Translation
TR71953|c0_g1_i1	Metabotropic glutamate receptor 1-like isoform X1	GO:0016020GO:0004871GO:0007165	1.1E-153	P: Signal transduction; C: Membrane; F: Signal transducer activity
TR96558|c0_g1_i2	Activating signal cointegrator 1	GO:0005634GO:0003713GO:0008270GO:0006366GO:0045893	0.0E0	C: Nucleus; F: Zinc ion binding; F: Transcription coactivator activity; P: Positive regulation of transcription, DNA-templated; P: Transcription from RNA polymerase II promoter
TR56671|c0_g1_i1	Uncharacterized protein C14orf80 homolog isoform X1	-	0.0E0	-

https://doi.org/10.1371/journal.pone.0213992.t002

amino acid). Since this substitutes a polar amino acid for a hydrophobic one, this SNP may lead to a change in protein function and should be further explored. Functional annotation revealed that most of these genes encoded proteins involved in transcription and translational regulation and structural organization of ribosome and mitochondria. Apart from these, the annotated gene Sprouty homolog 4-like was found to be involved in regulation of fibroblast growth and skeletal muscle fiber development, suggesting that the studied tench breeds might be adapted to different environments that affect growth related genes.

After removing the 6 outlier SNPs, a set composed of 86 SNP markers was used for studying neutral genetic differentiation and inbreeding. Pairwise F_{ST} estimates, within each breed, among E and W phylogroups and EW hybrid (H) were not significant; in contrast, all F_{ST} values were significant when pairwise comparisons between the two breeds were tested (Table 3). Overall, F_{ST} value between the two breeds was low but significant (F_{ST} = 0.0450, p-value < 0.0001). Additionally, F_{IS} within each breed was not significant, indicating homogeneity within breeds. In summary, individuals within breeds show homogeneous allele frequencies without regard to GH gene genotype, whereas individuals of the two breeds (even if they both are a mixture of E and W phylogroup haplotypes) are genetically different. Genotyping results of all 92 SNPs markers have also been included in the S19 Table.

Discussion

The major challenge of transcriptome-derived SNPs is marker "drop-out" during the validation step; the most significant factor is if a SNP spans an IEB. For instance, 64% of genotyping failures have been reported in EST-derived SNPs in catfish due to the proximity of SNPs to IEB [65]. The most evident cause for such genotyping failure is the presence of priming site at SNPs loci leading to non-base pairing of primers or expected amplification product is too large for amplification due to presence of intron between priming sites. Therefore, the key for successful SNP validation is avoidance of IEBs. In this study, the approach devised by [66] and applied successfully to European anchovy [38] was used to avoid the problem related to IEBs. In this method, the assembled transcript sequences were aligned to genome sequences of tench to identify the IEB. By selecting the SNPs not spanning an IEB, we obtained the highest conversion and validation rates of transcriptome-derived SNPs obtained to date for a non-model species.

In this study, using the validated SNPs we have demonstrated that the two tench breeds show low but significant genetic differentiation, even with their similar genetic structure concerning their phylogroup based gene pool. The ancestral populations that formed the two tench phylogroups separated about 0.064 to 1.6 million years ago as revealed from 1.6% sequence divergence of cytochrome b mitochondrial gene [21]. The western (W) and Eastern

Table 3. Pairwise F_{ST} (below diagonal) and p-values (above diagonal) among tench breeds (Hungarian, Tabor) and GH gene phylogroups genotype (H: Hybrid; W: Western; E: Eastern).

	Hungarian -H	Hungarian -W	Hungarian -E	Tabor-H	Tabor-W	Tabor-E
Hungarian-H	-	0.2022	0.1592	0.0000	0.0000	0.0000
Hungarian -W	0.0012	-	0.0429	0.0379	0.0000	0.0000
Hungarian -E	0.0025	0.0083	-	0.0787	0.0504	0.0000
Tabor-H	0.0619*	0.0000	0.0000	-	0.1973	0.3936
Tabor-W	0.0399*	0.0274*	0.0000	0.00538	-	0.0049
Tabor-E	0.0579*	0.0318*	0.0687*	0.00150	0.0218	-

* significant value

https://doi.org/10.1371/journal.pone.0213992.t003

(E) phylogroup significantly differs also in sequences of nuclear DNA e.g. the second intron of the actin gene, an intron of the gene coding for the ATP synthase β subunit, the first intron of the gene coding for the S7 ribosomal protein [21] and GH gene [6]. Due to the long history of tench phylogroup separation and individual evolution it is expected that the phylogroups would differ significantly also in physiological and biological functions resulting from nucleotide polymorphisms of functional genes. Therefore, our transcriptome-derived SNP array could be used for screening tench populations that still contain haplotypes of pure Western and pure Eastern phylogroup or F1 hybrid generation between pure W and E tench populations. Unfortunately, tench populations that bear pure Western haplotypes are very scarce or even absent [21] and we did not have such population in our collection. The Hungarian and Tabor breeds are, after several generations of mating fish with haplotypes of both phylogroups, a mosaic of both phylogroups due to free combination of chromosomes, crossing overs between homologous chromosomes and other possible processes that appear during formation of gametes. Based on F_{ST} values inferred from 86 SNPs it can be indirectly assumed that the SNPs genotypes were not significantly different for fish having Eastern, Western or hybrid GH gene genotype [22] within both Tabor and Hungarian breed. If the rate of phylogroup introgression within breeds were low, the degree of differentiation among fish displaying different GH gene genotype would be expected due to previously mentioned divergence between phylogroups in other genetic markers. On the other hand, significantly different F_{ST} values were observed between the two breeds with no matter to what GH gene genotype the fish belonged. The within-breed gene flow is corroborated by previous studies that show no negative fitness consequences derived from two phylogroup-mixed tench populations under cultured conditions [78]. In summary, six generations of within-breed isolated reproduction under cultured conditions allowed breed identity determination using the transcriptome-derived SNP array.

Moreover, apart from neutral levels of genetic differentiation, the SNPs in this study are transcriptome-derived markers and their variation in genes is informative for differential selection or adaptation in each breed. In this study, high allelic differentiation between both breeds was observed in growth-related genes, which might point to differential natural and human-affected selection, breeding and evolutionary history of Hungarian and Tabor tench breeds and/or stocks they were established from. Taking into account that the sequence of the GH (growth hormone) gene has 0.8% divergence in both tench phylogroups [6], we propose the following hypothesis: adaptive differences between breeds arise from differential composition of individuals from each phylogroup in each breed, giving to Hungarian and Tabor breeds different weight to their adaptation affecting growth related genes. However, further studies with protein sequencing of genes under selection are needed to corroborate the hypothesis presented here, as most of the SNPs found in the genes under selection have arisen due to synonymous mutations and will not lead to a change in the protein configuration. Insignificant association between GH gene genotype and SNP array also indicates that there is no linkage between our SNPs and the GH gene. However, this result does not say anything about association of these two markers to growth-related traits. It seems that effects of SNP array and GH gene genotype polymorphism on the growth-related traits will be (if any) independent of each other.

This study represents the first large-scale sequencing effort for SNP discovery and validation in tench. Although restriction-site associated DNA sequencing (RADseq) or double digest RADseq (ddRADseq) can generate large data set, SNPs derived from these approaches mostly fall into non-coding or unknown regions. Transcriptome derived SNPs are directly associated with functional regions in the genome and can give more information for 92 SNPs in coding region than hundreds or thousands of SNPs derived from non-coding or (not identified) regions. The validated SNPs can be used in further genetic studies for finding genes and/or DNA sequences associated with trait of importance.

Conclusions

The SNP discovery approach followed in the present study was developed for transcriptome-derived SNP discovery in European anchovy [38], and Atlantic mackerel [43] with successful conversion and validation rates. This approach can be used to discover large number of transcriptome-derived SNPs in any non-model species. In addition, our approach identifies SNPs in the transcriptome: these SNPs can be annotated and in some cases, as evidenced here, they are under natural selection. We showed that the SNPs array in tench is strong enough to distinguish tench breeds and that it might be useful for studies focused on searching the range of associations between DNA sequence and traits of importance. Overall, it was verified that transcriptome-derived SNPs may informs us not only about neutral genetic differentiation and population genetic structure (e.g. [37, 39]), but also about the functional role of the differences observed between populations or ecotypes

Supporting information

S1 Fig. GenomeScope profile.
(TIFF)

S2 Fig. Transcript-length distribution.
(TIF)

S3 Fig. Gene—Transcript distribution after TransDecoder prediction.
(TIF)

S4 Fig. Blastp hits distribution by organism. Organism is encoded by 5 letters (ASTMX *Astyanax mexicanus*, DANRE *Danio rerio*; ICTPU *Ictalurus punctatus*; LEPOC; ONCMY *Oncorryncus mykiss*; ORENI *Oreochromis niloticus*; POEFO; TAKRU *Takifugu rubripes*; TETNG; XIPMA).
(TIF)

S5 Fig. Blastx hits distribution by organism. Organism is encoded by 5 letters (ASTMX *Astyanax mexicanus*, DANRE *Danio rerio*; ICTPU *Ictalurus punctatus*; LEPOC; ONCMY *Oncorryncus mykiss*; ORENI *Oreochromis niloticus*; POEFO; TAKRU *Takifugu rubripes*; TETNG; XIPMA).
(TIF)

S6 Fig. Frequency distribution of Pfam domains (top 20).
(TIF)

S7 Fig. EggNOG distribution by ID (top 20).
(TIF)

S8 Fig. Gene Ontology summary.
(TIF)

S9 Fig. Joint GO level 2 distribution (top 30).
(TIF)

S10 Fig. Distribution of the Biological Process Gene Ontology terms.
(TIF)

S11 Fig. Distribution of the Cellular Component Gene Ontology terms.
(TIF)

S12 Fig. Distribution of the Molecular Function Gene Ontology terms.
(TIF)

S1 Table. Details of samples included in transcriptome sequencing.
(XLSX)

S2 Table. Details of samples included in genome sequencing.
(XLSX)

S3 Table. Transcriptome and genome sequencing results.
(XLSX)

S4 Table. Genome size estimation.
(XLSX)

S5 Table. Results obtained from the trimming performed in order to do transcriptome assembly.
(XLSX)

S6 Table. Backmapping of the reads.
(XLSX)

S7 Table. BUSCO.
(XLSX)

S8 Table. UniRef90 results from proteome querying (blastp).
(XLSX)

S9 Table. UniRef90 results from transcriptome querying (blastx).
(XLSX)

S10 Table. Summary of blastp results against UniRef90 by organism. Organism is encoded by 5 letters (ASTMX *Astyanax mexicanus*; DANRE *Danio rerio*; ICTPU *Ictalurus punctatus*; LEPOC; ONCMY *Oncorryncus mykiss*; ORENI *Oreochromis niloticus*; POEFO; TAKRU *Takifugu rubripes*; TETNG; XIPMA).
(XLSX)

S11 Table. Pfam hits.
(XLSX)

S12 Table. Pfam hits summarized by domain.
(XLSX)

S13 Table. EggNOG results.
(XLSX)

S14 Table. Summary of Gene Ontology results (level 1).
(XLSX)

S15 Table. Gene Ontology annotation of the transcriptome and proteome.
(XLSX)

S16 Table. Trimming for SNP discovery (DNA).
(XLSX)

S17 Table. Trimming for SNP discovery (RNA).
(XLSX)

S18 Table. Mapping.
(XLSX)

S19 Table. Genotyping results of all the 92 SNPs markers.
(XLSX)

Acknowledgments

The authors are thankful for the technical and human support provided by the Sequencing and Genotyping SGIker unit of UPV/EHU.

Author Contributions

Conceptualization: Martin Kocour, Andone Estonba.

Data curation: Girish Kumar.

Formal analysis: Girish Kumar, Jorge Langa, Iratxe Montes, Darrell Conklin.

Methodology: Girish Kumar.

Resources: Klaus Kohlmann.

Software: Girish Kumar, Jorge Langa, Darrell Conklin.

Supervision: Martin Kocour, Andone Estonba.

Writing – original draft: Girish Kumar.

Writing – review & editing: Girish Kumar, Jorge Langa, Iratxe Montes, Darrell Conklin, Martin Kocour, Andone Estonba.

References

1. Linhart O, Rodina M, Kocour M, Gela D. Insemination, fertilization and gamete management in tench, *Tinca tinca* (L.). Aquacult Int. 2006; 14(1–2):61–73.
2. Wolnicki J, Kaminski R, Sikorska J. Combined effects of water temperature and daily food availability period on the growth and survival of tench (*Tinca tinca*) larvae. Aquac Res. 2017; 48(7):3809–16.
3. Welcomme RL. International introductions of inland aquatic species. Rome: Food and Agriculture Organization of the United Nations; 1988. 318 p.
4. Kocour M, Gela D, Rodina M, Flajshans M. Performance of different tench, *Tinca tinca* (L.), groups under semi-intensive pond conditions: it is worth establishing a coordinated breeding program. Rev Fish Biol Fisher. 2010; 20(3):345–55.
5. Wang JX, Min WQ, Guan M, Gong LJ, Ren J, Huang Z, et al. Tench farming in China: present status and future prospects. Aquacult Int. 2006; 14(1–2):205–8.
6. Kocour M, Kohlmann K. Growth hormone gene polymorphisms in tench, *Tinca tinca* L. Aquaculture. 2011; 310(3–4):298–304.
7. FAO. Fishery Statistical Collections, Global aquaculture production 2017 April 20, 2017.
8. Flajshans M, Gela D, Kocour M, Buchtova H, Rodina M, Psenicka M, et al. A review on the potential of triploid tench for aquaculture. Rev Fish Biol Fisher. 2010; 20(3):317–29.
9. Kvasnicka P, Flajshans M, Rab P, Linhart O. Inheritance studies of blue and golden varieties of tench (Pisces: *Tinca tinca* L.). Journal of Heredity. 1998; 89(6):553–6.
10. Svobodova Z, Kolarova J. A review of the diseases and contaminant related mortalities of tench (*Tinca tinca* L.). Vet Med-Czech. 2004; 49(1):19–34.
11. Chen WJ, Mayden RL. Molecular systematics of the Cyprinoidea (Teleostei: Cypriniformes), the world's largest clade of freshwater fishes: further evidence from six nuclear genes. Molecular phylogenetics and evolution. 2009; 52(2):544–9. https://doi.org/10.1016/j.ympev.2009.01.006 PMID: 19489125
12. Arslan A, Taki FN. C-banded karyotype and nucleolar organizer regions of *Tinca tinca* (Cyprinidae) from Turkey. Caryologia. 2012; 65(3):246–9.

13. Leggatt RA, Iwama GK. Occurrence of polyploidy in the fishes. Rev Fish Biol Fisher. 2003; 13(3):237–46.
14. Šlechtová V, Šlechtá V., Valenta M. Genetic protein variability in tench (*Tinca tinca* L.) stocks in Czech Republic. Polish Archives of Hydrobiology. 1995; 42:133–40.
15. Kohlmann K, Kersten P. Enzyme variability in a wild population of tench (*Tinca tinca*). Polish Archives of Hydrobiology. 1998; 45:303–10.
16. Kohlmann K, Kersten P, Flajshans M. Comparison of microsatellite variability in wild and cultured tench (*Tinca tinca*). Aquaculture. 2007; 272:S147–S51.
17. Kohlmann K, Kersten P, Panicz R, Memis D, Flajshans M. Genetic variability and differentiation of wild and cultured tench populations inferred from microsatellite loci. Rev Fish Biol Fisher. 2010; 20(3):279–88.
18. Lo Presti R, Kohlmann K, Kersten P, Gasco L, Lisa C, Di Stasio L. Genetic variability in tench (*Tinca tinca* L.) as revealed by PCR-RFLP analysis of mitochondrial DNA. Ital J Anim Sci. 2012; 11(1).
19. Lajbner Z, Kotlik P. PCR-RFLP assays to distinguish the Western and Eastern phylogroups in wild and cultured tench *Tinca tinca*. Molecular ecology resources. 2011; 11(2):374–7. https://doi.org/10.1111/j.1755-0998.2010.02914.x PMID: 21429147
20. Lo Presti R, Kohlmann K, Kersten P, Lisa C, Di Stasio L. Sequence variability at the mitochondrial ND1, ND6, cyt b and D-loop segments in tench (*Tinca tinca* L.). J Appl Ichthyol. 2014; 30:15–21.
21. Lajbner Z, Linhart O, Kotlik P. Human-aided dispersal has altered but not erased the phylogeography of the tench. Evolutionary applications. 2011; 4(4):545–61. https://doi.org/10.1111/j.1752-4571.2010.00174.x PMID: 25568004
22. Kocour M, Kohlmann K. Distribution of five growth hormone gene haplogroups in wild and cultured tench, *Tinca tinca* L., populations. J Appl Ichthyol. 2014; 30:22–8.
23. Metzker ML. Sequencing technologies—the next generation. Nature reviews Genetics. 2010; 11(1):31–46. https://doi.org/10.1038/nrg2626 PMID: 19997069
24. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nature reviews Genetics. 2003; 4(12):981–94. https://doi.org/10.1038/nrg1226 PMID: 14631358
25. Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004; 13(4):969–80. PMID: 15012769
26. Morin PA, Luikart G, Wayne RK, Grp SW. SNPs in ecology, evolution and conservation. Trends Ecol Evol. 2004; 19(4):208–16.
27. Bester-Van Der Merwe A, Blaauw S, Du Plessis J, Roodt-Wilding R. Transcriptome-wide single nucleotide polymorphisms (SNPs) for abalone (Haliotis midae): validation and application using GoldenGate medium-throughput genotyping assays. International journal of molecular sciences. 2013; 14(9):19341–60. https://doi.org/10.3390/ijms140919341 PMID: 24065109
28. Li S, Liu H, Bai J, Zhu X. Transcriptome assembly and identification of genes and SNPs associated with growth traits in largemouth bass (Micropterus salmoides). Genetica. 2017; 145(2):175–87. https://doi.org/10.1007/s10709-017-9956-z PMID: 28204905
29. Liao Z, Wan Q, Shang X, Su J. Large-scale SNP screenings identify markers linked with GCRV resistant traits through transcriptomes of individuals and cell lines in *Ctenopharyngodon idella*. Sci Rep. 2017; 7(1):1184. https://doi.org/10.1038/s41598-017-01338-7 PMID: 28446772
30. Ogden R. Unlocking the potential of genomic technologies for wildlife forensics. Molecular ecology resources. 2011; 11 Suppl 1:109–16.
31. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, et al. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular ecology resources. 2011; 11 Suppl 1:123–36.
32. Helyar SJ, Limborg MT, Bekkevold D, Babbucci M, van Houdt J, Maes GE, et al. SNP discovery using Next Generation Transcriptomic Sequencing in Atlantic herring (Clupea harengus). Plos One. 2012; 7(8):e42089. https://doi.org/10.1371/journal.pone.0042089 PMID: 22879907
33. Lamichhaney S, Martinez Barrio A, Rafati N, Sundstrom G, Rubin CJ, Gilbert ER, et al. Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proc Natl Acad Sci U S A. 2012; 109(47):19345–50. https://doi.org/10.1073/pnas.1216128109 PMID: 23134729
34. Xu J, Ji P, Zhao Z, Zhang Y, Feng J, Wang J, et al. Genome-wide SNP discovery from transcriptome of four common carp strains. Plos One. 2012; 7(10):e48140. https://doi.org/10.1371/journal.pone.0048140 PMID: 23110192
35. Zarraonaindia I, Iriondo M, Albaina A, Pardo MA, Manzano C, Grant WS, et al. Multiple SNP markers reveal fine-scale population and deep phylogeographic structure in European anchovy (*Engraulis*

encrasicolus L.). Plos One. 2012; 7(7):e42201. https://doi.org/10.1371/journal.pone.0042201 PMID: 22860082

36. Hess JE, Campbell NR, Close DA, Docker MF, Narum SR. Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. Mol Ecol. 2013; 22(11):2898–916. https://doi.org/10.1111/mec.12150 PMID: 23205767

37. Montes I, Iriondo M, Manzano C, Santos M, Conklin D, Carvalho GR, et al. No loss of genetic diversity in the exploited and recently collapsed population of Bay of Biscay anchovy (*Engraulis encrasicolus*, L.). Mar Biol. 2016; 163(5).

38. Montes I, Conklin D, Albaina A, Creer S, Carvalho GR, Santos M, et al. SNP discovery in European anchovy (*Engraulis encrasicolus*, L) by high-throughput transcriptome and genome sequencing. Plos One. 2013; 8(8):e70051. https://doi.org/10.1371/journal.pone.0070051 PMID: 23936375

39. Montes I, Zarraonaindia I, Iriondo M, Grant WS, Manzano C, Cotano U, et al. Transcriptome analysis deciphers evolutionary mechanisms underlying genetic differentiation between coastal and offshore anchovy populations in the Bay of Biscay. Mar Biol. 2016; 163(10).

40. Laconcha U, Iriondo M, Arrizabalaga H, Manzano C, Markaide P, Montes I, et al. New Nuclear SNP Markers Unravel the Genetic Structure and Effective Population Size of Albacore Tuna (T*hunnus alalunga*). Plos One. 2015; 10(6):e0128247. https://doi.org/10.1371/journal.pone.0128247 PMID: 26090851

41. Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang H, et al. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife. 2016;5.

42. Robledo D, Rubiolo JA, Cabaleiro S, Martinez P, Bouza C. Differential gene expression and SNP association between fast- and slow-growing turbot (*Scophthalmus maximus*). Sci Rep. 2017; 7(1):12105. https://doi.org/10.1038/s41598-017-12459-4 PMID: 28935875

43. Alvarez P, Arthofer W, Coelho MM, Conklin D, Estonba A, Grosso AR, et al. Genomic Resources Notes Accepted 1 June 2015–31 July 2015. Molecular ecology resources. 2015; 15(6):1510–2. https://doi.org/10.1111/1755-0998.12454 PMID: 26452560

44. Flajshans M, Linhart O, Slechtova V, Slechta V. Genetic resources of commercially important fish species in the Czech Republic: present state and future strategy. Aquaculture. 1999; 173(1–4):471–83.

45. Lajbner Z, Kohlmann K, Linhart O, Kotlik P. Lack of reproductive isolation between the Western and Eastern phylogroups of the tench. Rev Fish Biol Fisher. 2010; 20(3):289–300.

46. Marçais G, Kinsgford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics. 2011; 27(6):764–770. https://doi.org/10.1093/bioinformatics/btr011 PMID: 21217122

47. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017; 33(14):2202–2204. https://doi.org/10.1093/bioinformatics/btx153 PMID: 28369201

48. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England). 2014; 30(15):2114–20.

49. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford, England). 2009; 25(11):1422–3.

50. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011; 29(7):644–52. https://doi.org/10.1038/nbt.1883 PMID: 21572440

51. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 2018; 35(3):543–548.

52. Felipe A. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717

53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

54. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research. 2003; 31(1):365–70. PMID: 12520024

55. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics (Oxford, England). 2007; 23(10):1282–8.

56. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011; 39(Web Server issue):W29–37. https://doi.org/10.1093/nar/gkr367 PMID: 21593126

57. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic acids research. 2014; 42(Database issue):D222–30. https://doi.org/10.1093/nar/gkt1223 PMID: 24288371

58. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research. 2012; 40(Database issue):D109–14. https://doi.org/10.1093/nar/gkr988 PMID: 22080510

59. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic acids research. 2012; 40(Database issue):D284–9. https://doi.org/10.1093/nar/gkr1060 PMID: 22096231

60. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology. 2016; 34(5):525–7. https://doi.org/10.1038/nbt.3519 PMID: 27043002

61. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nature methods. 2017; 14(7):687–90. https://doi.org/10.1038/nmeth.4324 PMID: 28581496

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009; 25(16):2078–9.

63. Koster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics (Oxford, England). 2012; 28(19):2520–2.

64. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9 (4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

65. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, et al. Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics. 2008; 9:450. https://doi.org/10.1186/1471-2164-9-450 PMID: 18826589

66. Conklin D, Montes, I., Albaina, A., Estonba, A. Improved conversion rates for SNP genotyping of non-model organisms. International Work-Conference on Bioinformatics and Biomedical Engineering; Granada, Spain 2013. p. 127–34.

67. Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A. GENECLASS2: a software for genetic assignment and first-generation migrant detection. The Journal of heredity. 2004; 95(6):536–9. https://doi.org/10.1093/jhered/esh074 PMID: 15475402

68. Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular ecology resources. 2008; 8(1):103–6. https://doi.org/10.1111/j.1471-8286.2007.01931.x PMID: 21585727

69. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945–59. PMID: 10835412

70. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005; 14(8):2611–20. https://doi.org/10.1111/j.1365-294X.2005.02553.x PMID: 15969739

71. Francis RM. pophelper: an R package and web app to analyse and visualize population structure. Molecular ecology resources. 2017; 17(1):27–32. https://doi.org/10.1111/1755-0998.12509 PMID: 26850166

72. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008; 180(2):977–93. https://doi.org/10.1534/genetics.108.092221 PMID: 18780740

73. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met. 1995; 57(1):289–300.

74. Weir BS, Cockerham CC. Estimating f-statistics for the analysis of population structure. Evolution; international journal of organic evolution. 1984; 38(6):1358–70. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x PMID: 28563791

75. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources. 2010; 10(3):564–7. https://doi.org/10.1111/j.1755-0998.2010.02847.x PMID: 21565059

76. Petit E, Balloux F, Goudet J. Sex-biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters. Evolution; international journal of organic evolution. 2001; 55(3):635–40. PMID: 11327171

77. Rice WR. Analyzing tables of statistical tests. Evolution; international journal of organic evolution. 1989; 43(1):223–5. https://doi.org/10.1111/j.1558-5646.1989.tb04220.x PMID: 28568501

78. Kumar G, Kohlmann, K., Gela, D., Kocour, M. Phylogroup origin of Tench Tinca tinca L. has no effects on main performance parameters. Aquaculture Europe-14; San-Sebastian, Spain 2014.