GENALPIC:

CONSERVATION GENOMICS OF ENDANGERED ALPINE ICHTHYOFAUNA

Barbara Sofia Ilardo

A thesis submitted in partial fulfilment of the
requirements of Liverpool John Moores University
for the degree of Doctor of Philosophy.

This research project was carried out in collaboration with
Fondazione Edmund Mach, Italy.

November 2022

Table of contents

Thesis abstract

Genomic erosion due to hybridisation is a problem which many endemisms face, including Italian pike *Esox flaviae* and marble trout *Salmo marmoratus*. These important freshwater species, which naturally occur in the Italian peninsula, are threatened by decades of stocking with non-native commercial lines, namely European pike *Esox lucius* and brown trout *Salmo trutta*. While supportive breeding programmes are in place for these species, screening of suitable breeders is based on traditional genetic methods such as microsatellite marker genotyping, which may have limited power to distinguish hybrids. Moreover, such markers are often surveyed through alternative laboratory protocols that can yield inconsistent results amongst research groups and hinder integration of molecular data from different local surveys on a larger biogeographical scale. In this doctoral thesis, I apply high-density Single Nucleotide Polymorphisms (SNPs) to the study of population substructure and genomic landscape within Italian pike, and I develop a large SNP panel to be implemented in a future genotyping array for large-scale genetic monitoring of Italian marble trout populations. In particular, I identified more than 20 million high-quality SNP markers from whole genome sequencing. Using these data sets, I found evidence of introgressive hybridisation in both species and reported a diminished ability of microsatellites to detect hybridisation compared to SNPs. Analyses of population structure confirm that at least four genetic clusters are present within Italian pike, and that anthropogenic translocations between geographically isolated basins have taken place. Moreover, I unveil genomic adaptations in Italian pike concerning olfactory perception, immune response and metabolism, emphasising the need to preserve the adaptive potential of endemic species. In marble trout, I filtered and validated *in silico* a set of more than 8 million high-quality SNPs after removing pseudo-SNPs from paralogous regions of the salmonid genome and potentially introgressed allochthonous alleles. Switching from current microsatellite-based

screening to a SNP genotyping array would not only increase resolving power for hybrid detection but also produce faster results in a less invasive manner. Indeed, while current screening methods require up to a week during which time marble trout are confined in low density pools, genetic analyses with the proposed technology would reduce time of confinement by several days, resulting in lower stress levels and higher post-release survival rates in marble trout wild breeders. Findings from this study will greatly facilitate detection of genetic introgression from introduced European pike and brown trout into these Alpine endemics which will inform regional and national conservation practices.

Contributions

Liverpool John Moores University, LJMU

Fondazione Edmund Mach (Italy), FEM – hosting institution

Barbara Sofia Ilardo, BSI – PhD candidate

Prof. Richard Brown, RB – Director of Studies and Supervisor

Dr. Andrea Gandolfi – Supervisor

Dr. Diego Micheletti – Advisor

Prof. Hazel Nichols – Advisor

Matteo Girardi and Stefano Casari – Field and Laboratory technicians at FEM

This thesis describes computational statistical analyses carried out by BSI on genomic data collected by AG, MG and SC. All analyses were done by BSI under the supervision of RB, AG, and DM. Bioinformatic pipelines were designed and carried out by BSI under the supervision of DM. The doctoral thesis was written entirely by BSI, who made modifications and amendments to an initial thesis draft following discussions with and comments from RB, AG and DM. Funding for this doctoral project was provided by LJMU and FEM.

Acknowledgements

I am grateful to LJMU and FEM for the opportunity to conduct research into a topic that is very dear to me, conservation genomics.

I would like to thank my supervisors and advisors for their inestimable support, guidance and understanding throughout the entire doctoral process amidst a global pandemic and personal tribulations including, but not limited to, a mid-PhD ADHD diagnosis.

I would like to acknowledge the important role LJMU and FEM PhD communities have played in encouraging my personal and professional development. A shout out to my fellow colleagues and friends whom I had the chance to meet along the way.

I am grateful to my family for bearing with me through the highs and the lows, and I am extremely thankful to my partner, who got me through the finish line.

Por último, quisiera dedicar esta tesis doctoral a mi querida nonna, porque se lo prometí y tengo que cumplir!

Declaration

No portion of the research described in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Chapter 1: General Introduction and Background

In this introductory chapter, I first illustrate what the genetic identity of a species is and why it must be protected, with particular regard to the threat posed by genetic hybridisation. I thus define the biological, ecological and economic characteristics as well as conservation status of Italian pike *Esox flaviae* Lucentini et al., 2011 and marble trout *Salmo marmoratus* Cuvier 1817. Then, I provide a more in-depth explanation of the field of conservation genetics and the technological advances leading up to conservation genomics, along with examples of successful applications in other species. I proceed to describe how to recognise signatures of molecular adaptation along the genome, and I conclude by summarising the research aim and objectives of this study.

1.1. Preserving the genetic identity of native species

A major goal for conservation biologists and geneticists is to preserve biodiversity by understanding, monitoring and protecting the genetic identity of threatened species, in order to guarantee their persistence in time (Mousseau, Sinervo and Endler, 2000). To this end, genetic variability is paramount in maintaining viable populations capable of sustaining themselves numerically across generations, and preserving their fitness, resistance and resilience in the event of external environmental changes (Ellegren and Sheldon, 2008). Such a topic is of growing interest and concern as we enter an era of tangible biological, ecological, and climatic alterations (Chown et al., 2016; Waldvogel et al., 2020).

In the context of this doctoral thesis, two cases were investigated: that of Italian pike and marble trout. Both species naturally occur in Italy and are threatened by the introduction of European pike *Esox lucius* Linnaeus 1758, and Atlantic brown trout *Salmo trutta* Linnaeus 1758, respectively. Such non-native species have been used in stocking to support angling practices for decades (Welcomme, 1988; Lucentini et al., 2006, 2009, 2011; Bianco, 2013;

Pedreschi et al, 2013; Genovesi et al, 2014), resulting in local population displacement and hybridisation and placing the genetic identity of both Italian species at risk (Fumagalli et al., 2002; Meraner et al., 2009, 2012, 2013; Gandolfi et al., 2015, 2017). As a direct consequence of this, the loss of endemic genetic variants underlying molecular adaptations to the native species' ecological niche may lower the chances of survival both in stable and dynamic environmental conditions (Harrisson et al., 2014; Hoelzel et al., 2019).

Supportive breeding programmes have so far mitigated the damage by identifying pure native individuals from wild populations through traditional conservation genetics methods and retaining them as breeders for artificial reproduction before releasing them back into the wild (Lucentini et al., 2011; Gandolfi et al., 2017; Martínez-Páramo et al., 2017; Eisendle et al., 2019). However, genetic approaches currently in use may lack power to detect complex patterns of introgression and to distinguish between native, non-native and hybrid individuals (Gandolfi et al, 2017, 2019). This translates into a risk of perpetrating artificial selection, allowing undetected allochthonous genetic variants to spread within endemic populations and possibly eradicate valuable genomic adaptations to local ecosystems. A need arises for the implementation of newer, high-resolution technologies in the field of conservation genomics capable of assessing the genomic landscape of threatened species and shed light on the molecular adaptations that may be compromised by hybridisation.

1.2. Italian pike

Two pike species are present in Italy: European pike (*E. lucius*) and Italian pike (*E. flaviae)*. However, the latter has been only recently described as an endemic species of the Italian Padano-Veneto and Northern Apennines basins, based on morphological and molecular evidence (Lucentini et al, 2011). The former, on the other hand, has been identified as a non-native species in Italy that has been widely stocked for angling purposes and which poses a threat to the genetic integrity of *E. flaviae* through hybridisation (Lucentini

et al, 2009, 2011, Gandolfi et al, 2017). Bianco (2013) argued that the name *E. flaviae* is a junior synonym of *E. cisalpinus* (Bianco & Delmastro, 2011), but because these authors described *E. cisalpinus* on the basis of phenotype and not genotype, in this thesis I refer to our Italian pike samples as *E. flaviae* in a cautionary manner (see Gandolfi et al, 2015).

Not much information concerning the specific biology and ecology of *E. flaviae* is available yet, but some characteristics can be deduced from scientific literature that predates its identification as a separate taxon (Eschbach et al, 2021). Like other pike species, it is an ambush predator that relies greatly on its sense of smell (Sternberg, 1992; Klein & Aylesworth, 1983) and on aquatic macrophytes for shelter and camouflage (Craig, 1996). Pike play an essential role in shaping freshwater communities due to its high position on the trophic chain (Craig, 1996, 2008). Floodplains are crucial habitats as they provide a spawning habitat for pike, which have phytophilic deposition (Casselman & Lewis, 1996; Raat, 1988). In Italy, habitat loss has been a threat to pike due to land reclamation since Roman times (Ciabatti, 1968) as well as in more recent times (Veggiani, 1974).

Although studies have shown widespread allochthonous introgression within Italian pike (Lucentini et al, 2011; Gandolfi et al, 2017), its conservation status is still “data deficient” according to the national Red Lists of the Italian Committee of the International Union for Conservation of Nature (IUCN, http://www.iucn.it/scheda.php?id=-897331167). The population in Trasimeno Lake has been used as a reservoir to restock other areas (Lucentini et al, 2006, 2009), and as of recent years is involved in a supportive breeding programme to contrast introgression with exotic *E. lucius* through screening practices considering mainly morphological traits (see Fig. 1.1) to select breeders (Lucentini et al, 2009).

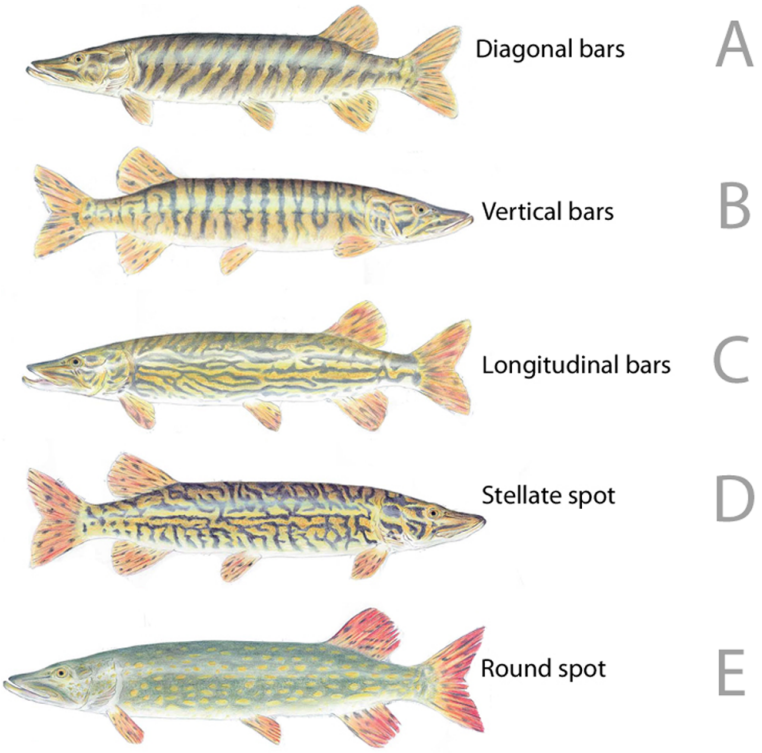


Fig. 1.1. Illustration from Lucentini et al. (2011) depicting phenotypic diversity across pike samples present in Italian freshwaters. A genetic association was found between phenotypes A through D and Italian pike, while round spots (E) were associated with European pike.

1.3. Marble trout

Marble trout (*Salmo marmoratus* Cuvier, 1817) is an important piscivorous species in the ecosystems it inhabits: the clean and cold waters of Alpine rivers and glacial lakes (Bianco, 1995). It is endemic to the Adriatic drainage system in Northern Italy, Southern Switzerland, Slovenia, Bosnia-Herzegovina and Montenegro (Bianco, 1995; Meraner & Gandolfi, 2017).

Fig. 1.2 Marble trout. Image source: https://www.pescafiume.it/trota-marmorata/

With its distinctive marbled colour pattern (Fig. 1.2), it is also a prized game fish for sport anglers. However, to support angling activities on trout through commercially accessible stocks, lakes and rivers have been stocked for decades with domestic lines of brown trout *Salmo trutta*, L. (Sommani, 1950; Tortonese 1970) which have led to genetic

erosion of marble trout by hybridisation (Caputo et al, 2004; Meraner et al, 2012; Gandolfi et al, 2019). To date, marble trout is listed as “Critically Endangered” on the IUCN Red List of Italian Vertebrates (Bianco et al, 2013).

Several efforts have been made towards its conservation, including cryopreservation of genetic material (Martínez-Páramo et al., 2017) and the establishment of “sanctuary” streams in Slovenia, where pure marble trout individuals are translocated to previously fishless watercourses (Martínez-Páramo et al, 2017; Crivelli et al, 2000). In Italy, supportive breeding programmes are in place, albeit locally, to facilitate reproduction of pure marble trout individuals following genetic screening using traditional conservation genetic methods such as microsatellite markers (Meraner et al, 2012; Meraner and Gandolfi, 2018; Eisendle et al, 2019). To ensure that marble trout populations are adequately preserved, accurate comprehension of population structure is necessary. Meraner and Gandolfi (2017) thoroughly summarise the current knowledge on this matter, highlighting the complex evolutionary history of the Salmo genus in Italy. Indeed, marble trout display pronounced genetic differentiation even at a micro-geopraphical scale (Fumagalli et al, 2002; Pujolar et al, 2011), emphasising the need to define management units (MU; Moritz, 1994) on the basis of genetic data.

1.4. Conservation Genetics

As its name suggests, conservation genetics is a branch of conservation biology aimed at studying and preserving biodiversity through population genetics. In particular, it seeks to investigate the dynamics of genetic variation in populations to ultimately prevent their extinction. Conservation actions vary depending on the nature of the threat a species might be facing, the most widespread case being declining effective population size due to climate change, habitat fragmentation and degradation and other anthropogenic threats.

The Italian wolf (*Canis lupus italicus*, Altobello 1921) and the Iberian lynx (*Lynx pardinus*, Temminck 1827) are two examples of endemisms that were on the brink of extinction which were able to recover thanks to an interdisciplinary approach combining genetics, field monitoring and habitat restoration (Fabbri et al, 2007, 2018; Kleinman-Ruiz et al, 2017, 2019). In some cases, locally extinct species can be reintroduced by mixing individuals from different sources to enhance their adaptive potential, as has been done successfully for the Asiatic wild ass, *Equus hemionus* Pallas 1775, in Israel (Zecherle et al, 2021) and the Eurasian otter, *Lutra lutra* Linnaeus 1758, in the Netherlands (Koelewijn et al, 2010). In other cases, such as in Italian pike and marble trout, allochthonous hybridisation erodes their endemic genetic diversity and threatens their long-term survival so conservation genetics and genomics strategies for these species focus on identifying and removing hybrids from the gene pool.

An overview of the most common and successful genetic techniques is presented hereafter.

1.4.1. Molecular markers

Allozymes

Molecular markers have revolutionised research in many scientific fields ranging from medical research to conservation ecology and biology. In the 1960s and 1970s, genetic diversity was assayed through protein variants, namely allozymes, which were separated by electrophoresis (Harris, 1966; Prakash, Lewontin and Hubby, 1969; Lewontin, 1974). Allozymes have been implemented in many species, including pike (Miller and Senanan, 2003) and trout (Giuffra et al, 1996). However, while this technique is time- and cost-efficient, its main drawbacks are the scarce number of polymorphic loci - as few as 2 out of 65 in *E. lucius*) (Seeb et al, 1987) - and the lethal sampling technique to obtain liver, muscle and eye tissue (Miller and Senanan, 2003).

Restriction Fragment Length Polymorphisms

The discovery of restriction enzymes, a type of endonuclease, paved the way for the first DNA-based markers: restriction fragment length polymorphisms, RFLP (Botstein et al, 1980). These enzymes catalyse the cleavage of DNA at specific nucleotide sequences, producing a variety of variously sized fragments with different electrophoretic mobility. If an allelic variant is present within the cleavage sequence, the enzyme does not recognize it and cleavage does not occur, leading to a longer fragment which can be detected through gel electrophoresis. Moreover, with the advent of the polymerase chain reaction (PCR) (Mullis, 1990), scarce amounts of DNA fragments were able to be amplified to measurable quantities. This allowed for minimally invasive sampling of fin tissue and the consequent release of fish. Splendiani et al. (2016) used mitochondrial (mtDNA) and nuclear (nDNA) RFLP loci to study Italian trout populations. A similar approach, amplified fragment-length polymorphism (AFLP) was used by Lucentini et al. (2011) to genetically characterise *Esox flaviae*. While RFLPs are widely available throughout the genome, previous knowledge about the species (i.e. specific PCR primers) is required to detect polymorphisms at a specific locus (Schlötterer, 2004).

Microsatellites

Also called Short Tandem Repeats (STR) or Simple Sequence Repeats (SSR), microsatellites are ubiquitous nuclear markers that consist of repetitive DNA motifs of two to six bases, with different alleles carrying a variable number of repeats (Richard et al, 2008). Their allelic variability coupled with the possibility to be amplified by simple PCR protocols contributed to the popularity of these markers, especially in population genetics. Indeed, the marble trout supportive breeding programme uses a combination of mtDNA and microsatellites to screen individuals (Meraner et al, 2013). On the other hand, genotyping microsatellites can be time-consuming as allele scoring is difficult to automate (Schlötterer, 2004).

Single Nucleotide Polymorphisms

Single Nucleotide Polymorphisms (SNPs) are the most abundant type of genetic marker (Nelson et al, 2004). Though usually biallelic, the strength of SNPs lies in their number: highly ubiquitous, a genome can contain several million SNPs which, as a whole, provide much higher resolution than other types of marker. Compared to microsatellites, genomic workflows for discovering and genotyping SNPs from genomic reads or genotyping arrays, respectively, are highly scalable and reproducible (Kerstens et al, 2009). Moreover, SNP arrays are becoming increasingly available for a great number of commercial breeds which can also be applied to wild populations (Kranis et al, 2013; Mattucci et al, 2019; Houston et al, 2014), making fast and large-scale genotyping possible at more accessible costs. SNPs are useful for mapping quantitative trait loci (QTL) in genome-wide association scans (GWAS) (Zargar et al, 2015) as well as population studies of demographic history and genomic divergence. Considering all of the above, SNPs are well-suited for integrating local population genetics surveys into large-scale studies. While Pustovrh et al (2012) have implemented a set of 47 nuclear SNPs in marble trout, to date, no population genomic studies using high-density SNPs have been published for either marble trout or Italian pike.

Next Generation Sequencing

The abovementioned discovery of high-density genomic marker sets would not be possible without technological advances in the field of high-throughput sequencing, which induced a revolutionary shift from traditional genetics to genomics. Next generation sequencing (NGS) produces massive amounts of information cyclically and in parallel (Park and Kim, 2016), as opposed to traditional Sanger capillary electrophoretic sequencing (Sanger et al, 1977). A main distinction can be made between whole genome sequencing (WGS) (Weber and Myers, 1997) and reduced-representation sequencing (RRS) (Altshuler et al, 2000), in which only a fraction of the genome (for instance, based on fragment size selection) is sequenced. Examples of the latter include Genotyping-by-Sequencing (GBS) (Elshire et al 2011), restriction site-associated DNA sequencing (RADseq) (Baird et al, 2008)

and double-digest RADseq (ddRADseq) (Peterson et al, 2012). Reads from both WGS and RRS can be assembled to form scaffolds or can be mapped to a reference genome if available, and genomic markers can be discovered, or “called”, by identifying sequence polymorphisms.

1.5. Genomic signatures of selection

The processes underlying molecular adaptation of populations to their environment leave traces along the genome, which can be detected using within- as well as between-population genomic differentiation data. Genome-Wide Selection Scans (GWSS) aim to identify chromosomal regions, and genes, under selective pressure through a variety of approaches.

In the absence of selection, random fluctuations in the frequency of a new mutation will be due to genetic drift alone, which is dependent on effective population size (Allendorf, 1986; Kimura, 1955; Lande, 1976; Nei & Tajima, 1981). Hence the frequency of a new neutral mutation in a large population may remain low for many generations and may be lost by chance. If it persists in the gene pool, it may slowly rise to higher frequencies in a manner that is directly proportional to its age (Kimura, 1955). The original chromosomal segment carrying the mutation, that is, the haplotype, will become progressively shorter over time as meiotic crossing-over events break it up at each generation (Hill & Weir, 1988; Weir, 1979).

However, when a new mutation arises, which bestows an increase in fitness coefficient on the individuals that carry it, it will be selected, meaning it will have a greater chance of being passed on to the next generations and it will rise to a high allelic frequency in a shorter amount of time compared to neutral ones (Smith & Haigh, 1974). It is this rapid expansion which characterises the selection signal. Moreover, any neutral alleles physically co-occurring on the haplotype will inevitably “hitch a ride” and follow the selected mutation into higher frequencies (Fig. 1.3). Such a process is termed genetic hitchhiking (Smith &

Haigh, 1974). A direct consequence of this is that haplotypes carrying a selected allele will remain distinctly long over time, compared to neutral haplotypes of the same age (Kim & Nielsen, 2004).

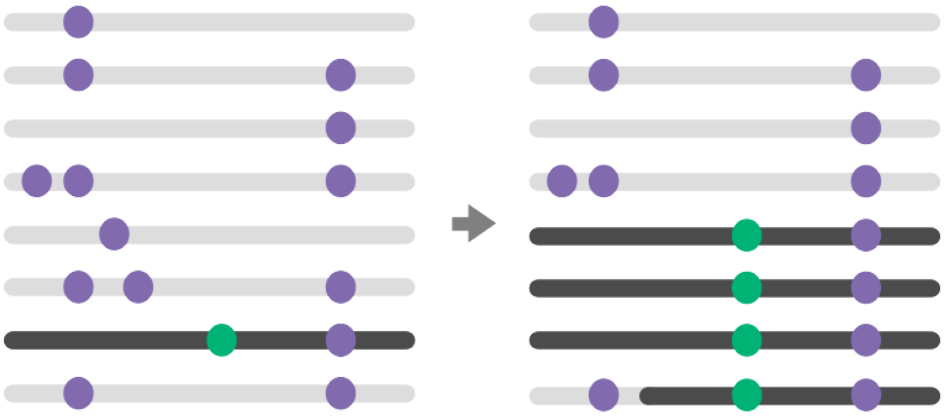


Fig. 1.3. An example of genetic hitchhiking due to the emergence of a beneficial mutation (green circle)**.** Purple circles correspond to pre-existing neutral mutations on chromosomes (grey bars). On the left, a set of randomly sampled chromosomes from a panmictic population are represented. The haplotype on which the beneficial mutation occurs (dark grey) is predominantly transmitted to offspring in subsequent generations on the right, and quickly spreads within the population. Thus, any neutral alleles close enough to the selected allele "hitchhike" and reach higher allelic frequencies.

Because genetic hitchhiking "sweeps away" genetic diversity, it results in a selective sweep, a term coined by Berry et al. (1991) and a mechanism extensively explored in literature (Braverman et al, 1995a; Charlesworth et al, 1993; Charlesworth & Charlesworth, 2018; Fu, 1997; Kim & Neilsen, 2004; Nielsen et al, 2005; Stephan, 2019). If the selective pressure is strong enough, it may ultimately lead to fixation of the alleles involved, thus a complete sweep. On the other hand, if the selective event does not persist

until fixation of beneficial allele(s), it is a partial or incomplete sweep (Ferrer-Admetlla et al., 2014; Nielsen et al., 2007).

Sweeps have also been distinguished into hard and soft sweeps (Hermisson & Pennings, 2005). The former, as previously described, corresponds to a single haplotype rising to high frequency thanks to the occurrence of a beneficial *de novo* mutation, while the latter can consist of more than one haplotype being selected due to either recurrent mutations or to pre-existing genetic variation that became advantageous following changes in the environment (Peter et al., 2012). Soft sweeps can also be partial or complete (Ferrer-Admetlla et al., 2014).

Overall, it has been postulated sweeps involve regions characterised by i) above-average haplotype length (Sabeti et al., 2002, 2007a; Tang et al., 2007) and ii) below-average genetic diversity compared to neutral genomic areas (Charlesworth et al., 1997). In cases of hard sweeps, below-average haplotypic diversity is to be expected as well (Sabeti et al., 2002). Unlike hard sweeps, soft sweeps do not always decrease haplotypic diversity (Sabeti et al., 2002), and may go unnoticed in certain selection scans if the implemented summary statistic is solely sensitive to, for example, allelic frequency as a discriminant factor.

Indeed, different statistics are designed to detect different types of sweeps (Williamson et al., 2007), or, in some cases, different stages of the same phenomenon, much like a wave as it grows upon nearing the shore and dissipates into the sand after it is gone. For example, if the sweep is complete and present only in one population, it may also produce a sufficiently strong signal in terms of allelic frequency differentiation when compared to other populations using allele frequency-based statistics such as F_{ST}, which, on the other hand, disregards haplotypes, i.e., the association between nearby variants. For this reason, F_{ST} is not sensitive to ongoing or partial sweeps, where allele frequencies across populations have not diverged as extremely. Other approaches are more adequate in the case of soft and partial sweeps, such as those implementing haplotype-based metrics (Stephan, 2019).

1.5.1. Genome-Wide Selection Scans

In Genome-Wide Selection Scans (GWSS), a common practice that is considered robust is to carry out analyses using different approaches and to then intersect outliers from each method (Hohenlohe et al., 2010; Vatsiou et al., 2016; Weigand & Leese, 2018). However, attention should be paid to which family of statistics are used and what kind of selection pattern these are sensitive to, as methods that detect mutually exclusive types of signals might yield a null or very limited intersection of candidate regions (Williamson et al., 2007; Zhong et al., 2022). Therefore, it is desirable to combine algorithms that look for compatible patterns and to separately analyse results from unrelated methods (Weigand & Leese, 2018). In this thesis, two complementary GWSS approaches – allelic frequency-based and haplotype-based – were implemented separately as a way to attain a broader understanding of how selection might be acting within and between species (Jónás et al., 2017; Zhong et al., 2022) (Fig. 1.4).

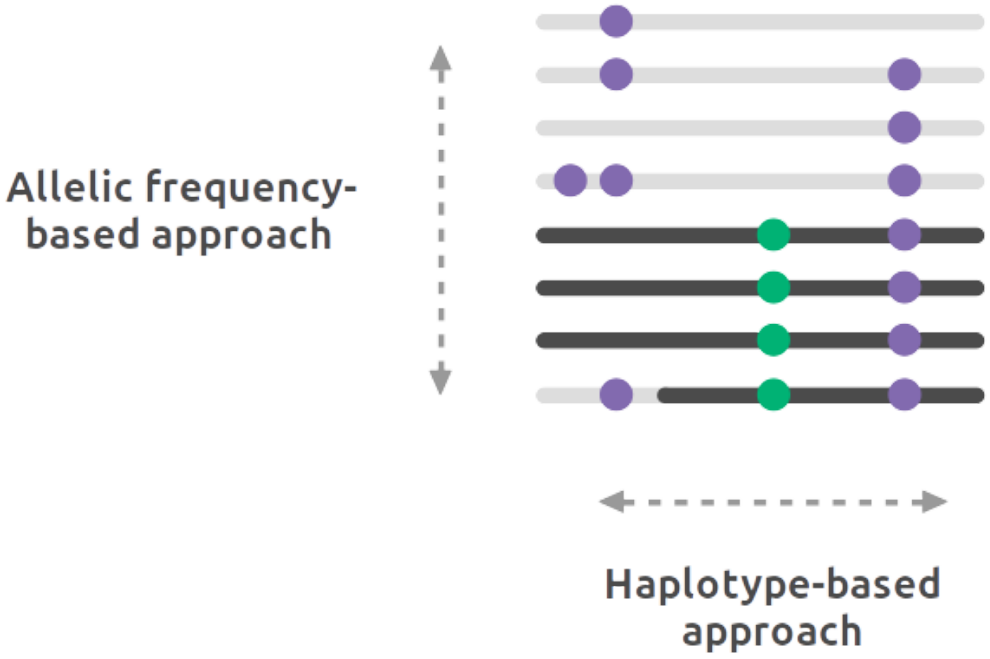


Fig. 1.4. Two possible approaches for detecting signals of selection in GWSS. For context, the concept of ”population” in population genetics is often visually represented as a collection of chromosomes or haplotypes stacked vertically, without distinguishing which belong to which individuals. Allelic frequency-based analytical methods can be referred to as “vertical” because they disregard physical association between alleles on a chromosome and only consider their frequency at a given locus across all sampled chromosomes. This means that even if chromosomes were hypothetically split into several shorter haplotypes and shuffled vertically, the vertical signal would remain unvaried. On the other hand, “horizontal” approaches such as haplotype-based methods give priority to linkage information between alleles and thus to haplotype lengths. Shuffling alleles vertically would disrupt this type of signal.

1.5.1.1. Allelic frequency-based approach

A popular approach in studies that explore signs of selection is to use statistics linked to allelic frequency. Often, these are either directly or indirectly correlated to heterozygosity, which can be assessed genome-wide either within or between populations to identify potential sweeps. When assessing a single population, areas with low heterozygosity may be indicative of selection (Charlesworth et al, 1997). On the other hand, the fixation index F_{ST} (Weir & Cockerham, 1984) is defined as the reduction in heterozygosity across two populations compared to expected values under Hardy-Weinberg equilibrium. It measures the amount of genetic divergence and can point to areas of the genome which are undergoing selective pressure in one population but not the other. F_{ST} ranges from 0, absence of allelic differentiation, to 1, distinct fixed alleles in the different populations.

In particular, F_{ST} is most sensitive to maximal differences in allelic frequencies such as those that take place when a beneficial mutation becomes fixed in one population but not in the other (Zhong et al, 2022). This can be the result of a completed hard sweep. On the other hand, because this statistic considers each marker independently, it disregards any physical linkage between markers. Therefore, while it can detect “vertical” signals (Fig.1.4), it

is blind to “horizontal” signals originating from direct association between linked markers. This means it might not be suitable for identifying partial, incomplete or soft sweeps where haplotype diversity is reduced, but overall allelic frequencies might not undergo detectable changes (Zhong et al., 2022). Nevertheless, this property of allelic frequency-based methods in general makes them useful when dealing with high-throughput sequencing of pooled or unphased samples (Rubin et al, 2010; Rubin et al, 2012; Bertolini et al, 2016), as they do not make assumptions about individual genotypes or haplotypes, respectively.

1.5.1.2. Haplotype-based approach

Complementary to F_{ST}, haplotype-based methods are capable of detecting “horizontal” signals along the genome (Fig. 1.4) and are often independent of allelic frequencies. These metrics place greater emphasis not on the relative abundance of a hypothetically beneficial mutation, but rather on the extent of the haplotype(s) on which it occurs, a feature that makes them sensitive to incipient or softer signals of selection.

The concept of Extended Haplotype Homozygosity (EHH) was introduced by Sabeti et al. (2002) and promptly became a widely used statistic for intra-population selection scans. EHH measures the decay of haplotype homozygosity, in other words, the breakdown of Linkage Disequilibrium as a function of distance from a site called focal or core SNP. EHH is calculated separately for each of the two alleles at the core SNP and is scaled from 0 to 1, being independent of core allele frequencies. Indeed, the main characteristic of EHH is its ability to describe the behaviour of haplotype conservation decay around an allele regardless of its frequency. The underlying hypothesis is that higher-than-expected EHH is associated with selected regions. In fact, according to the key postulates of natural selection, a selected allele will diffuse within a population in a relatively short amount of generations compared to neutral mutations. The scarce opportunities for meiotic recombination result in longer haplotypes around the selected locus. As a direct consequence of this, haplotype homozygosity decays slower, i.e., further from the considered locus.

Voight et al. (2006) further elaborated on this concept by proposing a test statistic of departure from normality for EHH, namely the integrated Haplotype Score (iHS), based on the standardised log-ratio of the integrals of EHH for either allele at a considered SNP. Another relevant statistic is the per-site EHH (EHHS), which linearly combines EHH at both alleles in order to yield a single value per SNP. Subsequently, Sabeti et al. (2007) and Tang et al. (2007) developed two ways of estimating EHHS that yield very similar results. Additionally, both publications proposed methods for comparing EHHS profiles between two populations, namely XP-EHH and RSB, depending on whether $EHHS_{Sabeti}$ or $EHHS_{Tang}$ are used, respectively. Essentially, XP-EHH and RSB are the cross-population adaptation of intra-population iHS: they are the normalised log-ratio of the integrals of EHHS for either population at a core SNP. Similarly to iHS, these tests are conducted per-site genome-wide and produce a signal which is capable of detecting the magnitude of the sweep. Moreover, the asymmetry of the signal indicates the population on which selection is acting.

1.6. Aim and structure of the thesis

The overall aim of this doctoral project is to assess the genomic landscape of Italian pike and marble trout populations in Italy and to provide relevant insight for conservation actions currently in place for these species using a whole genome sequencing approach. In particular, subsequent chapters of this thesis are structured as follows:

Chapter 2: Population Genomics in Italian Pike. In this chapter, I cover the bioinformatic pipeline used for the genomic read alignment and for the SNP discovery. I then investigate species differentiation and population subdivision within Italian pike using traditional population genetics approaches with high-resolution genomic markers. I also identify sets of ancestry-diagnostic SNPs as well as potentially introgressed alleles.

Chapter 3: Genome-Wide Selection Scans in Italian and European Pike. Here, I examine genomic footprints of selection underlying potential molecular adaptations in both pike species.

Chapter 4: Development of a High-Density SNP Panel for Marble Trout. This chapter covers the workflow of SNP discovery in pooled samples of Italian marble trout populations, fine-scale introgression analysis and the identification of paralogous loci as well as species- and population-diagnostic SNPs for the development of an informative SNP array.

Chapter 5: Final discussion. I contemplate the main findings of the thesis within the context of current conservation strategies, weighing the advantages and limitations of individual- and pooled-sequencing experimental designs. Last, I provide suggestions for further investigation.

Chapter 2: Population Genomics in Italian Pike

Abstract

The Italian pike was recently raised to species level and has been shown to have hybridised with stocked European pike (Esox lucius) leading to substantial genetic introgression. However, conventional genetic markers such as microsatellites generally lack the resolution required to disentangle the increasingly complex patterns of introgression across generations. Moreover, further research is needed to assess population substructuring and conservation status of this species, as it is currently considered “data deficient” by the International Union for Conservation of Nature (IUCN). To address these issues, whole genome sequencing data were analysed for both pike species, including Italian pike from six areas in Italy and two European pike populations. I assessed species divergence and population substructure using both principal component analysis and the model based algorithms implemented within the programme fastSTRUCTURE on 3.9 million SNPs. This provided further proof of species differentiation (the first principal component explained 67.1% of the observed variance) and revealed that endemic Italian pike could be subdivided into at least four clusters. I compared the power to detect hybrids across SNPs and microsatellites and found that detection of hybrids through microsatellites is biased towards lower European pike ancestry values (qEUR), especially in hatchery samples. Moreover, I investigated private alleles between and within pike species that can be used for diagnostic purposes. This study represents the first genomic study of Italian pike and provides insights that are relevant for conservation policy for this threatened species.

2.1. Introduction

Currently, a supportive breeding programme at the “Centro Ittiogenico del Trasimeno" in the italian province of Perugia is in place for the conservation of the genetic integrity of endemic Italian pike, *Esox flaviae* (Lucentini et al, 2009a). Early monitoring protocols for this species consisted of growth rate and mortality assessments (Lorenzoni et al. 2009), later followed by studies of overall effective population size (Ne) and genetic diversity through traditional genetic markers such as AFLP, mtDNA and microsatellites (Lucentini et al, 2009; Lucentini et al, 2011; Gandolfi et al 2017). Because the Italian pike is threatened by hybridisation with stocked, non-native European pike (Lucentini et al, 2011; Gandolfi et al, 2017), screening of spawners is aimed at identifying and removing hybrids from the gene pool. However, the screening process still greatly relies on a combination of mtDNA and phenotypic traits to differentiate between pure and hybrid individuals (Lucentini et al, 2009) and, after several generations, highly complex patterns of introgression are expected which are likely to elude detection with current monitoring methods. This highlights a need for the development of a novel technology for genetic monitoring capable of yielding high-resolution genotype data which are also standardised across research groups.

In this chapter, a large set of genomic SNPs was identified to assess genetic structuring at species and population level. Particular focus is placed on the detection of pike hybrids to assess introgression from *E. lucius.* An individual-based Whole Genome Sequencing approach was carried out to obtain abundant data for each of 61 pike individuals from eight localities including six italian ones, and to be able to compare results to previously generated microsatellite genotypes for the same individuals (Gandolfi et al, 2017). Lastly, sets of ancestry-diagnostic SNPs were identified for potential use in high-throughput genotyping technologies such as SNP arrays.

2.2. Methods

2.2.1. Genomic Data Preparation

2.2.1.1. Sample choice

A total of 61 pike individuals were chosen for this project from a larger cohort that had been genotyped with microsatellites for a previous study (Gandolfi et al, 2017). These individuals were sourced from eight areas (Fig. 2.1), six of which are located in Italy: Adda (n = 3) and Po (n = 11) river systems, Garda Lake (n = 10), Trentino region composed of Caldonazzo Lake (n = 10) and Terlago Lake (n =1), South Tyrol region with samples from the Adige River (n = 3) and Trasimeno Lake (n = 7). In addition, samples from one population in Austria (Danube River, n = 10), one in Germany (Elbe River, n = 3) and a hatchery located near Trasimeno Lake (n = 3) were also included.

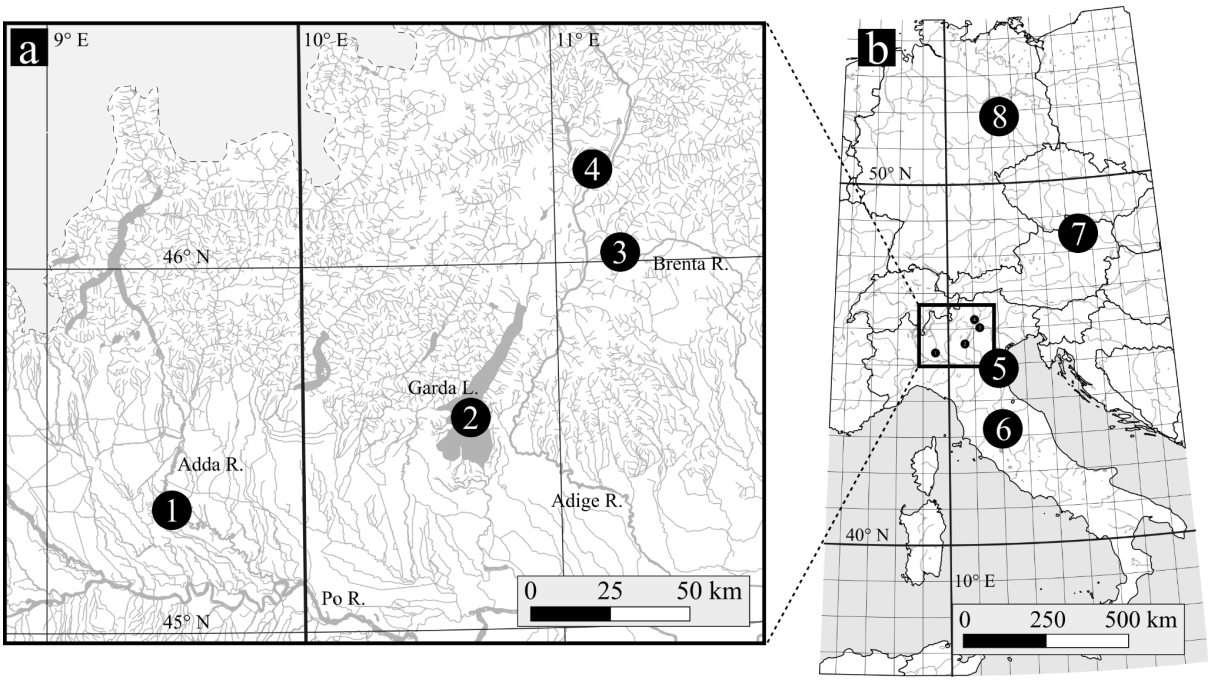


Fig. 2.1 Pike sampling locations. A: 1) Adda River; 2) Garda Lake; 3) Trentino (Caldonazzo and Terlago Lakes); 4) Adige River; **B:** 5) Valli di Argenta, Po Delta; 6)

Trasimeno Lake; 7) Danube River, Austria; 8) Elbe River, Germany. Figure adapted from Gandolfi et al, 2017.

2.2.1.2. Genomic read preparation

All 61 pike samples were sequenced at a target depth of 20X using Illumina 150 nt paired-end sequencing technology at Novogene Ltd, Cambridge, UK. This involved random fragmentation by sonication, with the DNA fragments being end-polished, A-tailed, and ligated with the full-length adapters of Illumina sequencing, and followed by further PCR amplification with P5 and indexed P7 oligonucleotides. The PCR products as the final construction of the libraries were purified with AMPure XP system. Then libraries were checked for size distribution by Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA), and quantified by real-time PCR (to meet the criteria of 3 nM).

I assessed the quality of generated paired-end reads with FastQC v0.11.5 (Andrews, 2010). To further increase quality, reads were trimmed with trimmomatic v0.39 (Bolger et al, 2014) using the parameters “PE -threads 4 -phred33 LEADING:28 TRAILING:28 MINLEN:100” to ensure that surviving paired-end (PE) reads were at least 100 base pairs long, with base quality encoding corresponding to the Phred+33 format and a quality of at least 28 at the beginning and at the end of the read.

2.2.1.3. Bioinformatic pipeline for read alignment and SNP discovery

Read alignment and variant calling was carried out as per the Genome Analysis Toolkit (GATK) Best Practices v4 by the Broad Institute (Poplin et al, 2017; Van der Auwera et al., 2013). First, the publicly available *E. lucius* genome (Ensembl genome build Eluc_v3, NCBI RefSeq identifier GCA_000721915.3) was decompressed and indexed both with the faidx command in samtools v1.14 (H. Li et al., 2009) and with BWA index (Heng Li & Durbin, 2010) using algorithm option “-a bwtsw”. Similarly, a dictionary of the reference genome was generated with the GATK tool “CreateSequenceDictionary”. Both forward and reverse reads

for each individual were mapped to the reference genome using a Burrows–Wheeler transform as implemented in the BWA mem aligner (Heng Li & Durbin, 2010), resulting in a Sequence Alignment Map file (SAM) for each sample. Identical reads originating from PCR duplication were then marked and removed with GATK MarkDuplicates. Read mate information was subsequently updated with GATK FixMateInformation, and BAM files were once again sorted and indexed with samtools index.

Variants were called separately for each individual with HaplotypeCaller (Poplin et al, 2017), which scans the BAM files, locally realigns reads to account for insertions and deletions (indels) and identifies allelic variants under the form of both indels and SNPs. Variants were saved in a gVCF (Genomic Variant Call Format) file that contains information about polymorphic sites as well as non-polymorphic genomic regions. All gVCF files were then merged into one cohort gVCF with GATK CombineGVCFs, which was indexed with GATK IndexFeatureFile. Lastly, GATK GenotypeGVCFs was used to obtain the final VCF (Variant Call Format) file, which only contains genotype information about polymorphic sites and is the main input file for many downstream genomic tools.

2.2.1.4. SNP filtering

A general filter was first applied to the SNP set to improve its informativity and to reduce the probability of including pseudo-markers arising from sequencing errors. Software VCFtools v0.1.15 (Danecek et al, 2011) was used to exclude markers which i) were located on unmapped scaffolds, ii) were not SNPs, iii) had more than two alleles, iv) had more than 10% missing genotypes across individuals, v) had a minor allele frequency (MAF) lower than 5%, vi) had sequencing depth (DP) lower than 8 and vii) had a genotype quality score (GQ) lower than 40.

Some of the analyses described in this chapter, namely inference of population structure and admixture with fastSTRUCTURE and Treemix software, require markers to be

statistically independent. Therefore, both SNP sets (between species and within-Italian pike) were pruned by removing the SNPs which are in approximate linkage disequilibrium (LD); this was done to ensure that remaining markers are almost independent. Besides software constraints, having a smaller, independent set of genomic markers reduces information redundancy and allows for faster computation times. Minimally-linked SNPs were selected by setting a maximum correlation threshold of 0.2 r^2 in windows of 250 variants (Plink v2 option "--indep-pairwise 250 1 0.2"). Within these LD-pruned SNP sets, further thinning by SNP count was also implemented ("--thin-count") followed by the amount of SNPs to randomly retain.

For the remaining analyses, such as the categorisation of population diagnostic alleles or genome-wide selection scans in the following chapter, the entire unpruned SNP set was used, as these methods are not impacted by the physical proximity of markers, rather, they benefit from an increased marker density.

2.2.2. Population Subdivision

2.2.2.1. Principal Component Analysis

Principal Component Analysis is a multivariate analysis which transforms the data by fitting orthogonal vectors, termed Principal Components. In genetics and genomics, input data usually corresponds to a matrix of individuals and their biallelic genotypes coded according to whether each marker is heterozygous or homozygous for the reference or alternative allele. The utility of PCA lies in the reduction of dimensionality, that is, its ability to synthesise large amounts of variables - i.e. genotypes - into few Principal Components, that explain most of the variability among individuals and can be easily represented as axes on a cartesian plane. PCA was carried out with Plink v1.9 to detect population subdivision both across the two pike species and within the Italian pike samples using the quality-filtered SNP data set in its entirety. For within-Italian pike PCA, only samples having a proportion of

Italian pike ancestry (qITA), as given by fastStructure, above 90% were considered, see the next section for further details.

2.2.2.2. fastSTRUCTURE

In addition to PCA, a model-based approach has been used to determine the fine-scale patterns of genetic variation. To this end, fastSTRUCTURE v1.0 (Raj, Stephens and Pritchard, 2014) was used to investigate admixture between species and within-Italian pike . Briefly, fastSTRUCTURE operates by estimating $\mathbf{q_k(i)}$, the proportion of individual *i*'s genome that originated from population *k* (Pritchard, Stephens and Donnelly, 2000). Individuals are not assigned to a population *a priori*, rather, the assignment is inferred by a Bayesian algorithm based on a given value of K (total number of clusters) and allelic frequencies. This software places one of two types of prior, simple or logistic, on allelic frequency. While the former is faster to compute, the logistic prior was preferred in this study because it is more flexible and thus more adequate for complex structure (Raj, Stephens and Pritchard, 2014).

Model complexity, that is the most likely number of populations, was assessed by running fastSTRUCTURE separately for an increasing value of K both between species and then within Italian pike alone. fastSTRUCTURE performs best for independent loci so sets of unlinked SNPs were used. Between-species admixture was investigated for each K from 1 to 10 through 20 replicate runs with 50 thousand SNPs. Preliminary tests showed that this marker set size was sufficient for robust variational inference across the two pike species given their genetic divergence. Resulting estimates of population ancestry qITA and qEUR were assessed for each individual, and hybrid samples with qEUR > 5% were removed for within-Italian pike analyses. Fine-scale population subdivision within Italian pike was then assessed using 100 thousand SNPs for each K from 1 to 9 through 16 replicate runs for K < 6 and 11 replicate runs for the rest.

Lastly, the proportion of European pike ancestry (qEUR) in hybrid individuals inferred through fastSTRUCTURE using 50 thousand SNPs was compared to previous admixture estimates using STRUCTURE (Pritchard, Stephens and Donnelly, 2000) with 17 microsatellites and published in Gandolfi et al, 2017.

2.2.2.3. Treemix

Hierarchical population structure and admixture was further analysed with Treemix v1.13 (Pickrell and Pritchard, 2012), which infers a bifurcating tree between the populations based on the maximum likelihood criterion using allelic frequencies. It improves the estimation of tree topology by accounting for gene flow between populations, modelled as migration events or edges (m) which can be specified *a priori* to the software.

The optimal number of edges was assessed using 100 thousand independent SNPs for each m from 0 to 5 through 10 replicates runs (parameters “-global -k 1 -bootstrap”), excluding any hybrids with qEU > 0.05 and setting the Danube *E. lucius* population as outgroup. Then, R package optM (Fitak, 2021) was used to infer the most likely number of migration events in the data using “Linear” and “Evanno” methods (see optM publication for details).

2.2.3. Population Diagnostic Alleles

Private alleles are variants which are found exclusively in one population (or a subset of populations) but not in others. They are useful for identifying from which population (or populations) an individual originates. Because of this property, private alleles can be included in genotyping arrays for diagnostic purposes, allowing for efficient discrimination of ancestry and admixture.

After estimating the individual proportion of Italian (qITA) and European pike (qEUR) ancestry in section 2.3.2.2, hybrids displaying a qEUR greater than 5% were removed for

this analysis. The VCF containing all remaining individuals was split into different files based on sampling location, and SNPs that became monomorphic after subsetting were removed from each set. Polymorphic SNPs from each subpopulation were then compared to identify private alleles occurring only in one, two and three subpopulations.

2.3. Results

2.3.1. Genomic Data Preparation

In total, 1.52 Terabytes of genomic data were generated, consisting of 7.9 billion total reads and about 87.6 million reads per individual. Mean quality across the entire length of trimmed reads was high, with an average Phred score of 37.3 (Fig. 2.2). On average, 99.4% of reads were mapped to the *E. lucius* reference genome, with no significant difference between *E. lucius* and *E. flaviae* samples. A total of approximately 4.8 million SNPs were identified through the variant calling pipeline after aligning reads to the *E. lucius* genome. An initial filter narrowed this set down to ~3.9 million high-quality, biallelic SNPs with a minimum allele frequency of 0.05. Of these, 472 thousand were located on unmapped scaffolds. After filtering out markers which were not on mapped chromosomes, the final data set consisted of ca. 3.5 million SNPs for 61 pike individuals.

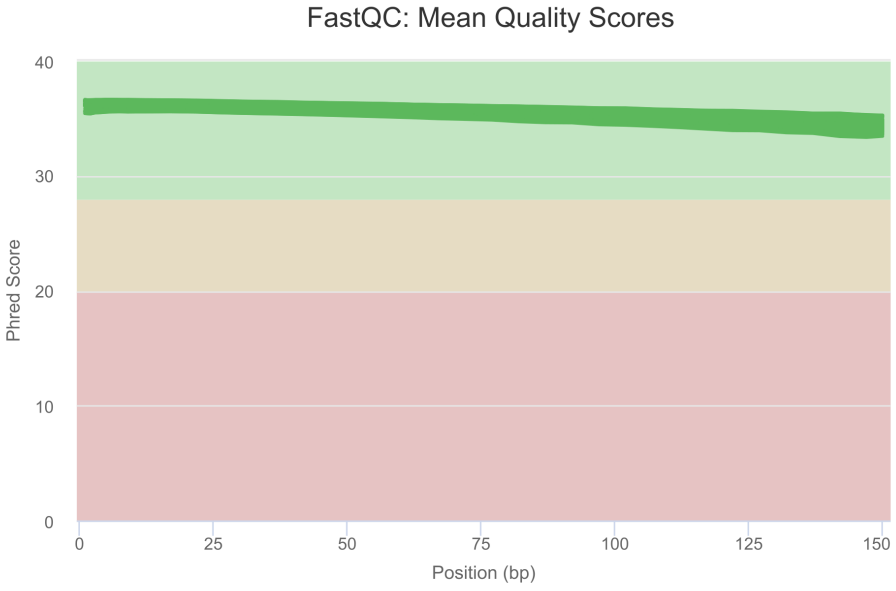


Fig. 2.2. Mean quality of genomic reads measured in Phred score. For context, a Phred score between 30 and 40 corresponds to a probability of correctly calling a base between 99.9% and 99.99%.

2.3.2. Population Subdivision

2.3.2.1. Principal Component Analysis

PCA including Italian pike and European pike using 3.5 million informative SNPs (Fig. 2.3.A) revealed conspicuous genetic divergence along the first principal component (PC1) which accounts for 67.1% of total observed variance and is in line with previous genetic studies showing divergence across the two species (Lucentini et al, 2011; Gandolfi et al, 2017). Hybrids are visible along PC1 between the European and Italian clusters. The second, third and fourth principal components correspond to within-species differentiation and collectively explain one-tenth of total variance in the sample. In European pike, there is greater within-population diversity in the Danube compared to the Elbe population.

The within-Italian pike PCA (Fig. 2.3.B) showed structuring along PC1 and PC2, accounting for 22.9% and 11.3% of total variance respectively. PC1 separates the Trentino group from the remaining populations, while Garda and Trasimeno form two clusters positioned at the extremes of PC2. Po and Adda individuals also present clustering along PC2, but to a lesser extent. PC3 predominantly reflects genetic variation within Trentino but also within Po populations.

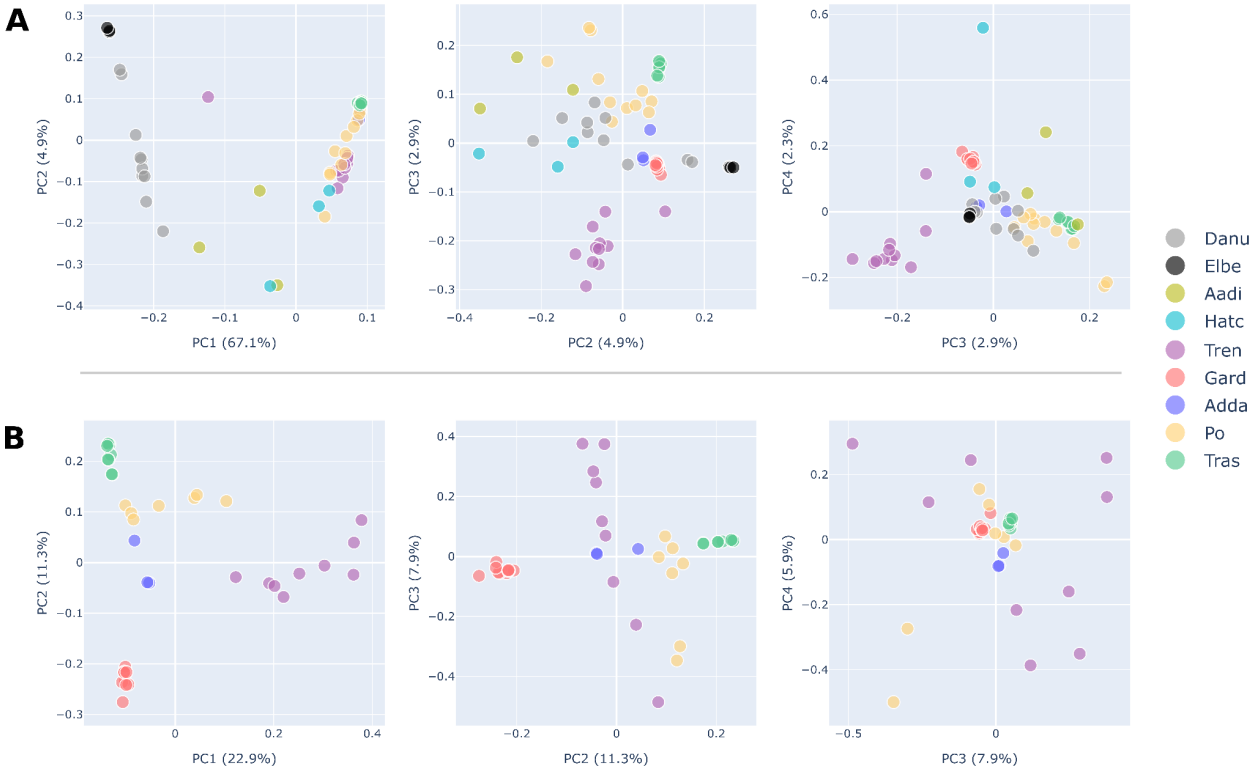


Fig. 2.3. A) PCA of all pike samples using 3.9 million informative SNPs reveals genetic divergence along PC1 between Italian pike populations (Trentino region, **Tren**; Garda lake, **Gard**; Adda river system, **Adda**; Po river system, **Po**; Trasimeno lake, **Tras**; Alto Adige, **Aadi**; Hatchery, **Hatch**) and European pike populations (Danube river, **Danu**; Elbe river, **Elbe**). Five potential hybrids are clearly visible along PC1 between the European and the Italian clusters: one from Trentino, three from Alto Adige river and one from a hatchery. B) Within-Italian pike PCA using 1.4 million polymorphic SNPs sheds further light into the hierarchical substructuring of this species.

2.3.2.2. FastSTRUCTURE

Variational inference of admixture between Italian and European pike confirms subdivision into two clusters corresponding to the two species (Fig. 2.4 upper plot). Samples with the highest levels of introgression from European pike were Alto Adige (qEUR = 0.60) and Hatchery (0.31), followed by lower but ubiquitous introgression in Trentino (0.11) and Po (0.07). After applying a 5% qEUR filter, all Alto Adige and Hatchery individuals were removed as well as approximately half of all Trentino (5 out of 11) and Po (6 out of 11) individuals.

Population subdivision within Italian pike varies depending on the level of model complexity analysed (Fig. 2.4 lower plots). When K is set to 2, Trentino and Garda samples cluster together while Trasimeno and Po form another cluster and Adda shows mixed ancestry between the two. At K = 3, Trentino separates from Garda, and Adda shows admixture mainly between Trasimeno+Po and Garda. For K greater or equal to 4, further light is shed on Adda and Po ancestries: although the variational algorithm converges on two different scenarios per K (shown as major and minor modes with their relatives frequencies), there is more support for Adda as a fourth separate cluster, and Po as an admixed population with contribution from Trasimeno and Adda. Indeed, the ChooseK.py script provided by the authors of fastStructure infers that four clusters are sufficient to optimally explain structure within the data.

When comparing variational inference of admixture using SNPs and microsatellites (Fig. 2.5), a trend emerges showing a systematic underestimation of qEUR when using 17 microsatellites as opposed to 50 thousand SNPs, and this is particularly evident in the hatchery samples.

Fig. 2.4. Variational inference of admixture between Italian and European pike (upper plot) and within Italian pike subpopulations (lower plots) using fastSTRUCTURE with 50K and 100 thousand SNPs, respectively. Each bar represents one individual and vertical partitions correspond to the proportion of membership (**q**) to each of K inferred clusters. **Between-species** analyses confirm previous findings of subdivision into two clusters corresponding to the two species and hybrids. **Within Italian pike**, population subdivision depends on the level of model complexity analysed. Results for K greater than 5 are not shown, as they vastly coincide with the major mode of K = 4.

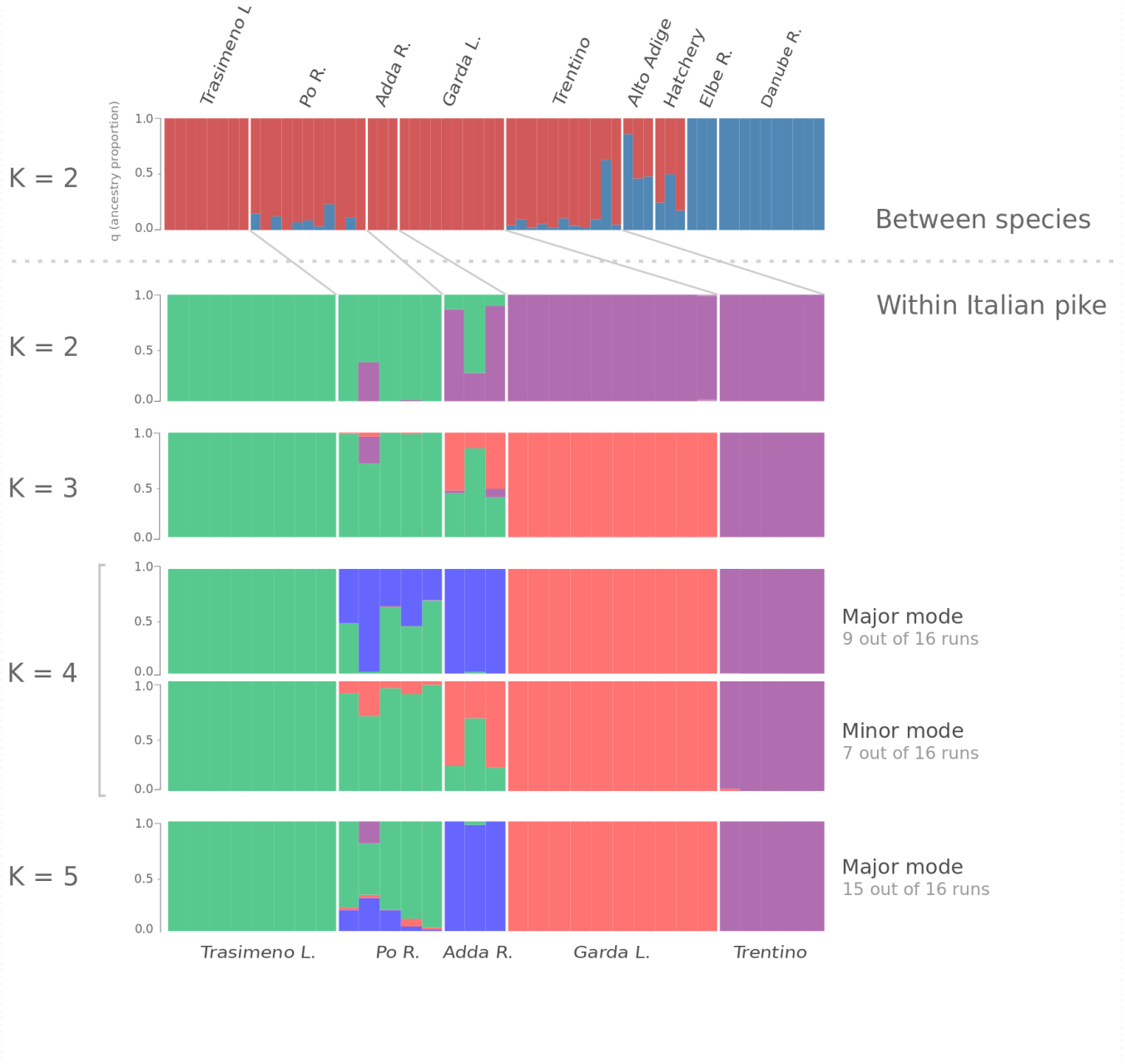
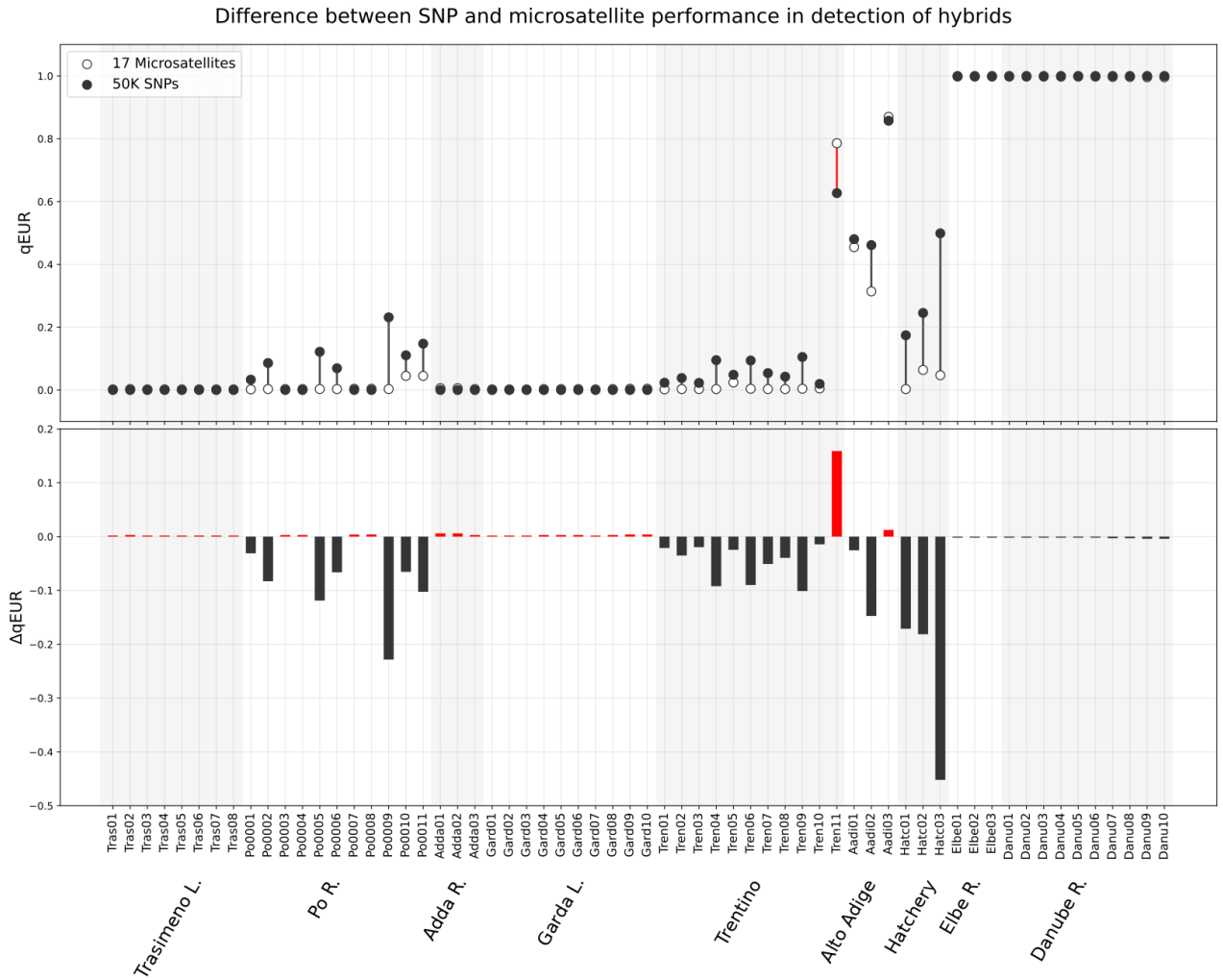


Fig. 2.5. Comparison between admixture analyses with SNPs versus microsatellites. Ancestry coefficients of 61 pike samples, including Italian, European and hybrid individuals, were obtained using fastSTRUCTURE and STRUCTURE respectively for 50 thousand SNPs and 17 microsatellites (from Gandolfi et al., 2017), setting K = 2. Estimates of European pike ancestry (qEUR) are shown in the upper plot for microsatellites (empty circles) and SNPs (full circles), and their delta ($qEUR_{microsat}$ - $qEUR_{SNP)}$ is shown as bars in the plot below, in black where SNPs were able to detect higher qEUR than microsatellites and in red for the opposite case. Overall, while both SNPs and microsatellites correctly identified pure individuals of either species, microsatellites underestimated qEUR in almost all hybrid individuals.



2.3.2.3. Treemix

Assessment of the optimal number of migration edges (m) through optM “Linear” and “Evanno” methods revealed the most likely scenario is one admixture event (Fig. 2.6). For this solution, almost all Treemix replicates showed admixture from Trasimeno into Po samples (see major mode in Fig. 2.7) while one presented gene flow from the Trentino-Adda cluster into Po (minor mode).

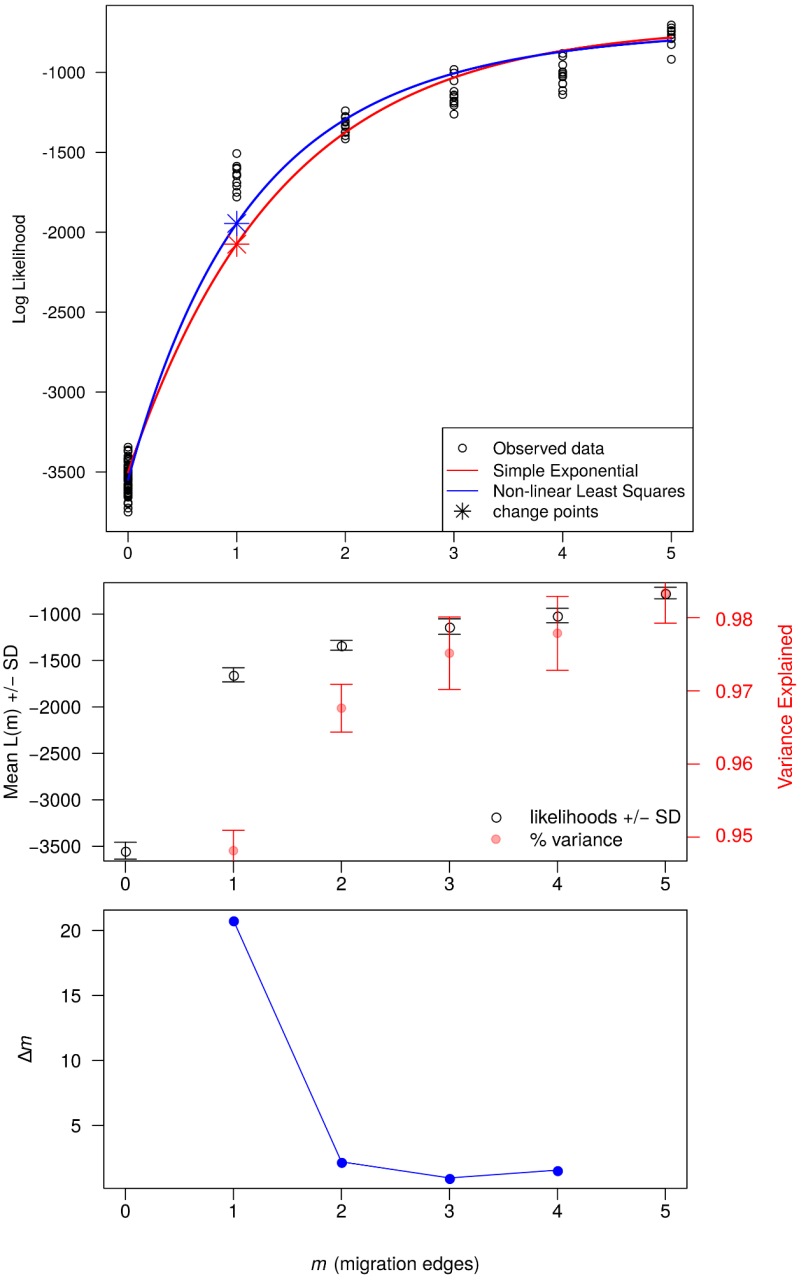


Fig. 2.6. The optimal number of migration edges, i.e. admixture events, between and within Italian and European pike populations was assessed with R package optM methods "Linear" (upper plot) and "Evanno" (lower plots). L(m), or Log Likelihood, is the log posterior probability of the data given *m*, and Δ*m* is the second-order rate of change in likelihood across incremental values of *m* (Fitak, 2021). Both methods converge on m = 1 being the optimal number of migration edges, which explains about 97% of total variance.

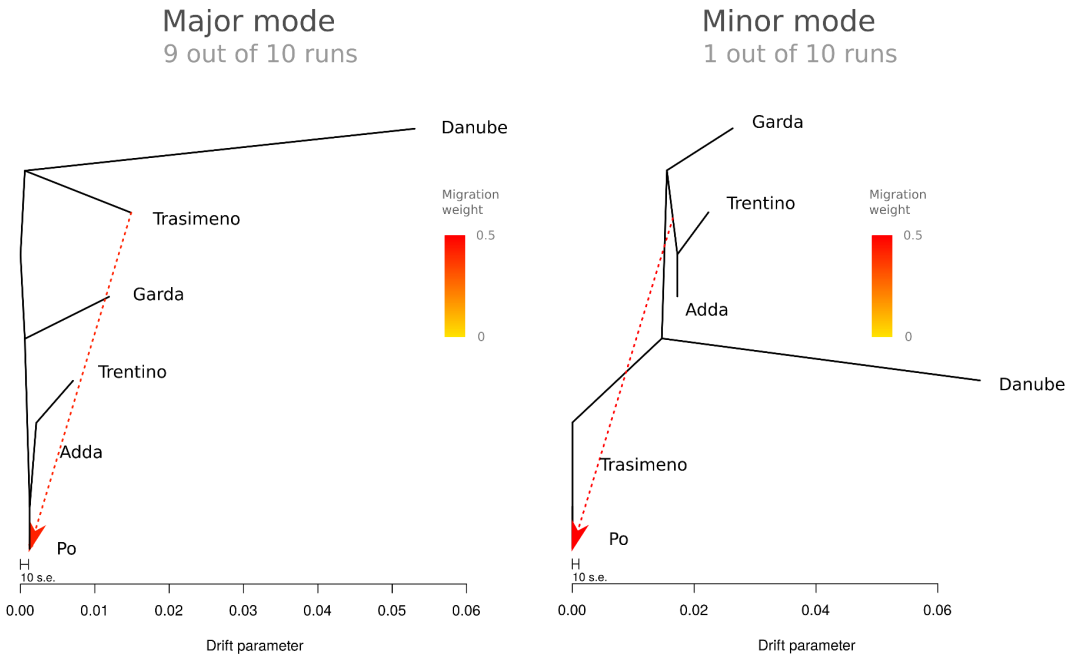


Fig. 2.7. Tree topology with Danube as outgroup, inferred with Treemix accounting for one migration edge (red arrow) representing gene flow between Trasimeno and Po (major mode) and between Trentino-Adda and Po populations. On the X axis, the drift parameter represents the amount of estimated genetic covariance due to random genetic drift in a Wright-Fisher model.

2.3.3. Population Diagnostic Alleles

Out of 3.5 million total SNPs, 213,734 (6.4%) and 415,679 (12.6%) segregated only in Italian or European pike, respectively (Fig. 2.8. B). The higher volume of European private alleles might be linked to a reference bias due to having mapped reads to the *Esox lucius* reference genome, which could lead to an under-representation of alternative alleles (Brandt et al, 2015). The Danube river *E. lucius* subpopulation showed the most private alleles (807,539) while the Elbe river *E. lucius* subpopulation presented just 1,734 private alleles. This notable difference is likely to be inflated because of sampling bias and a difference in sample size (10 Danube samples versus 3 Elbe) (Trask et al, 2011). Increasing the number of Elbe individuals would allow the discovery of more variants segregating in this population, some of which are currently labelled as private in the Danube sample due to lack of information. Nonetheless, these alleles are diagnostic of the European species and are therefore useful for identifying hybrids originating from admixture between the two species.

The analysed data set includes “pure” Italian pike having a qEUR lower than 5%, which is a rather conservative threshold for the purpose of identifying ancestry-diagnostic alleles as adequately as possible. Still, some alleles which are introgressed from *E. lucius* into *E. flaviae* populations evade this filter. A way to identify such alleles is to assess which of these are present exclusively in *E. lucius* populations plus only one of the *E. flaviae* subpopulations, as a consequence of *E. lucius* individuals being translocated into Italian locations. This category of alleles is shown in Fig. 2.8.C, with Trentino and Po samples having the greatest number of potentially introgressed alleles and cumulatively reaching about 15% of all polymorphic alleles. It is worth noting that an even more strict qEUR tolerance would likely decrease the amount of introgressed alleles identified with this approach. However, it would do so at the cost of further reducing the sample size, which could in turn enhance the effects of sampling and ascertainment bias. Rather, by identifying and removing potentially introgressed alleles, this method could prove useful for designing

an ancestry-informative set of SNPs based on a greater number of sampled individuals, even when remnant European pike ancestry is present in some of them.

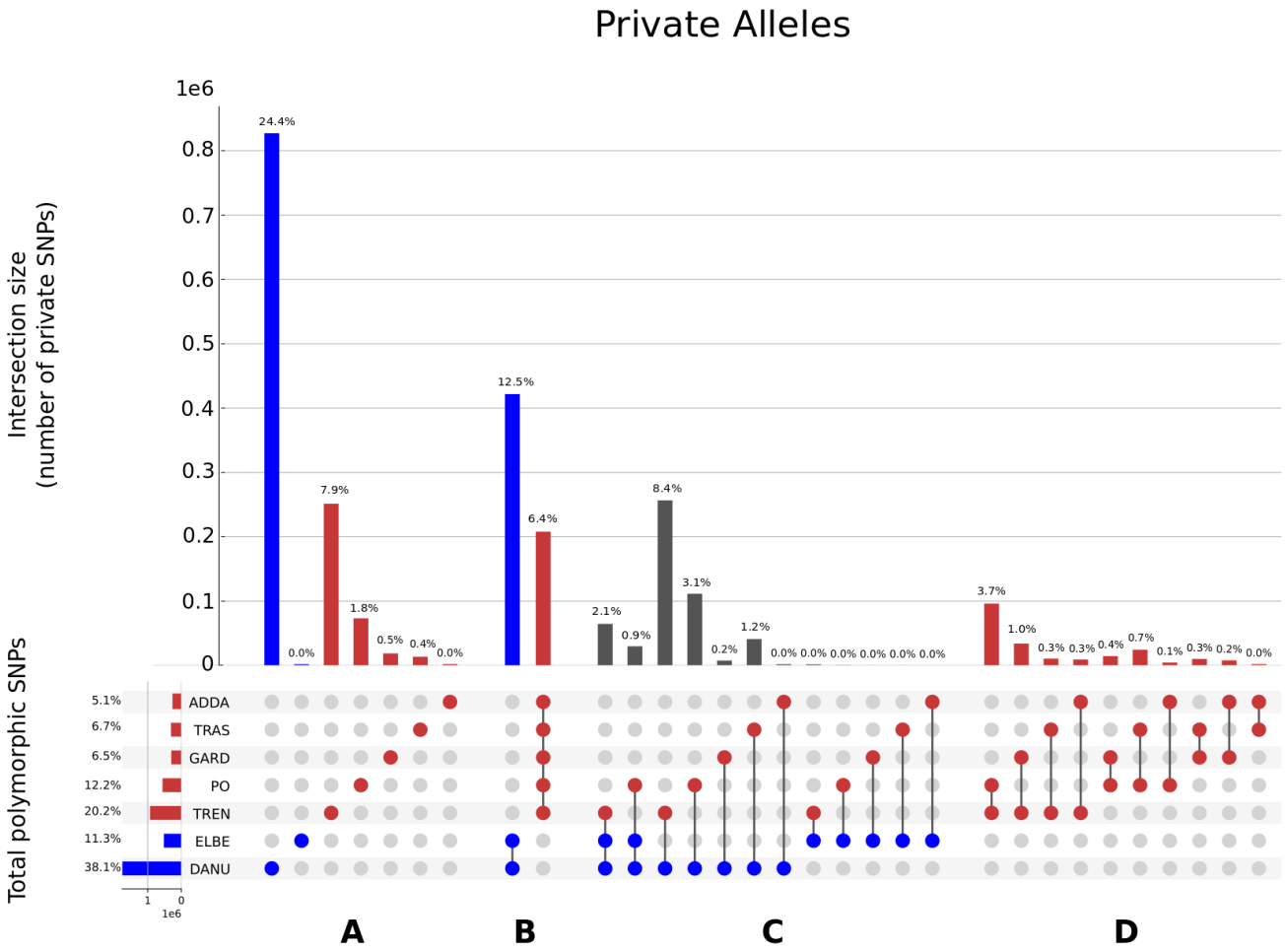


Fig. 2.8. Number of private alleles which segregate exclusively in the populations or set of populations shown below the bar plot as coloured circles. European and Italian pike populations are shown in blue and red, respectively. Percentages refer to the 3.5 million polymorphic SNPs considered after removing hybrids with qEUR > 5%. A) Subpopulation-diagnostic private alleles. The large difference between Danube and Elbe private alleles is likely due to sampling bias. These alleles are also diagnostic of ancestry at a species level. B) Species-diagnostic private alleles. C) These sets are likely to contain a proportion of introgressed alleles from European pike into Italian pike populations. D) Alleles which segregate exclusively in pairs of Italian subpopulations.

2.4. Discussion

In this chapter, between- and within-population subdivision in Italian and European pike was investigated with a resolution that was until now unavailable for the italian species. Principal Component Analysis and Bayesian inference of structure and admixture confirmed the genomic divergence of the two species that is consistent with the species status of *E. flaviae* suggested by previous studies (Lucentini et al, 2006, 2011; Gandolfi et al, 2017). When using PCA on two closely related species, the first component generally represents the highest taxonomic subdivision, with potential hybrids occur in the middle between the two species clusters, as is the case in wolf-dog (Stronen et al, 2022) and wildcat-domestic cat admixture studies (Mattucci et al, 2019). In this study, this is highlighted by a nearly 7-fold difference in genetic variance between species-level (PC1) and subpopulation-level (PC2, PC3 and PC4 combined) variation (Fig. 2.3.A).

A hierarchical subpopulation structure clearly emerges within Italian pike (Fig. 2.4), where four clusters were identified in agreement with previous findings (Gandolfi et al., 2017). In addition, this study sheds further light on gene flow across Italian pike populations. Indeed, tree topology inferred through Treemix (Fig. 2.7 major mode) accounts for gene flow from Trasimeno to Po, in line with anthropogenic translocations from Trasimeno Lake, which hosts the pike supportive breeding programme. In this model, the earliest population split corresponds to the divergence between Trasimeno and the remaining Italian populations, possibly because of either colonisation or vicariance due to the Apennines posing a barrier to gene flow. The Trasimeno-Po gene flow is compatible with within-Italian pike fastStructure results for k = 2 (Fig. 2.4) where Trasimeno and Po samples cluster together, although the Treemix minor mode topology seems to better support this clustering pattern. Gandolfi et al. (2017) identified the Adda River population as a genetic cluster on its own, with some individuals showing admixture from the Garda Lake and Trasimeno clusters (Gandolfi et al., 2017, Fig. 5e2). Support for Adda River pike as a separate cluster can also be seen in my

results (Fig. 2.4). These should thus be prudently considered as a subpopulation in future restocking activities to preserve within-species genetic architecture.

This first genomic study in *Esox flaviae* allowed for comparison between SNPs and traditional population genetics markers, namely microsatellites, currently used for monitoring *Esox flaviae* populations. One of the key results is the discrepancy between these two widely used techniques. In particular, estimates of European pike ancestry were systematically biased towards lower values in hybrid individuals when using microsatellites as opposed to SNPs. In other words, genomic markers were able to detect admixture better than microsatellites. This is undoubtedly due to the ubiquitous presence of SNPs along the genome, which prove useful in detecting complex patterns of introgression as haplotypic blocks inherited from European pike break down across generations. In contrast microsatellites only represent a small number of loci from different parts of the genome. A similar pattern has been observed by other studies of introgression using both SNPs and microsatellites, including species such as the spotted eagle (Väli et al, 2010), red deer (McFarlane et al, 2019) and steppe polecat (Szatmári et al, 2021).

Besides covering the genome in its entirety, the SNPs hereby discovered represent a feasible alternative to traditional methods because diagnostic analyses of individuals are easily standardised and upscaled using SNP arrays. The sets of population-diagnostic alleles identified in this chapter are indicative of ancestry both at species and subpopulation level, and can be further refined to be employed in a genotyping assay for quick discrimination of hybrids with low genotypic error rates (Anderson and Garza, 2006). While it can be challenging to determine precise reduction in expenses as they depend on various factors, such as sequencing and genotyping services, level of competence of technicians and analysts, etc, it can be estimated that implementing such a procedure would reduce the time and cost of data analysis by approximately an order of magnitude, increasing data transferability across research groups (Anderson and Garza, 2006).

2.5. Conclusion

- ~3.9 million high quality biallelic SNPs were discovered across both pike species, ~3.5 million of which are located on mapped chromosomes.

- Population subdivision analyses show a clear distinction between two divergent clusters that coincided with Italian and European pike, confirming the species-level status of *E. flaviae*.

- Within-Italian pike genetic structure is hierarchical, with the highest level representing a separation into two clusters, Trentino-Garda and Trasimeno-Po, with Adda showing admixture between the two.

- In genetic clustering analyses that assume 3 groups, Trentino and Garda separate into two different clusters while Trasimeno-Po remain associated.

- In genetic clustering analyses that assume 4 or more groups, the fastSTRUCTURE algorithm converges on different clustering solutions. Most of these identify Trentino, Garda, Trasimeno and Adda as separate clusters and Po as an admixed population between the Adda river system and Trasimeno Lake.

- Admixture between Italian subpopulations is likely driven by man-mediated translocations, especially where Trasimeno is concerned.

- Detection of hybrids through microsatellites is biased towards low qEUR values, especially in hatchery samples, possibly due to a domestication effect combined with the lower resolving power of microsatellites.

- Sets of ancestry-diagnostic SNPs were identified to be used in conservation actions, including alleles potentially introgressed from *E. lucius*.

Chapter 3: Genome-Wide Selection Scans in Italian and European pike

Abstract

One of the milestones of conservation genomics is to preserve the adaptive potential of a species or population, to allow it to better face environmental change. It provides the opportunity to identify regions of the genome that appear to have diverged under selective sweeps as described in section 1.5. In this chapter, two complementary approaches, F_{ST} and XP-EHH, were implemented to detect selection signals in *Esox flaviae* as well as in *Esox lucius* populations. In order to implement XP-EHH, genotypes were phased by computationally inferring haplotypes. Genes falling within candidate genomic regions were then assessed for functional enrichment to detect molecular changes that may have arisen under selection in either species. Gene Ontology Enrichment Analysis revealed a significant overrepresentation of gene functions involved in metabolism, immune system, genetic regulation, tissue repair, reproduction and sensory perception. In particular, a high number of genes were related to olfactory perception, which is an important aspect of the predatory tactics of Esocids, being ambush predators that rely greatly on olfaction. Interestingly, Italian and European pike showed signs of selection in different clusters of olfactory receptor genes, suggesting that this might be indeed due to molecular adaptations to their respective ecological niches. As a whole, these findings inform the need to preserve the genetic integrity of sister taxa, even when closely related, as the cost of their hybridisation could be the loss of specific genomic adaptations to the environment.

3.1. Introduction

3.1.1. Preserving the adaptive potential of threatened populations

Supportive breeding programmes mainly seek to maintain neutral genetic variation within threatened species (Frankham et al, 2002). However, there is a rising concern that naturally selected variants underlying adaptive traits could be neglected in captivity, instead favouring maladaptive traits by artificial selection, leading to a decrease in adaptive potential (Fraser 2008, 2017; Willoughby et al., 2017; Willoughby and Christie, 2018; Hoelzel et al, 2019). This is because the rather small sets of neutral genetic markers commonly used in captive and supportive breeding programmes do not detect functional loci (Reed and Frankham, 2001), resulting in two main consequences. First, even when phenotypic traits are not intentionally bred for, such as in hatcheries with conservation purposes (Vainikka et al, 2021), they can still undergo positive artificial selection because they are undetected and therefore uncontrolled. Second, traditional genetic markers are often not suitable for assessing and maintaining the adaptive potential of a population (Holderegger et al, 2006), so fitness may be decreased even in captive-bred lines destined for reintroduction in the wild (Fraser 2008, 2017; Willoughby et al, 2017; Willoughby and Christie, 2018; Hoelzel et al, 2019).

The need arises for informative genome-wide markers capable of assessing adaptive potential in terms of functional genetic adaptations (Harrisson et al, 2014). This is a novel concept that begins to find its way into state-of-the-art conservation practices (Funk et al, 2018; Hoelzel et al, 2019). In this study, adaptive potential is studied for the first time in *E. flaviae* through statistical detection of signatures of natural selection acting along the genome (see section 1.5 for details). If an annotated reference genome is available, candidate genes can be identified by Genome-Wide Selection Scans (GWSS) and overrepresented molecular functions can be detected through Gene Ontology Enrichment

Analysis (GOEA), as described in 3.2.3. As a whole, these functions provide insight as to what type of adaptations might be positively selected in this species and inform conservation actions for Italian pike.

3.1.2. Computational haplotype inference

One of the two approaches for GWSS implemented in this thesis is haplotype-based, which requires haplotypes to be resolved and thus genotypes to be phased. In other words, information is needed about which alleles were inherited together on the same molecule of DNA. Phase can be inferred through databases of known haplotypes in well-studied species, or through physical information from long NGS reads. SNP arrays and short genomic reads, however, yield unphased genotype calls, in which case computational phasing can be carried out (Browning & Browning, 2011).

Phasing algorithms infer haplotypes through shared genomic tracts that are identical-by-descent (IBD) between related individuals. IBD blocks can also be estimated by the algorithm through pairwise comparisons of unrelated individuals, such as those in the current data set, benefiting from cryptic relatedness present in the sample which originates from a common ancestor (Voight & Pritchard, 2005).

3.1.3. Functional Enrichment Analysis

Given the availability of a chromosome-scale assembled and annotated reference genome, it is possible to identify and functionally characterise genes within potentially selected regions. Gene Ontology Enrichment Analysis (GOEA) is a statistical method for identifying gene functions which are over-represented in a set of genes analysed, taking into account the entirety of genes present in the genome. This analysis owes its name to the Gene Ontology (GO) (Ashburner et al, 2000), which seeks to organise all current knowledge about gene and gene product attributes in a controlled vocabulary, that is, a network of biological descriptors or terms interconnected by hierarchical semantic relationships. Terms

are divided into one of three macrocategories: Molecular Function (MF), Cellular Component (CC) and Biological Process (BP).

Each gene or gene product is annotated with one or more GO terms describing its function as a whole and allowing for an analytical and thus automatable comparison across other genes. This type of analysis is widely used in gene expression studies, where the objective may be to determine which biological pathways are differentially regulated in various tissues (Tsvetkov et al, 2021), or to characterise the effects of certain drugs in pharmacological trials. When applied to GWSS, functional enrichment analysis can be used to attain an overall picture of processes undergoing selective pressure, which may hardly emerge by simply contemplating a heterogeneous list of genes (Song et al, 2022). It is, however, advisable to prudently portray findings without over-interpreting the meaning of enriched functions, as this can lead to rather speculative storytelling (Pavlidis et al, 2012).

Instead, GOEA should serve as a preliminary approach to understand which external selective pressures might be playing a part in the evolution of species, and pave the way for further and more targeted studies. For instance, if sufficient environmental data are available, principal coordinates analysis (PcoA) can be carried out to determine which key factors are driving the selection of genes with over-represented functions. If the correlation between external causes and selected genes is confirmed and if the species of interest can be studied in an experimental context, expression analyses and knockout experiments can further investigate the candidate genes. For endangered species where this is infeasible, GOEA findings can nevertheless provide valuable insight and inform conservation actions.

3.2. Methods

3.2.1. Phasing

Prior to conducting haplotype-based selection scans, computational phasing was carried out on the quality-filtered SNP data set described in Chapter 2, including 3.46 million SNPs for 44 Italian, 12 European and 5 hybrid pike. To assess which algorithm and parameters were most suitable for this study, chromosome 1 comprising 175K SNPs was phased using 3 different programmes: SHAPEIT v2 (Delaneau et al, 2013) and SHAPEIT v4 (Delaneau et al, 2019), run with default parameters, and Beagle v5.2 (Browning et al, 2021) with 3 different sets of parameters: 40cM windows and 12 iterations each, 20cM and 24 iterations, and 10cM and 48 iterations. Then, algorithm performance was compared through phase concordance and imputation accuracy. For the latter analysis, an increasing number of SNPs (from 1% to 40% of the total data set) were set as having a missing genotype and, after phasing and genotype imputation, compared to the original genotype calls. The most efficient algorithm was chosen based on a trade-off between computation time and accuracy, and was used to phase the remaining 24 chromosomes.

3.2.2. Genome-wide scans for selection

For these analyses, custom scripts were developed in Python 3.6 (Van Rossum & Drake, 2009) which make use of data analysis libraries pandas v1.1.5 (McKinney, 2010), numpy v1.19.5 (Harris et al., 2020) and statsmodels v0.12.2 (Seabold & Perktold, 2010), data visualisation libraries matplotlib v3.3.4 (Hunter, 2007) and plotly v5.5.0 (Plotly Technologies Inc, 2015), and genomic data manipulation software BEDTools v2.28.0 (Quinlan & Hall, 2010), VCFTools v0.1.15 (Danecek et al, 2011) and Plink v1.9 (Chang et al, 2015).

In both allelic-frequency and haplotype-based GWSS approaches, hybrids were excluded from the data set of 3.9 million quality-filtered genomic SNPs described in Chapter 2, in order to detect selection signals in either Italian (n = 43) or European pike (n=12). Briefly, outlier regions were detected and genes overlapping such candidate regions were identified based on the publicly available *Esox lucius* genomic annotation file version 3 (Ensembl genome build Eluc_v3, NCBI accession GCA_000721915.3). Lastly, functional enrichment analysis was carried out separately on sets of annotated genes identified through each approach. Specifics of the two methods are described in the following two sections.

3.2.2.1. Allelic frequency-based approach

F_{ST} (Weir & Cockerham, 1984) across the two species of pike was calculated for each SNP using Plink v1.9 option “--fst”. Species were distinguished by loading a cluster ID file using option “--within”, where the first two columns of the file contained individual names and the third column contained cluster IDs, either “0”, “1” or “NA” for Italian pike, European pike and hybrids, respectively (the latter being excluded from the analysis).

A custom Python script has been used to calculate the windowed average F_{ST} using a rolling window obtained by dividing each chromosome into windows of 150 Kbp with an overlap of 75 Kbp and calculating the average of the per-SNP F_{ST} values.

Following the common practice in F_{ST} scans, an outlier approach was implemented (Yang et al, 2014; Ford et al, 2015; Candy et al, 2015; Ahrens et al, 2018). The genome-wide empirical distribution of windowed F_{ST} was computed and windows having an F_{ST} above the 99th percentile were classified as outliers. Outlier windows that were less than 150 Kbp apart were merged into broader regions. Genes overlapping selected regions were identified through the intersection between the genomic annotation file and genomic intervals using BEDtools intersect command.

3.2.2.2. Haplotype-based approach

Genotype files were loaded onto Rstudio (RStudio Team, 2021) and values of EHHS, XP-EHH and Rsb, along with their respective P-values, were calculated with the R package rehh (Gautier et al, 2017; Gautier & Vitalis, 2012). Analyses were conducted separately for each chromosome for technical reasons.

Specifically, genomic data were first converted to objects of class haplohh with the function data2haplohh(). Then, scan_hh() was used on haplohh objects to compute integrated EHH (iHH), integrated EHHS (iES) and integrated normalised EHHS (inES) for all markers within each chromosome. XP-EHH and Rsb were calculated for each marker using ies2xpehh() and ines2rsb(), respectively, and indicating in both cases popname1 as “Italian” and popname2 as “European”. The order in which population labels are given determines the directionality of the statistics, with positive and negative values of XP-EHH and Rsb representing selection in Italian and European pike, respectively.

Outliers were identified following a custom pipeline based on Gautier et al. (2017) that I modified at several steps to be more conservative. First of all, the frequency of false positives due to multiple testing was reduced by applying Benjamini-Hochberg False Detection Rate (FDR) correction (Benjamini & Hochberg, 1995) to P-values and setting a significance threshold of 0.01. A rolling window approach was carried out by dividing the entire genome into windows of 150 Kbp with an overlap of 50%. The overlap across windows allows detection of signals which could otherwise be imperceptible if located at the boundary between windows.

Because XP-EHH and Rsb are directional, selection signals were detected separately in Italian and European pike by only considering either positive or negative values of either statistic at a time. Therefore, for each species, a window was considered as a statistical outlier if it contained at least 2 SNPs with an adjusted P-value less than 0.01 for any combination of the two statistics considered, XP-EHH and Rsb. Genes overlapping outlier windows were identified as described for F_{ST}. Candidate windows, regions and genes

detected through either positive or negative XP-EHH and Rsb values are henceforth referred to as the Italian (ITA) and European (EUR) sets, for simplicity. Lastly, EHHS profiles of both species were plotted for the SNP with the most extreme XP-EHH or Rsb value within particularly interesting regions showing functional homogeneity.

3.2.3. Functional Enrichment Analysis

The two GWSS methods used yielded three sets of candidate genes. Two of them corresponded to haplotype-based signals in either Italian or European pike, and another to allelic frequency divergence across species. To assess whether certain gene functions were statistically over-represented in each set, two approaches were implemented which I termed region-wise and genome-wise, to avoid confusion with the“genome-wide” method used in the context of GWSS.

The reason for this is to increase the probability of detecting true signals of functional enrichment due to the complexity of trait inheritance. Indeed, most phenotypic traits are polygenic rather than monogenic (Shi et al, 2016), meaning that they derive from the synergy of multiple alleles scattered along the entire genome. In enrichment analyses, it is possible for the functional signal of such genes to become diluted when considering the genome in its entirety. It may be overridden by a stronger signal from another polygenic system or by localised genic clusters which often present a great number of copies of similar genes.

As a consequence, a genome-wise GOEA analysis will yield only the strongest overall functional enrichment signals and, while these may or may not correspond to all of the traits truly under selection, they nonetheless provide an overall hint as to the main selective processes acting in a population. Consequently, carrying out GOEA in a region-wise manner increases the resolution at a local scale and allows for fainter signals to emerge.

Functional enrichment of GO terms was performed with a Python implementation of software gProfiler v1.0.0 (Reimand et al, 2007), which retrieves data from the Ensembl database corresponding to version 4 of the *E. lucius* genome (Eluc_v4, GenBank accession GCA_004634155) as well as annotations from the Gene Ontology (released 2021-12-15). It is also able to retrieve and integrate data from the Kyoto Encyclopedia of Genes and Genomes, KEGG (released 2021-12-27), which is a pathway-oriented source of genomic annotations.

In Jupyter lab, a gProfiler search instance was initiated and stored in an object with the command gp = Gprofiler(return_dataframe = True). For each gene set, gene ID codes in the form of numerical NCBI accession codes were passed as a list to the function gp.profile(organism = 'elucius', query = genelist). Default parameters were used, including P-value correction for multiple testing using the “g_SCS” method, developed ad hoc for this application (Reimand et al, 2007), with a significance threshold of 0.05.

In the region-wise approach, a label describing the general function of each region was determined, where possible, based on the GO term with lowest term size (i.e. lowest hierarchy) that encompassed all other GO terms in that region.

3.3. Results

3.3.1. Phasing

In the assessment of phasing methods, SHAPEIT2 was found to perform slightly poorer for some individuals compared to all other phasing programmes, yielding an overall diminished phase concordance (Fig. 3.1). Moreover, it could not handle more than 10% missing data when imputing genotypes (Fig. 3.2). SHAPEIT4 performed similarly to Beagle 5.2, but the latter was chosen because it provided higher phase concordance across different parameter combinations. Of the three configurations used for Beagle 5.2, the 40 cM windows with 12 iterations each was preferred due to the trade-off between computational

time and performance in terms of imputation accuracy. A total of 3.46 million SNPs were thus phased this way.

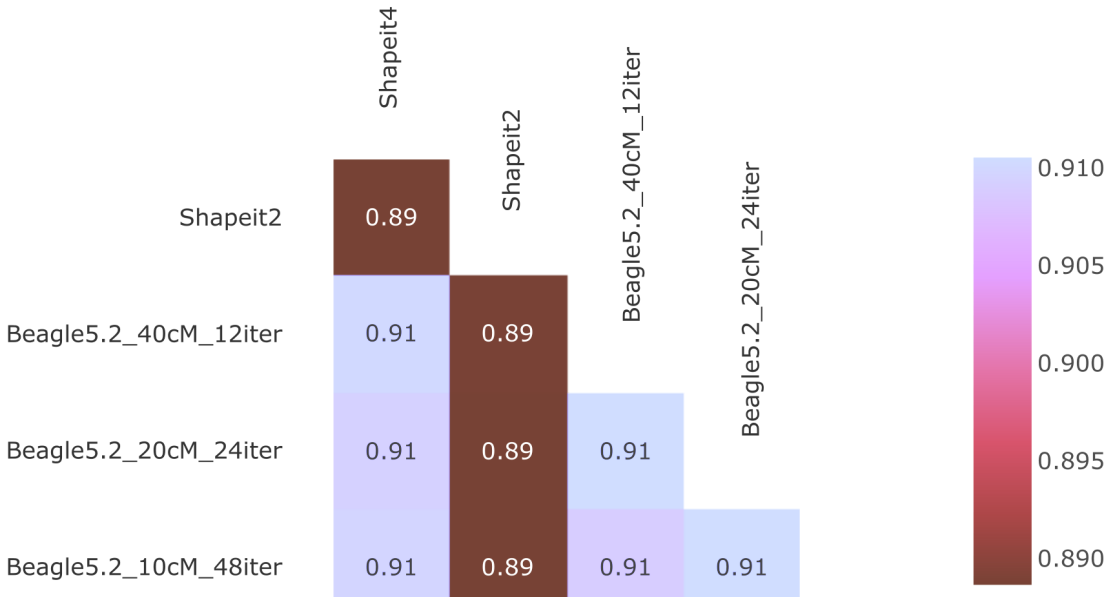


Fig. 3.1 Phase concordance across different methods. For each pairwise comparison between phasing methods, 175K genotypes from chromosome 1 were compared per-individual and phase concordance was calculated as the proportion of genotypes with matching phase. Phase concordance averages across all individuals is shown in the heatmap.

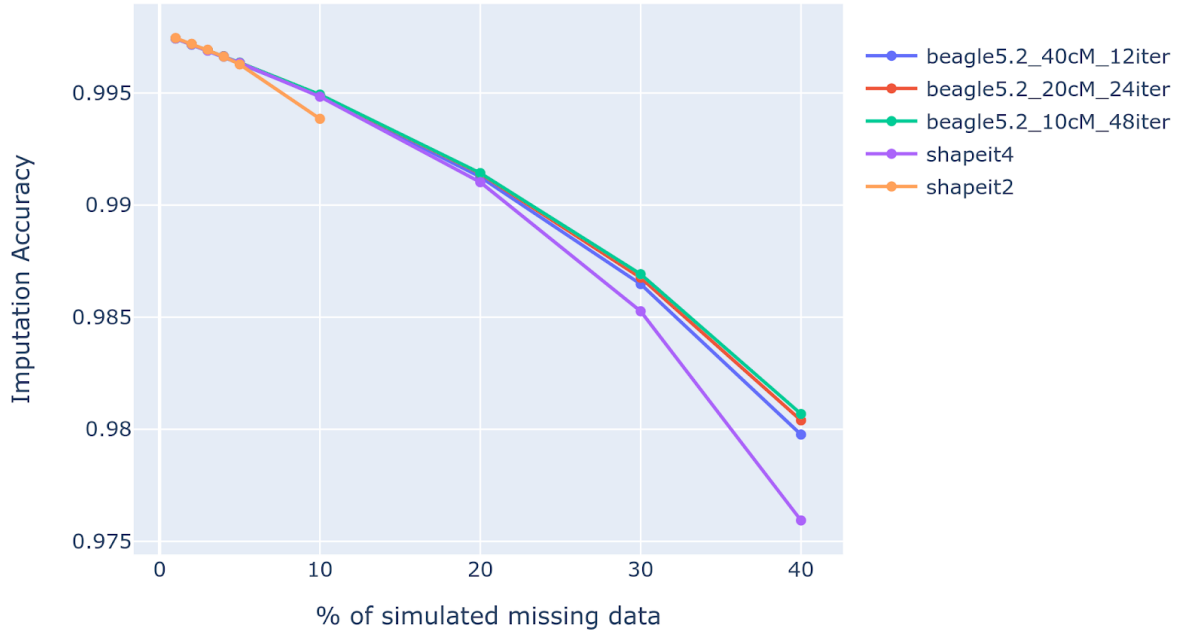


Fig 3.2 Genotype imputation. The performance of the different phasing methods used is plotted in different colours. The horizontal axis shows the increasing percentage of missing data simulated by masking known genotypes in chromosome 1, while the vertical axis is the proportion of correctly imputed genotypes.

3.3.2. Genome-wide scans for selection

3.3.2.1. Allelic frequency-based approach

Overall, general allelic differentiation between Italian and European pike estimated by averaging all values of F_{ST} was 0.53 ± 0.08 (Fig. 3.3A). Nearly all chromosomes showed a similar trend in terms of F_{ST} differentiation with Gaussian-like curves except for chromosome 24, which displayed more variance resulting in a wider and more uneven distribution (Fig. 3.3B).

Out of 10,565 overlapping genomic windows analysed, 107 (1.01%) had an F_{ST} value above the 99th percentile of the empirical distribution and were considered as potentially under selection. After merging windows that were less than 150 Kbp apart, 52 candidate regions were identified. Average length of regions was 532.2 ± 241.8 Kbp, the shortest spanning 300 Kbp and the longest 1.6 Mbp. Candidate regions were distributed in irregularly dispersed clusters along the genome with some chromosomes presenting a higher density of occurrences, such as chromosomes 5 (n = 11), 11 (n = 9) and 22 (n = 6), and some presenting none at all (Fig. 3.4). A total of 1005 genes were found that overlapped these regions.

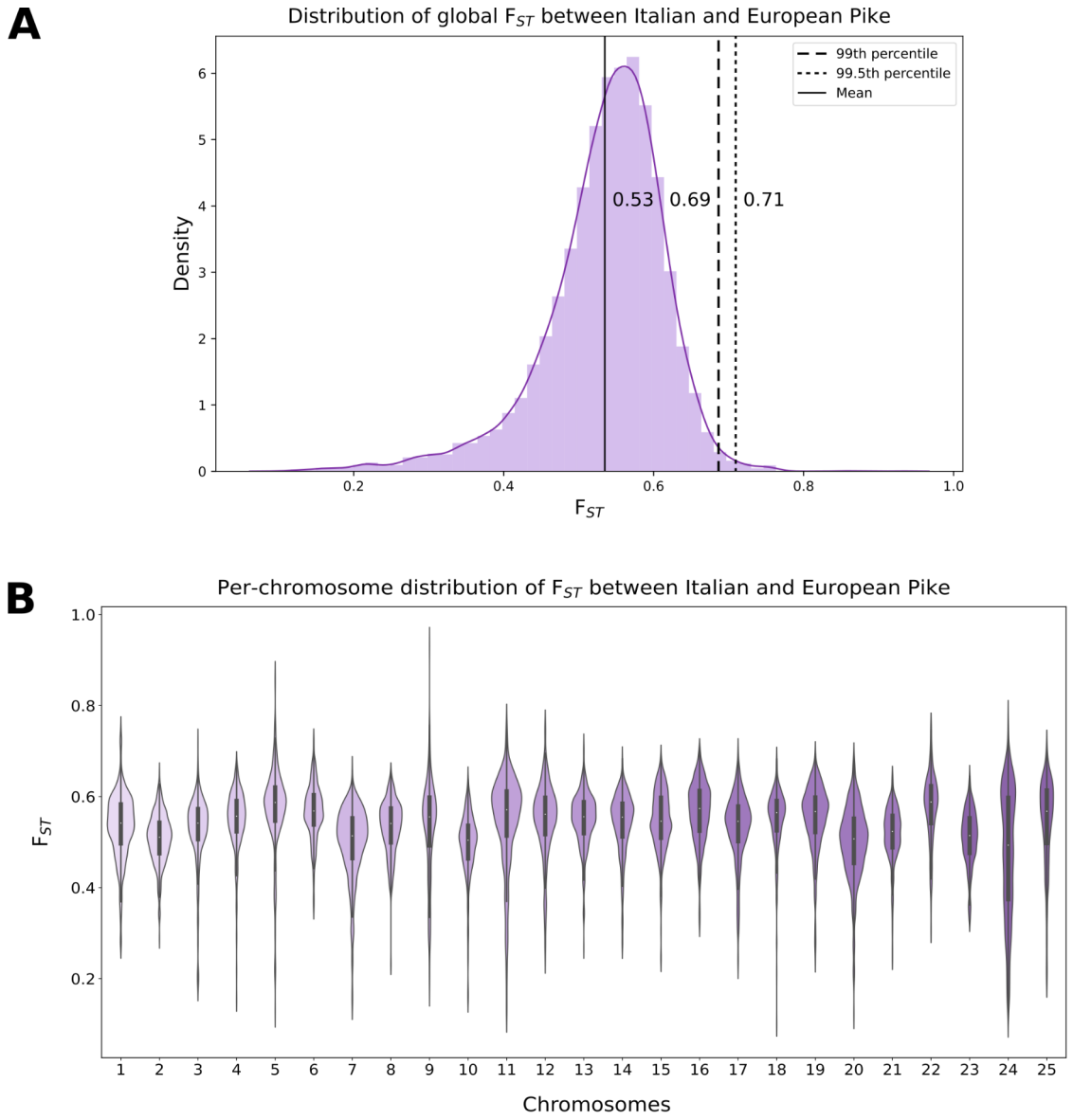


Fig. 3.3 Empirical distribution of genome-wide (A) and chromosome-wise (B) F_{ST} between Italian and European pike, estimated by averaging across windows of 150 Kbp with a 50% overlap. The thicker line inside each violin plot represents the interval from the first to the third quartile, that is, where 50% of values lie.

3.3.2.2. Haplotype-based approach

Using cross-population extended haplotype homozygosity metrics XP-EHH and Rsb, 161 (1.5%) and 191 (1.8%) out of 10,565 genomic windows were identified as potentially under selection in Italian and European pike, respectively. Merging windows by proximity produced 65 ITA and 83 EUR regions that were distributed in irregularly distanced clusters across the genome (Fig. 3.4 A). Region varied in length from 300 Kbp to 1.4 Mbp, with an average of 535.1 ± 177.2 Kbp.

Total lengths covered by candidate regions were calculated chromosome-wise for each type of signal considered, comparing both sets of species-specific outliers with non-directional F_{ST} outliers. Interestingly, chromosomes having the greatest frequency of species-specific signals had the fewest F_{ST} signals and vice versa (Fig. 3.4 A). Moreover, Fig. 3.5 details a section of a candidate region in chromosome 1 displaying EHHS profiles at two SNPs which have extreme XP-EHH and Rsb values. The delayed haplotype homozygosity decay in Italian pike compared to European pike in these instances point to a selective sweep in the former.

Lastly, a total of 2301 genes were found to overlap candidate regions: 1163 (49.28%) and 1167 (49.46%) were exclusive to Italian and European pike, respectively, while just 29 genes (1.26%) were common to both sets.

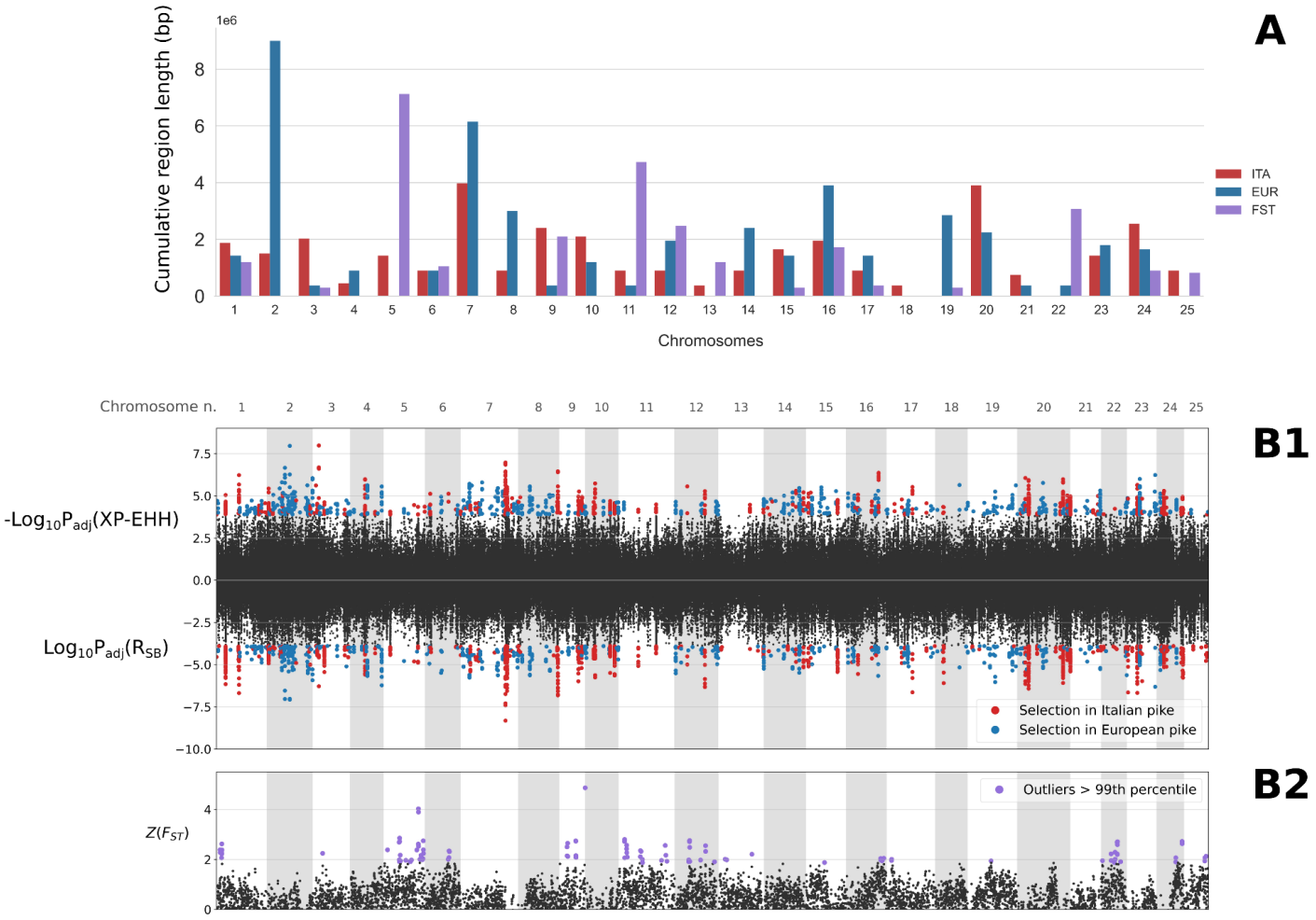


Fig. 3.4 A) Cumulative length of candidate regions across different selection signals: haplotype-based GWSS in Italian Pike ("ITA") and in European pike ("EUR"), as well as allelic frequency-based ("FST"). **B1) Genome-wide distribution of haplotype-based metrics**. Positive and negative values correspond to the logarithmic adjusted P-values of XP-EHH and and Rsb, respectively. The two metrics are displayed mirrored to emphasise similarities and differences across methods. Outliers are shown in red or blue depending on the sign of the metric, that is, if selection is acting in Italian or European pike, respectively. **B2) Genome-wide distribution of Z-scaled F_{ST}**. Each dot represents the average ZF_{ST} across a 150 Kbp rolling window with a 50% overlap. Outliers above the 99th percentile of the empirical distribution are colored in purple.

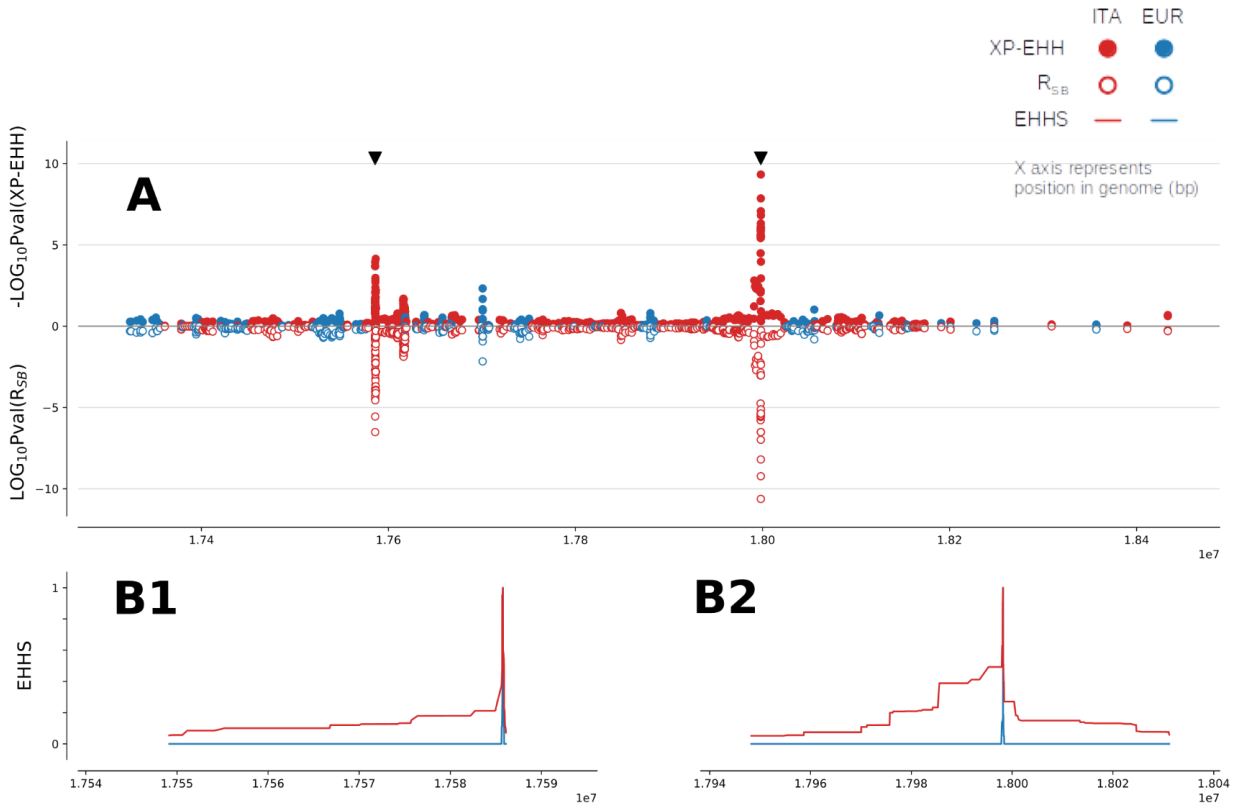


Fig 3.5 A) Region of chromosome 1 showing two peaks of XP-EHH and Rsb signals and **B) EHHS profiles for the corresponding focal SNPs**, indicated with black triangles in (A). Red and blue represent Italian and European pike, respectively. In particular, the colour of the dots in the Manhattan plot represents the sign of the haplotype-based metric. Genomic positions are indicated in bp.

3.3.3. Functional enrichment analysis

Overall the genome-wise GOEA only yielded results for the set of genes corresponding to selection in Italian pike, derived from the haplotype-based GWSS approach. Olfactory perception and functions related to the immune system were significantly enriched (adjusted P-value < 0.01).

However, the region-wise approach was much more fruitful. About one-third of ITA and EUR regions and one-sixth of FST regions presented significantly enriched gene functions. Among these, nearly all showed high functional within-region concordance, with only a few resulting in enriched GO terms that could not be directly traced back to the same pathway. Such was the case of one FST region in chromosome 5 (12.38 - 13.95 Mbp). This area displayed significantly overrepresented GO terms (P-value < 0.05) related both to muscle structure and/or function as well as metabolism.

Most regions were nevertheless functionally homogeneous (Tab. 3.1). When considering region-wise functions conjointly for each of the three sets (ITA, EUR and FST), a broader picture emerges. Italian pike-specific signals point to selection mainly in sensory perception pathways (olfactory and visual), immune system and synaptic transmission as well as in some metabolic pathways, genetic regulation and embryogenesis. While functions potentially under selective pressure in European pike show a similar pattern, especially with regard to olfaction, metabolism and genetic regulation, there are some differences. Indeed, there are several GO terms associated with tissue repair, molecular signalling and cell cycle regulation.

Interestingly, the signals of selection related to olfaction correspond to three different clusters of olfactory receptor genes. Chromosome 1 contains two such clusters, with the first undergoing selection in Italian pike (region 17,500 Kbp – 17,675 Kbp, visualized in Fig. 3.5.B1) and the second in European pike (region 30,775 Kbp – 30,950 Kbp), while the last

cluster occurs on chromosome 7 (region 36,900 Kbp – 37,200 Kbp) and is under selection in Italian pike.

Lastly, region-wise GOEA in F_{ST} outliers also revealed functions involved in metabolism, genetic regulation and cell cycle pathways. Unique to this set, though, there were several GO terms linked to nucleoside transmembrane transport and voltage-gated cation channel activity and a few related to the metabolism of xenobiotic agents, osmoregulation and melanogenesis.

		Functional Enrichment method	
GWSS method	**Species**	**Genome-wise**	**Region-wise**
Haplotype-based (XP-EHH and Rsb)	**Italian pike**	Olfactory perception **(8)** Immune system **(4)**	Olfactory perception **(37)** Immune system **(35)** Synaptic transmission **(30)** Metabolism **(10)** Vision **(7)** Retinoid metabolism **(6)** Genetic regulation, transcription and translation (3) Embryogenesis (2)
	European pike	No enrichment	Olfactory perception **(19)** Tissue repair **(7)** Metabolism (6) Molecular signaling (3) Genetic regulation, transcription and translation (3) Cell cycle (2) Immune system (1) Cell junction **(1)**
Allelic frequency-based (F_{ST})	**Either**	No enrichment	Metabolism **(22)** voltage-gated cation channel activity **(9)** Muscle structure or contraction (3) Nucleoside transmembrane transporter **(7)** Genetic regulation, transcription and translation **(4)** Melanogenesis **(1)** Cell cycle **(1)** Reproduction (1) Osmoregulation (1) Xenobiotic metabolism **(1)**

Table 3.1 Main results of Gene Ontology Enrichment Analysis for genes found in association with Italian-specific, European-specific and unpolarized selection signals. Functional descriptors are based on the lowest-grade GO term capable of encompassing all other GO terms within a region. Values between parentheses indicate the number of significantly enriched GO terms identified (P-value < 0.05). Bold values indicate the presence of at least one highly significant (P-value < 0.01) GO term.

3.4. Discussion

The allelic frequency- and haplotype-based GWSS approaches captured different aspects of selection signals that are potentially acting along the pike genome. Not surprisingly, identified regions were almost entirely complementary across the two methods (Fig. 3.4), as the former is likely indicative of either ancient or stronger selection, while the latter reveals ongoing or softer instances of selective pressure. Moreover, the relative abundance of putative soft and partial sweeps detected through XP-EHH and Rsb compared to the scarcity of signals compatible with hard sweeps is in line with considerations regarding soft sweeps as the prevalent form of evolution (Csilléry et al, 2018; Pritchard et al, 2010; Schrider & Kern, 2017). In fact, it is more likely for standing genetic variation to gain a beneficial function as a result of changes in the environment than for a favourable *de novo* mutation to arise *ad hoc* in the right place, at the right time. Furthermore, pleiotropic alleles underlying an advantageous trait will also display a pattern resembling that of a partial rather than a hard sweep, and it is well known from human studies that complex polygenic traits are much more common than monogenic ones (Shi et al, 2016).

It must be noted that selective pressures are not the only factors capable of leaving footprints in the genome. Certain demographic events such as bottlenecks, migration and population expansion may give rise to similar signals. For example, while genetic hitch-hiking produces an excess of high- and low-frequency alleles, hence skewing the Site Frequency Spectrum (SFS) towards extreme frequencies (Braverman et al, 1995; Fay & Wu, 2000), rapid population growth can lead to an excess of low-frequency variables as well (Fu, 1997; Ramírez-Soriano et al, 2008). This is because of increased opportunities for new mutations, and because the resulting larger effective population size minimises random genetic drift that would otherwise tend to remove rare variants from the gene pool (Hartl & Clark, 2007). Bottlenecks can also produce analogous patterns in the genome depending on the strength, duration and age of the event. During an abrupt and drastic contraction in

population size, much genetic diversity is lost and thus there is a reduction in the quantity of low-frequency alleles (Nei et al, 1975). However, if the bottleneck is not as pronounced or is very recent, or if the population subsequently undergoes a significant expansion, the SFS curve will more closely resemble that of a selective sweep, with a decrement in intermediate-frequency alleles (Ramírez-Soriano et al., 2008).

In the case of well-studied species, or species with rather simple demographic histories, data simulated from models can be used as the null hypothesis of selective neutrality based on which to identify outliers. In our case, previous analyses portrayed a complex picture likely due to numerous human-mediated translocations spanning several decades (Welcomme, 1988; Lucentini et al., 2006, 2009, 2011; Bianco, 2013). As an alternative to model-based outlier detection, we opted to use distributions of genome-wide statistics to represent the null hypothesis of neutrality, on the basis that demographic processes have genome-wide effects while selection acts locally (Cavalli-Sforza, 1966; Lewontin & Krakauer, 1973; Akey et al, 2002). Naturally, disposing of a large set of widespread genomic variants is essential for the statistical rigour of this type of approach. Compared with early multi-locus studies (Eckert et al, 2009; Hamblin et al, 2004; Kayser et al, 2003; Orengo & Aguadé, 2004; Schlötterer, 2002), FDR-controlled detection based on hundreds of thousands of SNPs can greatly increase resolution when disentangling selection from other confounding signals (Akey et al, 2002).

The GWSS methods implemented in this study detect signals of positive selection, as opposed to negative or background selection. Positive selection is associated with selective pressure from direct interactions with the environment, including ecological and climate conditions, pathogens and other organisms (Levasseur et al., 2007; Fumagalli et al., 2011). Thus, functional enrichment analysis provides indications of which external factors may have contributed to shaping the genomic landscape of Italian and European pike.

As previously mentioned, selection signals detected through F_{ST} are presumably more ancient than those detected by haplotype-based methods. Hence, it seems likely that functions linked to osmoregulation and melanogenesis underwent evolution when the two

species diverged, compatible with a model of allopatric speciation. Indeed, this is known to be the case in well studied species of stickleback (*Gasterosteidae* Bonaparte, 1831), which show evidence of major adaptive radiation during colonisation of new niches after the last glaciation just 12,000 years ago (Divino et al, 2016; Kirch et al, 2021; Marques et al, 2018).

Likewise, findings from haplotype-based methods are likely linked to dynamic and perhaps ongoing processes of adaptation to environments and biocenoses. In pike, factors such as water turbidity, which are in turn determined by a combination of biotic and abiotic elements (e.g. local vegetation, hydrogeography and temperature), appear to influence phenotypic traits (Jepsen et al, 2001; Lehtiniemi et al, 2005). Further, being an ambush predator, pike rely heavily on sensory perception, and turbidity has been shown to have a negative effect on visually-guided but not on olfaction-driven predation success (Lunt & Smee, 2015). This is consistent with evidence of selection on olfactory receptors identified in this study. Nonetheless, more robust understanding of the processes driving the selection of biological functions will require further investigation using experimental or environmental data.

3.5. Conclusion

- Different approaches for detecting signals of genome-wide selection were carried out in Italian and European pike to better understand the basis of past and current molecular adaptation.
- Haplotype-based methods using metrics XP-EHH and Rsb revealed species-specific regions with ongoing genomic divergence, while a genomic scan implementing F_{ST} shed light on evolutionary processes that might have taken place at the divergence between the two species.
- Haplotype-based signals were more frequent than allelic frequency-based ones, and they were also complementary, being located in rarely overlapping regions of the genome.
- Functional enrichment analysis points to certain key molecular pathways as potentially under selective pressure, including olfactory perception, metabolism, and immune response.
- While more in-depth studies are needed to provide experimental support, this study represents the first comprehensive analysis of genomic selection in Italian pike and highlights the need to preserve this endangered species by providing evidence of unique molecular adaptation.

Chapter 4: Development of a High-Density SNP Panel for Marble Trout

---

Abstract

Besides being a prized game fish (Meraner & Gandolfi, 2017), the marble trout (*Salmo marmoratus*, G. Cuvier 1829) plays an important role as top predator in Italian alpine ecosystems (Klemetsen et al, 2003). However, it is threatened by genetic erosion due to hybridisation with stocked brown trout, and determination of its complex ancestry is often hindered by the limited resolving power of traditional genetic markers (Gratton et al, 2014). The aim of the work described in this chapter was to identify a high-density SNP panel from pooled Whole Genome Sequencing (WGS) data to be included in an Axiom SNP assay for assessment of Italian marble trout management units (MU) and detection of hybrids. A total of 19.6 million high-quality SNPs were discovered in one exotic brown trout and six marble trout populations from the Italian Alps. I present a cost- and time-efficient method for filtering out potentially introgressed alleles as well as pseudo-SNPs from paralogous regions of the trout genome, which is a common issue in salmonid genomic studies. To increase genotyping success rate, probes centered around each SNP were tested *in silico* for specificity in the trout genome, yielding a final set of 8.4 million unique probes. Furthermore, species and population ancestry-informative SNPs were identified which can be used to further filter the selected probes, depending on the final target size of the array. The SNP panel that I developed paves the way for large-scale and resource-efficient high-throughput genotyping efforts that will help clarify the current genetic integrity of marble trout populations, define management units (MUs) and provide insight for future conservation actions for marble trout.

4.1. Introduction

Marble trout (*Salmo marmoratus*, G. Cuvier 1829) is a freshwater salmonid native to the Alpine freshwater system, present in mountainous areas of Croatia, Slovenia, Austria, Switzerland but mostly in Northern Italy (Povz, 1995). A highly prized species for anglers, it is currently endangered in most of its native distribution range due to habitat alteration and hybridization with farmed brown trout of Atlantic lineage (*Salmo trutta*, Linneus 1758), the stocking of which has been ongoing for decades if not centuries (Caputo et al, 2004; Caputo, Giovannotti and Splendiani, 2010). Although the species is listed as Least Concern in the IUCN Red List of Endangered Species (Crivelli, 2006), its genetic composition is so heavily compromised that it is estimated no pure marble trout remain in most localities (Meraner & Gandolfi, 2017).

Conservation programmes to locally rehabilitate marble trout populations are mostly based on phenotypic or genetic marker-assisted supportive breeding activities. However, as hybridisation progresses across generations, markers currently employed in screening procedures, mainly microsatellites and mitochondrial DNA (mtDNA), often lack the resolution needed to disentangle the true extent of fine-scale introgression. While Saint-Pé et al. (2019) recently identified 12,204 SNPs in brown trout of the Atlantic and Mediterranean lineages, these might not be sufficiently informative in the marble trout populations hereby assessed. Indeed, Palombo et al. (2021) found that only about 900 out of 57 thousand SNPs from the Affymetrix rainbow trout SNP array (Palti et al, 2014) were polymorphic and thus informative about genetic identity in the Italian trout lineages they studied. Moreover, Pustovrh et al. (2012) previously identified 41 polymorphic SNPs in Slovenian populations of marble trout, but this number is incompatible with the objective of providing high-resolution insight into the genomic structure of marble trout. As a result of this, conservation programmes may be currently lacking comprehensive information to prevent marble trout populations from

permanently losing genomic adaptations essential to their survival under anthropogenic threats and climate change.

In this chapter, I develop a high-density Single Nucleotide Polymorphism (SNP) panel to be implemented in a genotyping array allowing for estimation of introgression with non-native brown trout and distinction of six Italian marble trout populations. Low coverage Whole Genome Sequencing (WGS) data were generated for each individual and then pooled by population. Generated SNPs were filtered to ensure that the final panel consisted only of high quality, informative genomic markers compatible with ThermoFisher Axiom SNP assay technology and excluding any pseudo-SNPs deriving from paralogous regions of the trout genome.

4.2. Methods

4.2.1. Samples, Sequencing and Variant Calling

Prior to the onset of this doctoral project, genetic material had been collected from the trout individuals shown in Table 4.1. Small portions of caudal fin were clipped and stored in 95% ethanol until the processing phase. Whole genomic DNA was extracted and purified from the fin clips with a KingFisher Cell and Tissue DNA Kit (Thermo Fisher Scientific Inc., Fremont, CA, USA), according to manufacturer protocols.

A per-population pooling approach had been implemented to maximise the number of sequenced individuals while still allowing for robust population genetic analyses, with an average cumulative per-population sequencing depth of ~30X. Thus, Illumina 150 nt paired-end Whole Genome Sequencing (WGS) reads had been generated at Novogene Ltd (Cambridge, UK), individually at 3X depth for 58 marble trout and 8 brown trout of Atlantic lineage corresponding to a total of eight populations. The marble trout originated from six Italian localities, including rivers Adda, Adige Bozen (AdigeBZ), Adige Trento (AdigeTN), Avisio, Noce and Passirio, with an average of 9 individuals per locality (Tab. 4.2). These individuals presented a marble trout ancestry coefficient of at least 0.95 according to a

previous admixture survey using STRUCTURE software (Pritchard et al., 2000) and 15 microsatellites (data from Meraner & Gandolfi, 2018).

Individual	Population	Sampling date
ADD_1803	Adda	03/2005
ADD_1804	Adda	03/2005
ADD_1805	Adda	03/2005
ADD_1806	Adda	03/2005
ADD_1816	Adda	03/2005
ADD_1823	Adda	03/2005
ADD_1826	Adda	03/2005
ADD_1834	Adda	03/2005
ADD_1836	Adda	03/2005
ADD_1841	Adda	03/2005
APD_0177	Adige TN	03/2006
MUR_0768	Adige TN	03/2007
MUR_0777	Adige TN	03/2007
OSS_0737	Adige TN	03/2007
OSS_0742	Adige TN	03/2007
OSS_0744	Adige TN	03/2007
PDV_0042	Adige TN	03/2006
RMO_0666	Adige TN	11/2006
ZAM_0370	Adige TN	08/2006
CAL_0530	Avisio	11/2006
CAL_0532	Avisio	11/2006
CAL_0548	Avisio	11/2006
CAL_0551	Avisio	11/2006
IAS_0717	Avisio	11/2006
LAV_0596	Avisio	11/2006
MEZ_0309	Avisio	05/2006
PiP_1237	Avisio	06/2007
PoA_1240	Avisio	06/2007
PoA_1246	Avisio	06/2007
CA1_0226	Noce	03/2006

CA1_0227	Noce	03/2006
CA1_0232	Noce	03/2006
CA2_0444	Noce	10/2006
CA2_0448	Noce	10/2006
CA2_0456	Noce	10/2006
PES_0497	Noce	11/2006
PES_0500	Noce	11/2006
PRA_0289	Noce	03/2006
PRA_0290	Noce	03/2006
St150431	Adige BZ	10/2015
St150435	Adige BZ	10/2015
St150438	Adige BZ	10/2015
St150448	Adige BZ	10/2015
St150450	Adige BZ	10/2015
St150465	Adige BZ	10/2015
St150467	Adige BZ	10/2015
St150476	Adige BZ	10/2015
St150544	Adige BZ	11/2015
St150391	Passirio	10/2015
St150407	Passirio	10/2015
St150426	Passirio	10/2015
St150478	Passirio	11/2015
St150479	Passirio	11/2015
St150482	Passirio	11/2015
St150485	Passirio	11/2015
St150487	Passirio	11/2015
St150496	Passirio	11/2015

Tab. 4.1. Individual trout codes, sampling locations and date of sampling.

Population	Sample size (n)
Adda	10
AdigeBZ	9
AdigeTN	10
Atlantic	8
Avisio	10
Noce	10
Passirio	9

Tab. 4.2. Trout populations and sample sizes. Atlantic refers to brown trout of Atlantic lineage.

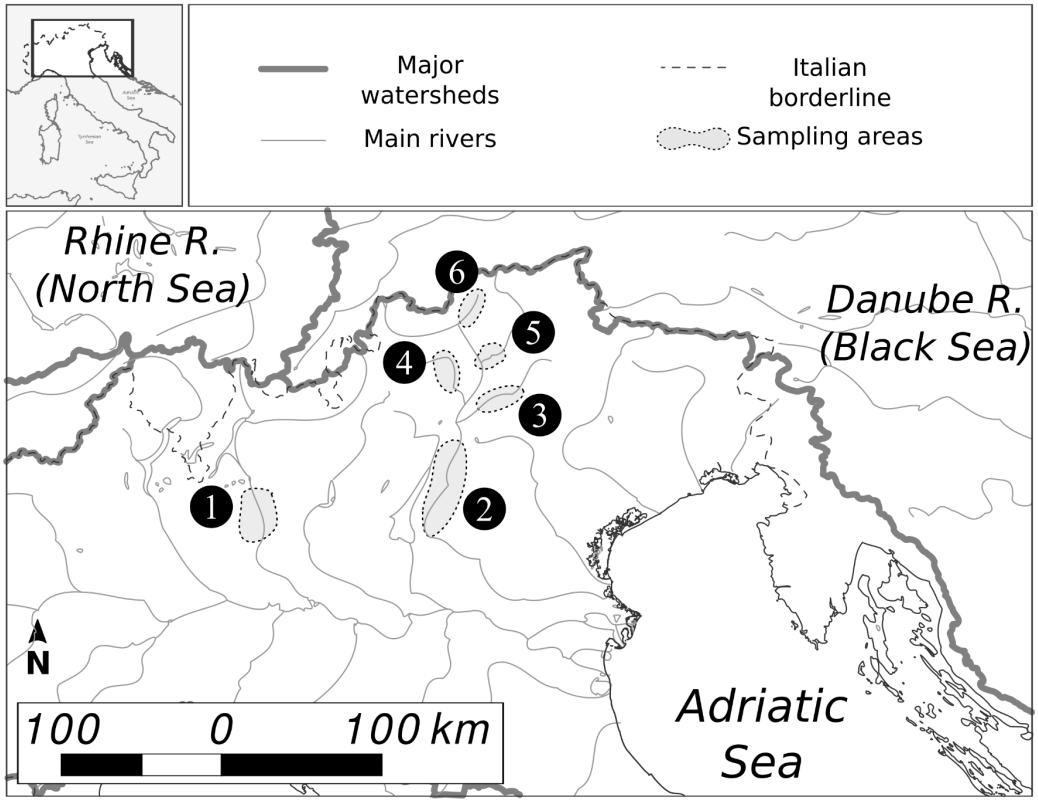


Fig. 4.1 Marble trout sampling locations. 1) Adda River; 2) Adige River (Trento); 3) Avisio River; 4) Noce River; 5) Adige River (Bozen); 6) Passirio River. Figure adapted from Meraner & Gandolfi (2018).

Read quality was checked with FastQC (Andrews, 2010) and poor quality bases were trimmed with Trimmomatic (Bolger et al., 2014) using the parameters “PE -threads 4 -phred33 LEADING:28 TRAILING:28 MINLEN:100” to ensure that surviving paired-end (PE) reads were at least 100 base pairs long, with base quality encoding corresponding to the Phred+33 format and a quality of at least 28 at the beginning and at the end of the read. Reads from individuals were aligned to the brown trout reference genome v1 (GenBank Assembly Accession GCA_901001165.1) using the bwa-mem aligner (Li & Durbin, 2009). Mapped reads were then pooled by population and variants were called according to GATK Best Practices v4 (McKenna et al, 2010; Poplin et al, 2017).

Genomic variants were filtered by only retaining markers that i) were biallelic SNPs, ii) were located on anchored chromosomes and iii) had a pooled read depth greater than 8 reads within populations, which is a rather conservative threshold chosen based on previous analyses to ensure a very low error rate.

4.2.2. Selection of SNP panel for species and population identification

4.2.2.1. Fine-scale introgression analysis

Because the marble trout is critically threatened by hybridisation with non-native brown trout and no single population is completely free of introgression (Meraner and Gandolfi, 2018), it is likely that some brown trout alleles are present in our pooled samples. Although all individuals have been selected on the basis of a previous microsatellite screening, genome-wide SNPs can reveal undetected introgression, as seen in Chapter 2 for the Italian pike. I hypothesise that reference alleles having low frequency (AF < 0.05) in marble trout populations while, at the same time, high frequency (AF > 0.5) in brown trout (therefore having a high probability of being derived from the introduced species) are

potentially introgressed. Because samples are pooled by population, allelic depth is used as a proxy of allelic frequency.

4.2.2.2. Identification of paralogous loci in Marble trout

Compared to other teleosts, Salmonids underwent a further genome duplication event approximately 96 million years ago (Allendorf & Thorgaard, 1984). Despite diploidization, the trout genome still presents some paralogous regions which have tetrasomic inheritance (Berthelot et al, 2014) and which can hinder correct variant calling. Indeed, alignment algorithms often fail to detect duplicated regions as paralogous and instead mistake them for homologous loci (Christensen et al, 2013). That is, if the amount of accumulated mutations is not enough for an algorithm to distinguish two loci, these are collapsed into one locus and mismatches are erroneously called as single nucleotide variants (SNVs), namely pseudo-SNPs in the context of this thesis. As a consequence of the experimental design favouring genetic diversity over sequencing depth, the population-pooled samples do not allow for discrimination of individual genotypes. Therefore, variants were considered pseudo-SNPs and filtered out when all populations showed a minor allele frequency (MAF) greater than 0.45 or a missing genotype at a given locus. Read depth was assessed to verify whether filtered loci met the requirement of having double the target sequencing depth.

4.2.2.3. Technical filters

After applying the filters previously described, sequences (henceforth termed probes) consisting of 71 bp regions with each remaining SNP in the centre were extracted from the brown trout reference genome. The exact length was determined following indications of the SNP assay manufacturer (ThermoFisher, Santa Clara, California, United States).

A Blast (Altschul et al, 1990) against the brown trout reference genome was used to exclude non-specific binding of probes (parameters “-task blastn -word_size 17 -max_target_seqs 10 -max_hsps 1”). Probes were considered unique and retained if there were no additional non-specific hits in the genome with percentage of identity greater than 90% and other requirements depending on the location of the match within the query. In

particular, for upstream sequences (first 35 nt of the probe), no more than one blast match was allowed after filtering for “query end ≥ 35” and “query start ≤ 3”, because any non-specific alignments which end before the 35th base imply at least one mismatch at the 3’ end of the probe, and those which start after the 3rd base would have an alignment length inferior to 90% of the probe length; both of these conditions impair *in vitro* annealing of probes. Following the same logic for the reverse strand, downstream (last 35 nt of the probe) filters were “query start ≤ 37” and “query end ≥ 68”. Probes were then labelled as either “upstream”, “downstream” or “across” depending on whether one part or both bind uniquely to the desired region.

After excluding non-unique matches, probes with SNPs presenting an inter-marker distance shorter than probe length (35 bp) were discarded, as well as loci with either G/C or A/T alleles, as these would require twice as many probes due to the architecture of the ThermoFisher Axiom SNP array.

4.2.2.4. Ancestry diagnostic SNPs

SNP sets which are informative of membership to different marble trout populations, thus diagnostic of ancestry, were identified based on criteria corresponding to the seven categories described below.

- **Category 1: Alternative alleles polymorphic in both trout species**

  These alleles have an alternative allele frequency (AF) > 0.1 in brown and in at least 3 out of 6 marble trout populations, with AF < 0.3 in all populations to exclude any paralogous regions which might have eluded previous filters. This category is useful for minimising ascertainment bias especially when genotyping new populations not included in this study.

- **Category 2: Species-specific monomorphic alleles**

    - **2A: The alternative allele is monomorphic in and exclusive to marble trout**. These are most adequate SNPs for identifying species.
    - **2B: The alternative allele is monomorphic in and exclusive to brown trout samples in this study.** Because the reference brown trout genome contains, by definition, the reference allele, these SNPs highlight differences between brown trout samples sequenced for this study and the brown trout reference genome.
- **Category 3: Alternative allele exclusive to marble trout, polymorphic in only one marble trout population and monomorphic elsewhere**

  This category is useful for distinguishing between species albeit with less precision compared to 2A.

- **Category 4: Alternative allele polymorphic in and exclusive to brown trout**

  Like 2B, this SNP subset provides insight into brown trout but not marble trout lineages. In other words, if the alternative allele is present it's likely from brown trout, but the opposite is not true.

- **Category 5: Alternative allele polymorphic in and exclusive to Marble trout**

  Alternative alleles present with AF > 0.1 in at least 3 out of 6 marble populations but absent in brown trout.

- **Category 6: Alternative allele monomorphic in and exclusive to only one Marble trout population**

  This is the most efficient category for identifying individual marble trout populations.

- **Category 7: Alternative allele polymorphic in and exclusive to only one marble trout population**

These SNPs are similar to those specified in category 6, but with less power to diagnose population ancestry.

The most informative loci are those that present an alternative allele which is fixed (i.e. monomorphic) in one species or subpopulation and completely absent from others. Examples of these are SNPs from categories 2A and 6, where the presence of the alternative allele is immediately indicative of marble trout ancestry. It must be noted that while category 2A reference alleles are diagnostic of brown trout ancestry, category 6 reference alleles are not necessarily so, as they can be found in other marble trout lineages.

4.3. Results

4.3.1. WGS read alignment and SNP discovery

Illumina Whole Genome Sequencing yielded a total of 6.99 billion raw 150 bp paired-end reads. Approximately 5.94 billion reads (85.1%) remained after trimming. Of these, 5.89 billion reads (99.1%) were mapped to the brown trout genome and a high percentage (90.4%) was properly paired. The variant calling pipeline generated an initial set of 30.1 million variants, comprising 23.4 million SNPs (96.7% of which are biallelic) and 6.4 million indels (70.7% biallelic). Filtering out low-quality, multiallelic variants occurring on unanchored scaffolds resulted in a final dataset of 19.6 million high quality biallelic SNPs.

4.3.2. Selection of SNP panel for species and population identification

4.3.2.1. Fine-scale introgression analysis

A variable number of potentially introgressed alleles from brown trout was detected in all sampled marble trout populations, with Adda and Adige Trento showing the least amount (0.013% and 0.018%, respectively, compared to all 19.6 million SNPs), and Passirio (0.052%) and AdigeBZ (0.051%) presenting the highest estimated introgression, as shown in Fig. 4.2. In total, 36,012 unique introgressed alleles were detected, corresponding to 0.183% of the SNP data set. The physical distribution of such alleles along the genome is shown in Fig. 4.3 and, with greater resolution and following a per-individual approach, in Fig. 4.4. A clustered pattern is often present, which can be attributed to the inheritance of haplotype blocks from brown trout. For practical purposes, only chromosome 1 is shown, however the trend is similar in all 40 chromosomes.

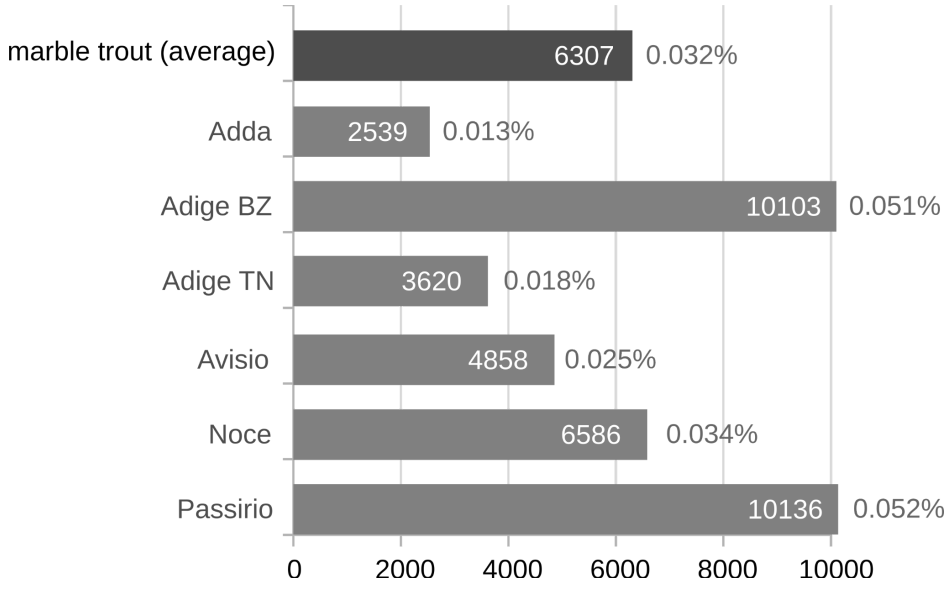


Fig 4.2. Relative amount of detected introgression in all sampled marble trout populations. Per-population and average number and percentage of potentially introgressed alleles within 19.6 million genome-wide SNPs.

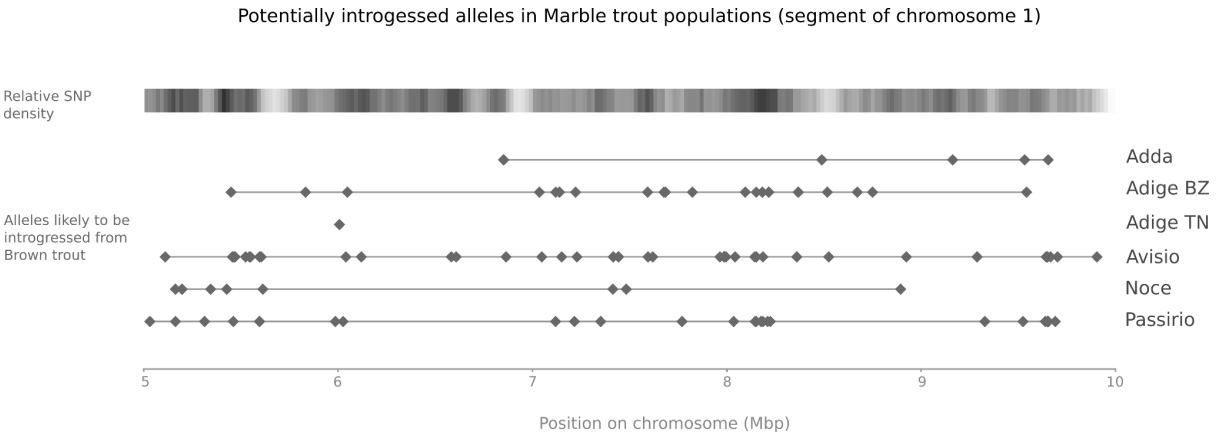


Fig 4.3. Putative introgressed alleles in marble trout populations in a 5 Mbp segment of chromosome 1. Above, the heatmap shows the relative SNP density. Below, diamond markers indicate the positions of alleles likely to be introgressed from brown trout into different marble trout populations, which fulfil the requirement of being rare in marble trout (frequency < 0.05) and common in brown trout (frequency > 0.5).

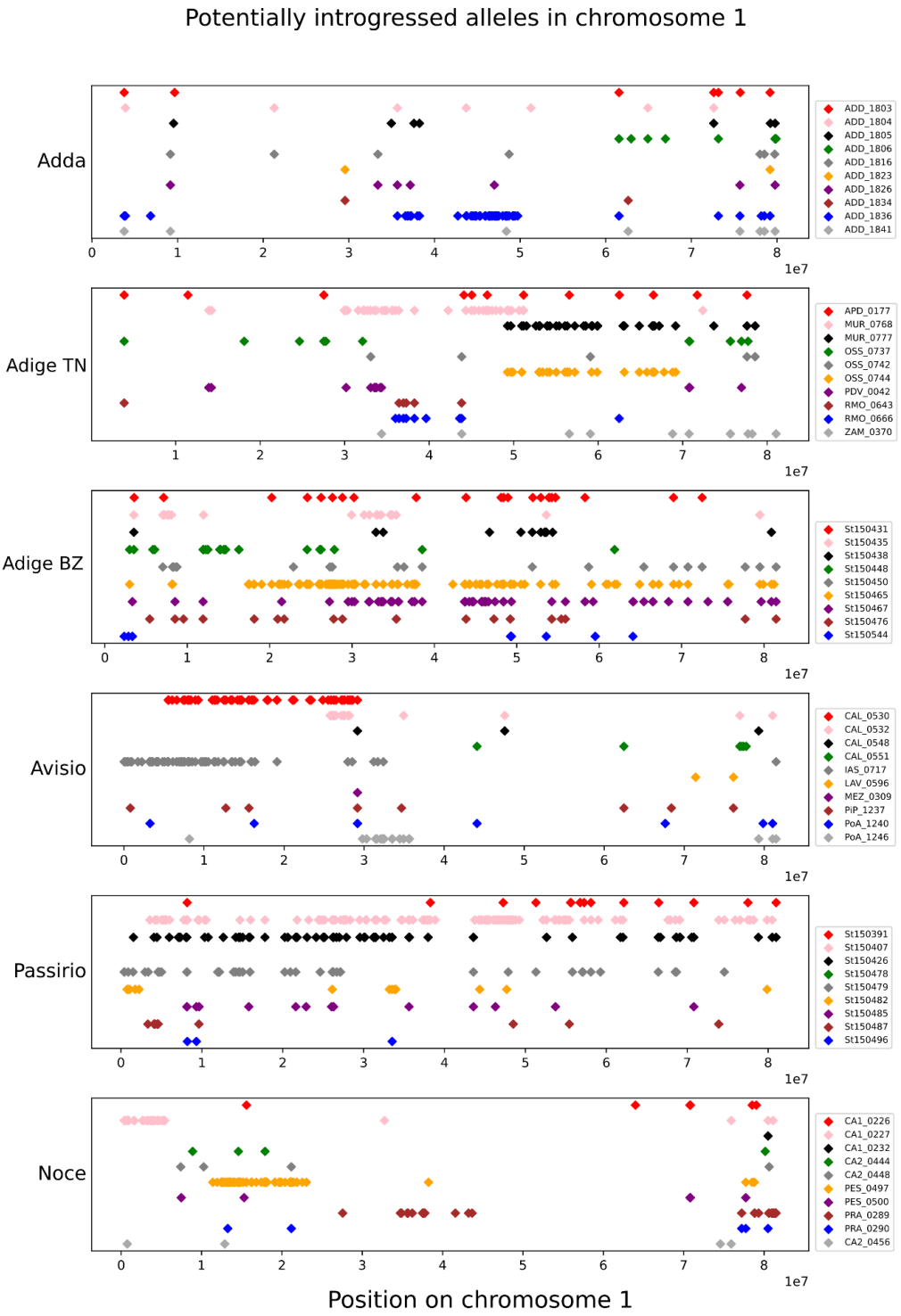


Fig 4.4. Putative introgressed alleles in chromosome 1 for each individual (shown in legend) and for each marble trout population. Chromosomal position is shown on the X axis as a multiple of 1e7 base pairs (bp). Clustered alleles are likely to be haplotype blocks introgressed from Atlantic brown trout. Such introgression patterns are present in all chromosomes (not shown), and in all sampled marble trout populations.

4.3.2.2. Paralogous loci

A total of 424 thousand loci containing putative pseudo-SNPs from paralogous regions (having allelic frequency between 0.45 and 0.55) were identified in all chromosomes. In particular, *ex-post* evaluation of per-population sequencing depth at these loci highlighted a peak at ~60X (termed as C in Fig. 4.5) in about half of the chromosomes, corresponding to double the sequencing depth at which genomic data was generated (peak B in Fig 4.5). This is indicative of pseudo-SNPs deriving from paralogous regions of the genome.

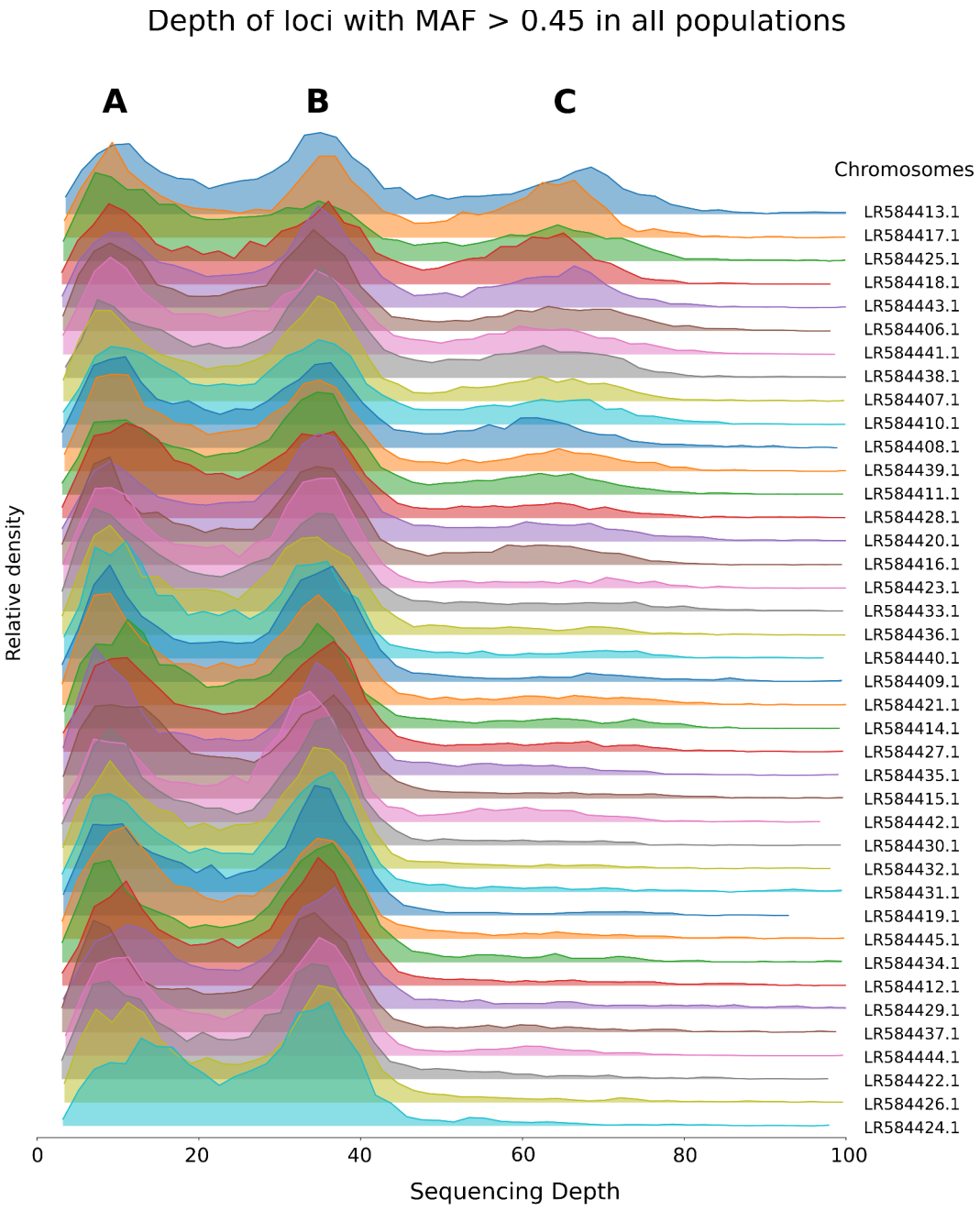


Fig 4.5. Sequencing depth of potentially paralogous pseudo-SNPs.
The joyplot shows the sequencing depth distribution of SNPs having MAF > 0.45 in all trout populations. The first peak at a depth of ~10X (A) is due to missing data, while the second (B) corresponds to the target depth at which data was sequenced (~30X). The peak at ~60X (C) corresponds to loci that have approximately twice the sequencing depth at which genomic data was generated, so they likely originate from paralogous regions.

4.3.2.3. Technical filters

The SNP set was further refined to 13.4 million SNPs after excluding loci that were potentially introgressed, pseudo-SNPs, had G/C or A/T alleles and were less than 35 bp apart. After generating 71 bp probes centred around these SNPs and discarding non-specific matches within the brown trout genome, 8.4 million unique probes were identified (Tab. 4.2), on average 210 thousand per chromosome. Moreover, approximately 7.3 and 7.2 million unique probes were identified when considering only the upstream or downstream half, respectively.

Chrom. number	Chrom. ID	Unique 35 bp upstream probes	Unique 35 bp downstream probes	Unique 71 bp probes
1	LR584410.1	256574	251881	298618
2	LR584445.1	239853	236728	272572
3	LR584416.1	223677	218999	258467
4	LR584420.1	273312	267938	318119
5	LR584433.1	278287	274746	316313
6	LR584406.1	220202	215901	253504
7	LR584430.1	197233	194859	225833
8	LR584407.1	179875	176569	206776
9	LR584409.1	180048	177315	204887
10	LR584419.1	188220	185722	216809
11	LR584438.1	67608	65815	82551
12	LR584441.1	265391	260466	308621
13	LR584428.1	287473	283339	326828
14	LR584411.1	300200	295140	348827
15	LR584415.1	196905	193954	226097
16	LR584431.1	219957	216863	252615

17	LR584426.1	205318	202230	235354
18	LR584435.1	167873	165023	194345
19	LR584427.1	193686	190997	221708
20	LR584429.1	204627	201538	235274
21	LR584437.1	204698	201523	232738
22	LR584440.1	176129	173448	199361
23	LR584421.1	183646	181340	207946
24	LR584412.1	188032	185557	213861
25	LR584436.1	183079	180440	209361
26	LR584439.1	159288	156745	182356
27	LR584424.1	193488	190669	219803
28	LR584422.1	184673	181970	209357
29	LR584418.1	88358	86109	106688
30	LR584432.1	164122	161694	185873
31	LR584423.1	172564	170493	197799
32	LR584408.1	95039	92601	115565
33	LR584414.1	176910	174404	200360
34	LR584434.1	163182	160908	185341
35	LR584444.1	157880	155729	179457
36	LR584442.1	151951	149560	172568
37	LR584417.1	80027	77821	96125
38	LR584425.1	96078	93468	115728
39	LR584413.1	65573	63610	79359
40	LR584443.1	71886	70099	86242
Total		**7302922**	**7184211**	**8400006**

Tab. 4.3. Unique probes identified for each chromosome, after excluding putative introgressed loci, pseudo-SNPs, SNPs with G/C or A/T alleles and inter-SNP distance less than 35 bp.

4.3.2.4. Ancestry diagnostic SNPs

Population- and species-diagnostic SNPs were identified based on criteria for 7 different categories (Fig 4.6). Of these, the most informative for inferring ancestry are category 2A (35.2 thousand marble trout species-specific SNPs) and category 6 (~11 thousand SNPs exclusive to each marble trout subpopulation for a total of 105.5 thousand SNPs). About 674.4 thousand SNPs were polymorphic both in brown trout and at least 3 marble trout populations (category 1), while 3.4 million SNPs segregated only in marble trout populations but not in brown trout.

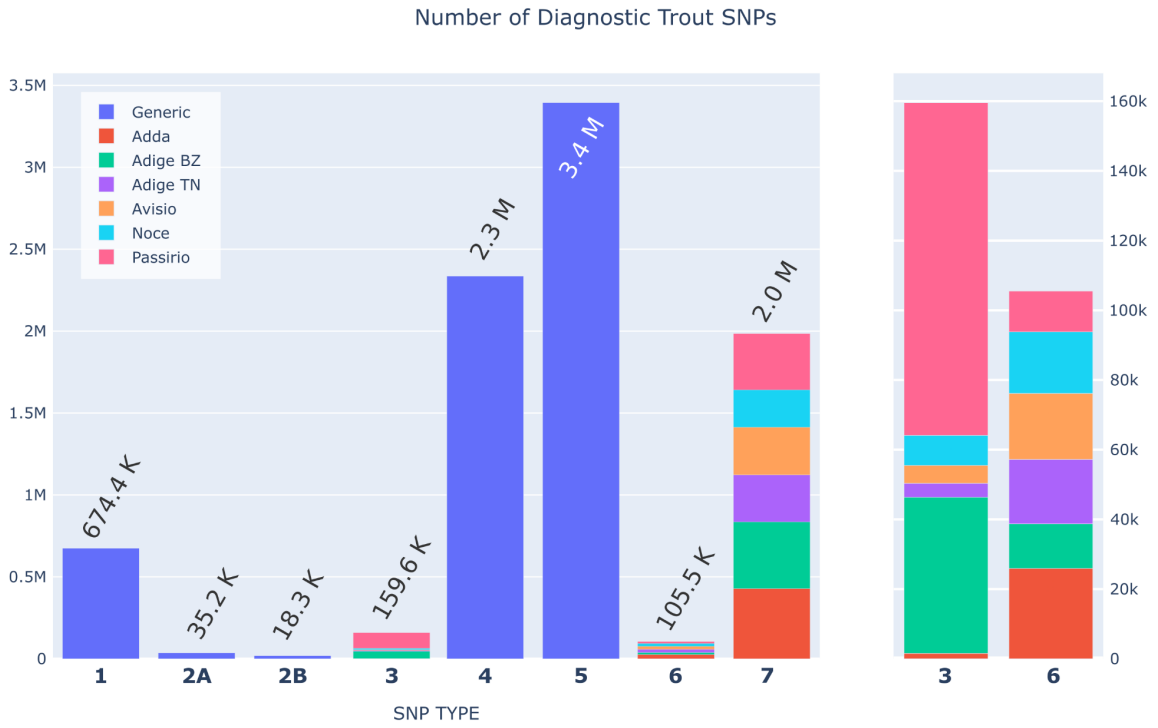


Fig 4.6. Ancestry diagnostic SNPs. Count of SNPs ascribed to different categories as described in section 4.3.2.4. In particular, subtype 2A is the most informative of marble trout as a species, while type 6 is the most suitable for identifying marble trout subpopulation.

4.4. Discussion

Although the genetic diversity of the *Salmo trutta* L. complex in the Italian region is difficult to disentangle, there is clear consensus that Italian lineages must be shielded from further genomic erosion due to introgression from stocked brown trout (Meraner et al, 2012; Gratton et al, 2014; Meraner & Gandolfi, 2017; Splendiani et al, 2019). To this end, Palombo et al. (2021) implemented the Affymetrix 57 K SNP array (Palti et al, 2014) derived from rainbow trout (*Oncorhynchus mykiss,* Walbaum 1792) to genotype two lineages of Mediterranean trout (*Salmo cettii*, Rafinesque 1810) in South-Central Italy but only ~900 SNPs were polymorphic in these populations. Likewise, Saint-Pé et al. (2019) identified a set of 12,204 SNPs informative in brown trout from Atlantic and Mediterranean lineages, ~700 of which were polymorphic in our trout populations (data not shown). Still, studies in Italian marble trout greatly rely on mtDNA and few nuclear markers for ancestry assessment and detection of hybrids (Meraner et al, 2012), emphasising the need for a standardised, high-density genotyping tool. The present study aimed to close this gap and began by discovering 19.6 million high-quality SNPs from Whole Genome Sequencing of one brown and six marble trout populations. Among these, over 4 million were polymorphic in at least 3 of the six italian samples (SNP categories 1 and 5, as described in section 4.3.2.4.), and more than 2 million SNPs were polymorphic in at least one population (categories 3 and 7). Moreover, alleles that segregate in brown trout but not in marble trout (categories 2A and 4) are useful for countering ascertainment bias when using the SNP array to genotype populations different from the ones considered in its development (Bianco et al, 2014; Geibel et al, 2021).

One of the challenges of this study was to ensure that, although individuals were previously selected on the basis of genetic screening to exclude hybrids, as far as reasonably possible, SNPs that contained introgressed alleles did not make their way into the final SNP panel. The presence of a significant number of these SNPs would confound

and weaken the power to detect individual ancestry. Filtering criteria were thus designed to detect rare alleles in marble trout which had the highest probability of deriving from admixture with domestic brown trout. Visual (Fig. 4.4) and analytical (tests of Complete Spatial Randomness, data not shown) assessment of the physical distribution along chromosomes suggested that such alleles were often present in clusters compatible with inherited brown trout haplotypic blocks. Interestingly, although trout individuals selected for this study presented a nearly 100% marble trout ancestry according to a previous survey using 15 microsatellites (published in Meraner & Gandolfi, 2018), I observed putative introgressed alleles in all populations which matched the admixture patterns found by Meraner & Gandolfi (2018). Specifically, in both studies Northern Adige (Adige BZ) and Passirio rivers showed the highest level of average introgression, followed by Noce, Avisio and Southern Adige (Adige TN) rivers and ending with Adda river as the least admixed population. The fact that my estimates are in line with previously reported admixture trends observed with mtDNA and nuclear microsatellites (Meraner et al, 2012; Splendiani et al, 2016, 2019) represents an *a posteriori* validation of the filtering method used for identifying introgressed SNP alleles.

Another obstacle was posed by residual paralogous regions originating from WGD (Allendorf & Thorgaard, 1984; Comai, 2005), which I was able to overcome by discarding 424 thousand SNPs having an approximate allelic frequency of 0.50 ± 0.05 in all populations. While not all these loci are pseudo-SNPs (most are likely true SNPs as suggested by peak B in Fig. 4.5), my conservative approach was facilitated by the high availability of genomic markers (19.6 million).

Finally, in the light of increasing genotyping success rate in a future SNP genotyping array, flanking sequences around SNPs were tested for specificity in the brown trout genome considering parameters that influence annealing capacity (see section 4.3.2.3 Technical filters), and only probes with unique matches were retained (Tab. 4.2). Further *in silico* evaluation of probes is normally carried out by the manufacturer and a *p*-convert value is assigned to each probe (Liu et al, 2017). Additionally, should the 8.4 million 71 bp probes not

suffice, I identified more than 7.2 million unique 35 bp upstream or downstream half-probes that can also be tested for inclusion in the Axiom SNP array, as per manufacturer instructions.

Thus, the present study contributes to the conservation of endemic marble trout by providing a high-density SNP panel informative of species and population ancestry in six different marble trout populations from Northern Italy. These SNPs can be further filtered using information from ancestry-informative SNP categories hereby identified and depending on the target amount of trout SNPs allocated within the Axiom SNP array.

4.5. Conclusion

- A total of 19.6 million high quality biallelic SNPs were generated from pooled WGS data in one brown trout and six marble trout populations.

- Fine-scale introgression analysis revealed a widespread effect of admixture from brown trout into all sampled marble trout populations, accounting for a total of ~36 thousand putative introgressed alleles (0.18% of the initial SNP data set).

- Passirio and AdigeBZ samples showed higher genome-wide estimates of introgression according to the filtering approach used in this study: 0.052% and 0.051%, respectively. Adda (0.013%) and AdigeTN (0.018%) showed the least amount of introgression.

- 424 thousand pseudo-SNPs from paralogous regions of the trout genome were identified with a conservative approach and removed from the SNP panel.

- SNPs which had G/C or A/T alleles or an inter-SNP distance smaller than 35 bp were discarded, compatibly with Axiom SNP Assay technology.

- A total of 8.4 million unique 71 bp-long probes centred around remaining SNPs were retained after discarding probes with non-specific genomic alignments.

- Moreover, sets of ancestry-informative SNPs both at species and population level were identified for future use in the genotyping array. SNPs which are less informative of current samples but nonetheless polymorphic in both species were also identified in order to contrast ascertainment bias when genotyping additional populations, not included in this study

Chapter 5: Final Discussion and Concluding Remarks

5.1. Introgression by non-native species

It is estimated that more than one-fifth of freshwater fish species in Italy are recognised as “Critically Endangered” (IUCN Red List 2013), and at least a one quarter-fourth are classified as either “Endangered”or “Vulnerable”. Moreover, as many as 18.3% of the 93 fish species occurring in the Italian peninsula are endemisms (Bianco et al, 2013), which are particularly at risk not only due to the increasing climatic oscillations observed in Mediterranean peninsular ecosystems (Griffiths, 2006) and to habitat degradation (Lucentini et al, 2006) but also due to restocking activities with allochthonous species (Horreo et al, 2014). In particular, this anthropogenic threat is linked to absence or inadequacy of legislation aimed at preserving local biodiversity, often caused by an underlying lack of knowledge concerning actual MUs. This was the case for marble trout and italian pike, of great value in recreational fishing, whose populations have undergone stocking with commercial lines for decades (Lucentini et al, 2011; Meraner and Gandolfi, 2017), long before the negative consequences of such actions were known. Despite strict European, national and regional bans against the propagation of alien species, release of domestic strains into the wild is still ongoing in many areas, including protected Natura 2000 sites (https://ec.europa.eu/environment/nature/natura2000/; Splendiani et al., 2019).

Long-term fitness of introgressed populations is expected to decrease due to the loss of genomic adaptations to local ecosystems (Muhfel et al, 2009). My genome-wide selection scans in Italian pike highlight potential species-specific molecular adaptations regarding olfactory perception, metabolism- and immune system-related functions, supporting the need to preserve the adaptive traits of endemic species.

Furthermore, my results confirm the persistence of non-native hybridisation both in Italian pike and in wild marble trout populations, albeit through different approaches given the different experimental designs. I identified traces of exotic admixture in the former through Bayesian inference of demography, while in the latter I filtered SNPs based on criteria concerning allelic frequency. Although individuals from both species had been selected for this project on the basis of previous microsatellite screening, my analyses revealed hybridisation in putatively "pure" populations that was not detected by traditional genetic markers, which I further discuss in Section 5.3.

The ability to detect hybrids is a crucial topic in conservation genetics studies across many species (Stronen et al, 2022; Mattucci et al, 2019; Halbert and Derr, 2006; Meraner, Unfer and Gandolfi, 2013). A key question in this regard, for practical and applied purposes, is where to trace the line between hybrids and pure individuals. While there is no universal rule, some studies adopt the 90% ancestry membership threshold (Barilani et al., 2006; van Wyk et al., 2017; Gandolfi et al., 2017) which has been determined as the limit of detection (LOD) through analysis of empirical and simulated data (Barilani et al, 2006), but the true LOD may vary depending on factors such as marker set size, type and the genetic similarity between parental species (Vähä and Primmer, 2006). Moreover, the criterion for defining hybrids may vary depending on the aim of the survey, as different analyses may have a different sensitivity to introgressed alleles. For example, phylogenetic or demographic analyses might be compromised to a greater extent by undetected introgression (Holder, Anderson and Holloway, 2001; Leaché and McGuire, 2006) than genome-wide selection or association scans, where it could, at worst, dilute true selection signals (Medina-Gomez et al., 2015). For this reason, I implemented two different hybrid detection thresholds in Italian pike: 5% and 10% in Chapters 2 and 3, respectively. Most importantly, conservation programmes should place special emphasis on which threshold they use when defining and removing hybrids from wild populations, as a lower tolerance of admixture may result in the

exclusion of allochthonous alleles at the expense of endemic genetic diversity, by eliminating rare native alleles.

5.2. On man-mediated versus natural gene flow

In the present study, I contribute to the genetic conservation of two Italian alpine endemic fish, the Italian pike and marble trout, both directly, by determining levels of introgression from a non-native congeneric and natural genetic structuring within the former, and indirectly by discovering high-density genomic markers to be used in future screening and monitoring practices for both species.

Understanding population substructure is essential for defining management units, which can be considered a reservoir of genomic adaptations to specific biogeographic areas and therefore to particular ecosystemic conditions. Supportive breeding programmes for these two important teleosts do not as yet prioritise distinctions between local lineages, partly because uncertainty still pervades the characterisation of pike and trout MUs. Indeed, my demographic analyses reveal admixture between Italian pike populations that should be separated by the Apennines, an insurmountable obstacle for most aquatic species. This was the case for the Po River population showing introgression from Trasimeno Lake pike, which can only be attributed to man-mediated translocations and should be discouraged. I also detected other instances of within-species admixture in native pike, which, on the other hand, can be examples of natural migration events due to spatial continuity as is the case between Garda Lake, Adda River and Po River, given that they all belong to the Po basin.

Similar dynamics have been documented for other endemic freshwater species such as grayling (Meraner, Cornetti and Gandolfi, 2014), barbel (Berrebi et al, 2014) and the Italian “vairone” *Telestes muticellus*, Bonaparte 1837 (Stefani et al, 2004; Marchetto et al., 2010), which show evidence of secondary contact between populations that are now geographically separated. This can be understood in the light of climatic oscillations over the past 2 million years (Hewitt, 1999): for freshwater fish, glacial maxima often provided

ecological corridors instead of obstacles to gene flow. In fact, during glaciations in the Alpine region throughout the Quaternary period, the sea level periodically dropped below 120 meters causing the North Adriatic shoreline to recede by hundreds of kilometres (Maselli et al, 2011). As a consequence, the Po basin expanded considerably (Fig. 5.1) and many catchments reached confluence before draining into the Adriatic Sea, generating an ecological corridor for freshwater species to colonise (or re-colonise) the Italian peninsula from Eastern Europe (Giuffra et al, 199; Kotlik and Berrebi, 2001; Bernatchez, 2001, Gouskov and Vorburger, 2016; Meraner and Gandolfi, 2017). Contrary to terrestrial species, deglaciation then caused inland ichthyofauna to evolve separately due to vicariance, leading to the high endemic biodiversity observed in Italy today.

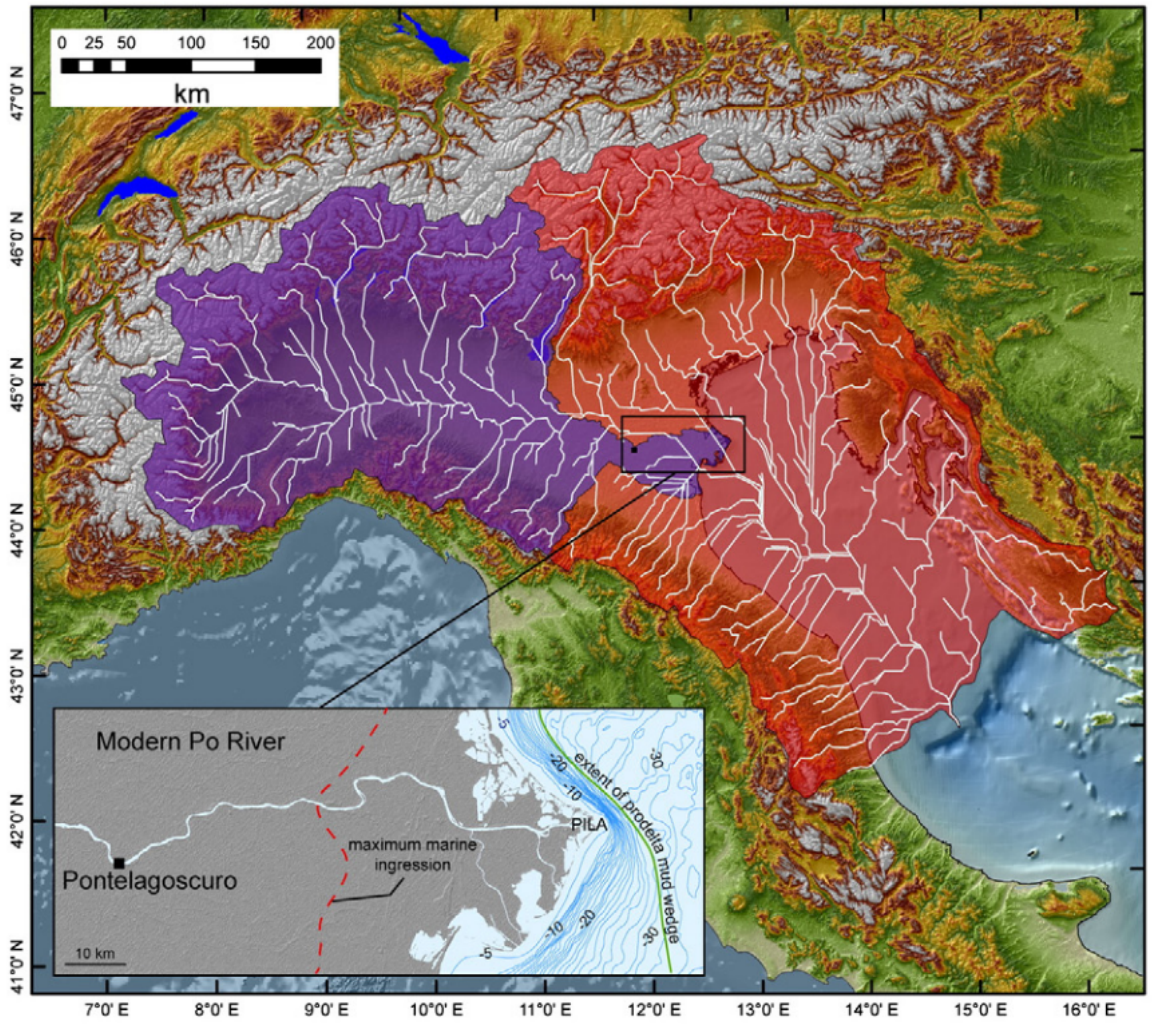


Fig. 5.1. Extent of the Po basin in the present day (shown in purple) and during maximum marine ingression ~5.5 thousand years ago (shown in red). Figure taken from Maselli et al., 2011.

5.3. Microsatellites versus single nucleotide polymorphisms

Genetic monitoring has been paramount in the investigation and preservation of biodiversity, especially through the use of traditional genetic markers such as microsatellites (Hodel et al, 2016). However, they may provide limited resolving power to discern between certain competing biological, ecological or evolutionary hypotheses as well as to detect admixed individuals in introgressed populations (Haasl and Payseur, 2011). In fact, in Chapter 2 I compared the performance of both marker types in detecting introgression from non-native European pike (qEU) and found that in all but one case microsatellites considerably underestimated hybridisation. One possible explanation for this lies in the different software used for demographic inference: Gandolfi and colleagues (2017) analysed their microsatellite data with STRUCTURE (Pritchard et al., 2000) while I used fastSTRUCTURE (Raj, Stephens and Pritchard, 2014) for SNPs. These two software are closely related but differ in their ability to handle different marker set sizes and types.

Indeed, they take into consideration the different mutation dynamics underlying SNPs and microsatellites. The former arise due to point mutations and follow the infinite-site model (ISM) (Kimura, 1969), according to which it's highly unlikely for a mutation to occur twice at the same site. On the other hand, microsatellites have a greater mutation rate (Li et al, 2002) and occur due to DNA strand slippage during replication, which generates variability in the number of tandem repeats (Bhargava and Fuentes, 2009), and are better described by other models such as the stepwise mutation model (SMM) (Slatkin, 1995). The main difference between the ISM and SMM is that allele homoplasy is allowed within the second; in other words, a second mutation can revert a microsatellite allele back to its ancestral state. Homoplastic alleles are an example of convergent evolution, being identical by state (number of tandem repeats) but not identical by descent (Ohta and Kimura, 1973).

A practical implication of this phenomenon is that microsatellite markers may not be as reliable as SNPs for phylogeographic studies where divergence occurred earlier than 3 – 30 thousand years ago (Paetkau et al, 1997; Beaumont et al, 1999). In fact, Berrebi, Jesenšek and Crivelli (2016) found that microsatellites cannot distinguish ancient (and thus natural) admixture between marble trout and brown trout, likely due to their high mutation rate. Vice versa, microsatellites may, in some cases, provide greater insight into fine population structure and recent gene flow given their greater mutation rate and per-locus informativity (Angers and Bernatchez, 1998). Thus, another possible interpretation of the dissonance between SNPs and microsatellites could be linked to incomplete lineage sorting, ILS, (Komarova and Lavrenchenko, 2022), because different markers can give rise to different gene trees and confounding results (Kutschera et al., 2014). Indeed, gene flow and ILS can lead to the same type of signal and this adds a layer of complexity to the inference of introgression, which can be contrasted by a larger number of genomic SNPs as opposed to fewer microsatellites (Komarova and Lavrenchenko, 2022).

5.4. Choice of experimental design

In this doctoral thesis, I implemented two different whole genome sequencing approaches: individual-based for Italian pike and population-based for marble trout. The workflows implied in both methods greatly overlap, but also consist of specific advantages and disadvantages. This experimental design reflects similarities and differences between the two case studies.

Firstly, no genomic studies existed in literature for Italian pike, being very recently identified as a distinct endemic species, so there was an interest to gain high resolution insight into the genomic landscape of this species. Because of this, an individual sequencing approach was chosen with an average sequencing depth of 20X to yield reliable genotypes and limit the amount of markers that do not exceed quality assessments. This approach allowed for haplotype phasing which was used in studies of selection to determine genomic

regions that fitted the patterns expected under selection which may provide insights into genomic adaptations. Moreover, because current supportive breeding programs for pike are based predominantly on morphological traits as opposed to genetic ones (Lucentini et al, 2009), it was important to assess the extent of hybridisation through a high-resolution, per-individual genomic approach.

However, the higher sequencing costs initially posed a constraint on the sample size which in turn limited the statistical power of certain analyses. In fact, the original data set consisted of only 28 pikes, 16 of which had been previously identified as *Esox flaviae* through microsatellites. As further funding became available, it was possible to increase the pike data set to 61 individuals and reach significance thresholds for outlier detection in GWSS analyses. *Per contra*, upscaling the data set required more time for data generation, read alignment, SNP discovery and re-analysis. Thus, it is wise to consider aimed resolving power when defining the optimal sample size in function of downstream analyses, sequencing depth and project costs.

In marble trout, given the necessity to disentangle the complexity of the *Salmo* genus in Italy (Meraner and Gandolfi, 2017) as well as to improve the resolution and automation of current screening methods used in supportive breeding, the primary aim was to produce a cost-effective large-scale genotyping tool, namely a SNP panel for future use in a genotyping array. Unlike for pike, individuals were sequenced at low coverage (3X) and reads were consequently pooled together by sampling location. The experimental design in this case favoured the exploration of allelic frequencies in a higher number of populations as opposed to the investigation of the genomic landscape of individuals. In population genomics, this approach is common because of the advantageous ratio between informativity and sequencing costs (Ferretti, Ramos-Osins and Perez-Enciso, 2013).

5.5. Implications of the study and directions for future research

Amidst the growing concern of anthropogenic threats to biodiversity, technological advances have brought new hope to the field of conservation biology. In this study, I generated millions of informative SNPs from whole genome sequencing data to be used in genetic monitoring and supportive breeding practices for two key alpine species, marble trout and Italian pike. My findings advance the understanding of genetic clusters within Italian pike, which should be taken into consideration by national and regional wildlife management protocols. Moreover, I provide novel insight into the genomic landscape of Italian pike and potential species-specific molecular adaptations, emphasising the need to preserve the adaptive potential of endemic species. Future investigation of the unique adaptations within local genetic clusters will be beneficial for the acknowledgement and protection of management units. In marble trout, I filtered and validated *in silico* a set of more than 8 million high-quality SNPs for inclusion in a genotyping array aimed at assisting current supportive breeding procedures. Switching from current microsatellite-based screening to a SNP genotyping array would not only increase resolving power for hybrid detection but also provide faster results, translating in less time in captivity, lower stress levels and higher post-release survival rates in marble trout (Lucarda et al, 2007). Overall, the conservation of both species will benefit from this study, though further monitoring efforts and phylogenetic research are needed to ensure the long-term survival of these endemisms.

References

Ahrens, C. W., Rymer, P. D., Stow, A., Bragg, J., Dillon, S., Umbers, K. D. L., & Dudaniec, R. Y. (2018). The search for loci under selection: Trends, biases and progress. *Molecular Ecology*, *27*(6), 1342–1356. https://doi.org/10.1111/mec.14549

Akey, J. M., Zhang, G., Zhang, K., Jin, L., & Shriver, M. D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, *12*(12), 1805–1814. https://doi.org/10.1101/gr.631202

Allendorf, F. W. (1986). Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, *5*(2), 181–190. https://doi.org/10.1002/zoo.1430050212

Allendorf, F. W., & Thorgaard, G. H. (1984). Tetraploidy and the evolution of salmonid fishes. In *Evolutionary Genetics of Fishes* (pp. 1–53). Springer US. http://dx.doi.org/10.1007/978-1-4684-4652-4_1

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., & Lander, E. S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, *407*(6803), 513–516. https://doi.org/10.1038/35035083

Anderson, E. C., & Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, *172*(4), 2567–2582. https://doi.org/10.1534/genetics.105.048074

Andrews, S. (2010). *Babraham Bioinformatics*. FastQC A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Angers, B., & Bernatchez, L. (1998). Combined Use of SMM and Non-SMM Methods to Infer Fine Structure and Evolutionary History of Closely Related Brook Charr (Salvelinus fontinalis, Salmonidae) Populations from Microsatellites. *Molecular Biology and Evolution*, *15*(2), 143–159. https://doi.org/10.1093/oxfordjournals.molbev.a025911

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1). https://doi.org/10.1002/0471250953.bi1110s43

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Barilani, M., Sfougaris, A., Giannakopoulos, A., Mucci, N., Tabarroni, C., & Randi, E. (2006). Detecting introgressive hybridisation in rock partridge populations (Alectoris graeca) in Greece through Bayesian admixture analyses of multilocus genotypes. *Conservation Genetics*, *8*(2), 343–354. https://doi.org/10.1007/s10592-006-9174-1

Beaumont, M., Bruford, M., Goldstein, D., & Schlötterer, C. (1999). *Microsatellites: evolution and applications*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bernatchez, L. (2001). The evolutionary history of brown trout (Salmo trutta L.) Inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial dna variation. *Evolution*, *55*(2), 351–379. https://doi.org/10.1111/j.0014-3820.2001.tb01300.x

Berrebi, Jesenšek, & Crivelli. (2016). Natural and domestic introgressions in the marble trout population of Soča River (Slovenia). *Hydrobiologia*, *785*(1), 277–291. https://doi.org/10.1007/s10750-016-2932-2

Berrebi, P., Chenuil, A., Kotlík, P., Machordom, A., Tsigenopoulos, C. S., Alves, M., Cartaxana, A., Correia, A., & Lopes, L. (2014). Disentangling the evolutionary history of the genus Barbus sensu lato, a twenty years adventure. *Professor Carlos Almaça (1934–2010)–Estado Da Arte Em Áreas Científicas Do Seu Interesse*, 29–55.

Berry, A. J., Ajioka, J. W., & Kreitman, M. (1991a). Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics*, *129*(4), 1111–1117. https://doi.org/10.1093/genetics/129.4.1111

Berry, A. J., Ajioka, J. W., & Kreitman, M. (1991b). Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics*, *129*(4), 1111–1117. https://doi.org/10.1093/genetics/129.4.1111

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., Aury, J.-M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G. H., … Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, *5*(1). https://doi.org/10.1038/ncomms4657

Bertolini, F., Geraci, C., Schiavo, G., Sardina, M. T., Chiofalo, V., & Fontanesi, L. (2016). Whole genome semiconductor based sequencing of farmed European sea bass (Dicentrarchus labrax) Mediterranean genetic stocks using a DNA pooling approach. *Marine Genomics*, *28*, 63–70. https://doi.org/10.1016/j.margen.2016.03.007

Bhargava, & Fuentes. (2009). Mutational dynamics of microsatellites. *Molecular Biotechnology*, *44*(3), 250–266. https://doi.org/10.1007/s12033-009-9230-4

Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., Poncet, C., Micheletti, D., Kerschbamer, E., Di Pierro, E. A., Larger, S., Pindo, M., Van de Weg, E., Davassi, A., Laurens, F., Velasco, R., Durel, C.-E., & Troggio, M. (2016). Development and validation of the Axiom®Apple480K SNP genotyping array. *The Plant Journal*, *86*(1), 62–74. https://doi.org/10.1111/tpj.13145

Bianco, P. G. (1995). Mediterranean endemic freshwater fishes of Italy. *Biological Conservation*, *72*(2), 159–170. https://doi.org/10.1016/0006-3207(94)00078-5

Bianco, P. G. (2013). An update on the status of native and exotic freshwater fishes of Italy. *Journal of Applied Ichthyology*, *30*(1), 62–77. https://doi.org/10.1111/jai.12291

Bianco, P. G., Caputo, V., Ferrito, V., Lorenzoni, M., Nonnis Marzano, F., Stefani, F., Sabatini, A., Tancioni, L., Rondinini, C., Battistoni, A., & Peronace, V. (2013). *Lista Rossa IUCN dei Vertebrati Italiani*.

Bianco, P. G., & Delmastro, G. B. (2011). *Recenti novità tassonomiche riguardanti i pesci d'acqua dolce autoctoni in Italia e descrizione di una nuova specie di luccio*. IGF publishing.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*, 314–331.

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3 Genes|Genomes|Genetics*, *5*(5), 931–941. https://doi.org/10.1534/g3.114.015784

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, *140*(2), 783–796. https://doi.org/10.1093/genetics/140.2.783

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, *108*(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Candy, J. R., Campbell, N. R., Grinnell, M. H., Beacham, T. D., Larson, W. A., & Narum, S. R. (2015). Population differentiation determined from putative neutral and divergent adaptive genetic markers in Eulachon (Thaleichthys pacificus, Osmeridae), an anadromous Pacific smelt. *Molecular Ecology Resources*, *15*(6), 1421–1434. https://doi.org/10.1111/1755-0998.12400

Caputo, V., Giovannotti, M., Nisi Cerioni, P., Caniglia, M. L., & Splendiani, A. (2004). Genetic diversity of brown trout in central Italy. *Journal of Fish Biology*, *65*(2), 403–418. https://doi.org/10.1111/j.0022-1112.2004.00458.x

Caputo, V., Giovannotti, M., & Splendiani, A. (2010). Pattern of gonad maturation in a highly stocked population of brown trout (Salmo truttaL., 1758) from Central Italy. *Italian Journal of Zoology*, *77*(1), 14–22. https://doi.org/10.1080/11250000802589576

Casselman, J. M., & Lewis, C. A. (1996). Habitat requirements of northern pike (Essox lucius). *Canadian Journal of Fisheries and Aquatic Sciences*, *53*(S1), 161–174. https://doi.org/10.1139/f96-019

Cavalli-Sforza, L. (1966). Population structure and human evolution. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *164*(995), 362–379. https://doi.org/10.1098/rspb.1966.0038

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). https://doi.org/10.1186/s13742-015-0047-8

Charlesworth, B., & Charlesworth, D. (2018). Neutral variation in the context of selection. *Molecular Biology and Evolution*, *35*(6), 1359–1361. https://doi.org/10.1093/molbev/msy062

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*(4), 1289–1303. https://doi.org/10.1093/genetics/134.4.1289

Charlesworth, B., Nordborg, M., & Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research*, *70*(2), 155–174. https://doi.org/10.1017/s0016672397002954

Chown, S., Hodgins, K., & Griffin, P. (2016). Biological Invasions, Climate Change, and Genomics. *Crop Breeding*, 59–114. https://doi.org/10.1201/9781315365084-12

Christensen, K. A., Brunelli, J. P., Lambert, M. J., DeKoning, J., Phillips, R. B., & Thorgaard, G. H. (2013). Identification of single nucleotide polymorphisms from the transcriptome of an organism with a whole genome duplication. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-325

Christensen, M., Sunde, L., Bolund, L., & Ørntoft, T. F. (1999). Comparison of three methods of microsatellite detection. *Scandinavian Journal of Clinical and Laboratory Investigation*, *59*(3), 167–177. https://doi.org/10.1080/00365519950185698

Ciabatti, M. (1968). *Gli antichi delta del Po anteriori al 1,600 [The ancient Po Delta before AD 1,600].* (pp. 23–33). Atti del convegno internazionale di studi sulle antichità di Classe-Ravenna [Proceedings of the International Meeting on the history of Classe-Ravenna].

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics*, *6*(11), 836–846. https://doi.org/10.1038/nrg1711

Craig, J. F. (1996). *Pike - biology and exploitation. Fish and Fisheries Series*. Chapman & Hall, London, UK.

Craig, J. F. (2008). A short review of pike ecology. *Hydrobiologia*, *601*(1), 5–16. https://doi.org/10.1007/s10750-007-9262-3

Crivelli, A. J. (2006). Salmo marmoratus. *IUCN Red List of Threatened Species*. https://doi.org/10.2305/iucn.uk.2006.rlts.t19859a9043279.en

Crivelli, A. J., Poizat, G., Berrebi, P., Jesensek, D., & Rubin, J. F. (2000). Conservation biology applied to fish: The example of a project for rehabilitating the marble trout (Salmo marmoratus) in Slovenia. *Cybium : Revue Internationale d'Ichtyologie*, *24*(3). https://doi.org/https://hal.archives-ouvertes.fr/halsde-00332632

Csilléry, K., Rodríguez-Verdugo, A., Rellstab, C., & Guillaume, F. (2018). Detecting the genomic signal of polygenic adaptation and the role of epistasis in evolution. *Molecular Ecology*, *27*(3), 606–612. https://doi.org/10.1111/mec.14499

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. https://doi.org/10.1038/nmeth.2307

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13225-y

Divino, J. N., Monette, M. Y., McCormick, S. D., Yancey, P. H., Flannery, K. G., Bell, M. A., Rollins, J. L., Von Hippel, F. A., & Schultz, E. T. (2016). Osmoregulatory physiology and rapid evolution of salinity tolerance in threespine stickleback recently introduced to fresh water. *Evolutionary Ecology Research*, *17*(2), 179–201.

Eckert, A. J., Wegrzyn, J. L., Pande, B., Jermstad, K. D., Lee, J. M., Liechty, J. D., Tearse, B. R., Krutovsky, K. V., & Neale, D. B. (2009). Multilocus Patterns of Nucleotide Diversity and Divergence Reveal Positive Selection at Candidate Genes Related to Cold Hardiness in Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii). *Genetics*, *183*(1), 289–298. https://doi.org/10.1534/genetics.109.103895

Eisendle, D., Wieser, J., Meraner, A., & Gandolfi, A. (2019). Supportive breeding program of Marble trout (Salmo marmoratus) in the Province of Bolzano-Italy. *Advances in the Population Ecology of Stream Salmonids V*, 40–41. http://hdl.handle.net/10449/55438

Ellegren, H., & Sheldon, B. C. (2008). Genetic basis of fitness differences in natural populations. *Nature 2008 452:7184*, *452*(7184), 169–175. https://doi.org/10.1038/nature06737

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Eschbach, E., Nolte, A. W., Kohlmann, K., Alós, J., Schöning, S., & Arlinghaus, R. (2021). Genetic population structure of a top predatory fish (northern pike, Esox lucius) covaries with anthropogenic alteration of freshwater ecosystems. *Freshwater Biology*, *66*(5), 884–901. https://doi.org/10.1111/fwb.13684

Fabbri, E., Miquel, C., Lucchini, V., Santini, A., Caniglia, R., Duchamp, C., Weber, J.-M., Lequette, B., Marucco, F., Boitani, L., Fumagalli, L., Taberlet, P., & Randi, E. (2007). From the Apennines to the Alps: Colonization genetics of the naturally expanding Italian wolf (Canis lupus) population. Molecular Ecology, 16(8), 1661–1671. https://doi.org/10.1111/j.1365-294x.2007.03262.x

Fabbri, E., Velli, E., D'Amico, F., Galaverni, M., Mastrogiuseppe, L., Mattucci, F., & Caniglia, R. (2018). From predation to management: Monitoring wolf distribution and understanding depredation patterns from attacks on livestock. Hystrix, the Italian Journal of Mammalogy, 29(1), 101–110. https://doi.org/10.4404/hystrix-00070-2018

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–1413. https://doi.org/10.1093/genetics/155.3.1405

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291. https://doi.org/10.1093/molbev/msu077

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. https://doi.org/10.1111/mec.12522

Fitak, R. R. (2021). OptM: Estimating the optimal number of migration edges on population trees using Treemix. *Biology Methods and Protocols*, *6*(1). https://doi.org/10.1093/biomethods/bpab017

Ford, A. G. P., Dasmahapatra, K. K., Rüber, L., Gharbi, K., Cezard, T., & Day, J. J. (2015). High levels of interspecific gene flow in an endemic cichlid fish adaptive radiation from an extreme lake environment. *Molecular Ecology*, *24*(13), 3421–3440. https://doi.org/10.1111/mec.13247

Frankham, R., Ballou, J. D., Briscoe, D. A., & Ballou, J. D. (2002). *Introduction to conservation genetics*. Cambridge university press.

Fraser, D. J. (2008). How well can captive breeding programs conserve biodiversity? A review of salmonids. *Evolutionary Applications*, *1*(4), 535–586. https://doi.org/10.1111/j.1752-4571.2008.00036.x

Fraser, D. J. (2017). Genetic diversity of small populations: Not always "doom and gloom"? *Molecular Ecology*, *26*(23), 6499–6501. https://doi.org/10.1111/mec.14371

Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, *147*(2), 915–925. https://doi.org/10.1093/genetics/147.2.915

Fumagalli, L., Snoj, A., Jesenšek, D., Balloux, F., Jug, T., Duron, O., Brossier, F., Crivelli, A. J., & Berrebi, P. (2002). Extreme genetic differentiation among the remnant populations of marble trout (Salmo marmoratus) in Slovenia. *Molecular Ecology*, *11*(12), 2711–2716. https://doi.org/10.1046/j.1365-294x.2002.01648.x

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, *7*(11), e1002355. https://doi.org/10.1371/journal.pgen.1002355

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, *27*(9), 489–496. https://doi.org/10.1016/j.tree.2012.05.012

Gandolfi, A., Eisendle, D., Wieser, J., Girardi, M., Casari, S., Crestanello, B., & Meraner, A. (2019). Do it right or don't do it at all! Genetic screening of S. marmoratus exemplifies the need to revise many or most salmonid conservation and restocking programmes. *Advances in the Population Ecology of Stream Salmonids V*, 42. http://hdl.handle.net/10449/54388

Gandolfi, A., Ferrari, C., Crestanello, B., Girardi, M., Lucentini, L., & Meraner, A. (2017). Population genetics of pike, genus Esox (Actinopterygii, Esocidae), in Northern Italy: Evidence for mosaic distribution of native, exotic and introgressed populations. *Hydrobiologia*, *794*(1), 73–92. https://doi.org/10.1007/s10750-016-3083-1

Gandolfi, A., Fontaneto, D., Natali, M., & Lucentini, L. (2015). Mitochondrial genome of Esox flaviae (Southern pike): Announcement and comparison with other Esocidae. *Mitochondrial DNA Part A*, *27*(4), 3037–3038. https://doi.org/10.3109/19401736.2015.1063123

Gautier, M., Klassmann, A., & Vitalis, R. (2016). rehh2.0: A reimplementation of the R packagerehhto detect positive selection from haplotype structure. *Molecular Ecology Resources*, *17*(1), 78–90. https://doi.org/10.1111/1755-0998.12634

Gautier, M., & Vitalis, R. (2012). rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, *28*(8), 1176–1177. https://doi.org/10.1093/bioinformatics/bts115

Geibel, J., Reimer, C., Weigend, S., Weigend, A., Pook, T., & Simianer, H. (2021). How array design creates SNP ascertainment bias. *PLOS ONE*, *16*(3), e0245178. https://doi.org/10.1371/journal.pone.0245178

Genovesi, P., Angelini, P., Bianchi, E., Dupré, E., Ercole, S., Giacanelli, V., Ronchi, F., & Stoch, F. (2014). *Specie e habitat di interesse comunitario in Italia: distribuzione, stato di conservazionee trend. Odonati (Riservato E., Fabbri R., Festi A., Grieco C., Hardersen S., Landi F.)* (Vol. 194). ISPRA. http://www.isprambiente.gov.it/it/pubblicazioni/rapporti/specie-e-habitat-di-interesse-comunitario-in-italia-distribuzione-stato-di-conservazione-e-trend

Giuffra, E., Guyomard, R., & Forneris, G. (1996). Phylogenetic relationships and introgression patterns between incipient parapatric species of Italian brown trout (Salmo truttaL. complex). *Molecular Ecology*, *5*(2), 207–220. https://doi.org/10.1046/j.1365-294x.1996.00074.x

Gouskov, & Vorburger. (2016). Postglacial recolonizations, watershed crossings and human translocations shape the distribution of chub lineages around the Swiss Alps. *BMC Evolutionary Biology*, *16*(1), 1–13. https://doi.org/10.1186/s12862-016-0750-9

Gratton, P., Allegrucci, G., Sbordoni, V., & Gandolfi, A. (2014). The evolutionary jigsaw puzzle of the surviving trout (Salmo trutta L. complex) diversity in the Italian region. A multilocus Bayesian

approach. *Molecular Phylogenetics and Evolution*, *79*, 292–304. https://doi.org/10.1016/j.ympev.2014.06.022

Griffiths, D. (2006). Pattern and process in the ecological biogeography of European freshwater fish. *Journal of Animal Ecology*, *75*(3), 734–751. https://doi.org/10.1111/j.1365-2656.2006.01094.x

Haasl, & Payseur. (2010). Multi-locus inference of population structure: A comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171. https://doi.org/10.1038/hdy.2010.21

Halbert, N. D., & Derr, J. N. (2006). A comprehensive evaluation of cattle introgression into US federal bison herds. *Journal of Heredity*, *98*(1), 1–12. https://doi.org/10.1093/jhered/esl051

Hamblin, M. T., Mitchell, S. E., White, G. M., Gallego, J., Kukatla, R., Wing, R. A., Paterson, A. H., & Kresovich, S. (2004). Comparative Population Genetics of the Panicoid Grasses: Sequence Polymorphism, Linkage Disequilibrium and Selection in a Diverse Sample of Sorghum bicolor. *Genetics*, *167*(1), 471–483. https://doi.org/10.1534/genetics.167.1.471

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harris, H. (1966). C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *164*(995), 298–310. https://doi.org/10.1098/rspb.1966.0032

Harrisson, K. A., Pavlova, A., Telonis‐Scott, M., & Sunnucks, P. (2014). Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, *7*(9), 1008–1025. https://doi.org/10.1111/eva.12149

Hartl, D. L., & Clark, A. G. (2007). *Principles of population genetics. 4th edition*. Sinauer Associates.

Hermisson, J., & Pennings, P. S. (2005). Soft sweeps. *Genetics*, *169*(4), 2335–2352. https://doi.org/10.1534/genetics.104.036947

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, *68*(1–2), 87–112. https://doi.org/10.1111/j.1095-8312.1999.tb01160.x

Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, *33*(1), 54–78. https://doi.org/10.1016/0040-5809(88)90004-4

Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., Gitzendanner, M. A., Douglas, N. A., Germain-Aubrey, C. C., Chen, S., Soltis, D. E., & Soltis, P. S. (2016). The Report of My Death was an Exaggeration: A Review for Researchers Using Microsatellites in the 21st Century. *Applications in Plant Sciences*, *4*(6), 1600025. https://doi.org/10.3732/apps.1600025

Hoelzel, A. R., Bruford, M. W., & Fleischer, R. C. (2019). Conservation of adaptive potential and functional diversity. *Conservation Genetics*, *20*(1), 1–5. https://doi.org/10.1007/s10592-019-01151-x

Hohenlohe, P., Phillips, P., & Cresko, W. (2010). Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. *International Journal of Plant Sciences*, *171*(9), 1059–1071. https://doi.org/10.1086/656306

Holder, M. T., Anderson, J. A., & Holloway, A. K. (2001). Difficulties in detecting hybridization. *Systematic Biology*, *50*(6), 978–982. https://doi.org/10.1080/106351501753462911

Holderegger, R., Kamm, U., & Gugerli, F. (2006). Adaptive vs. neutral genetic diversity: Implications for landscape genetics. *Landscape Ecology*, *21*(6), 797–807. https://doi.org/10.1007/s10980-005-5245-9

Horreo, J. L., Machado-Schiaffino, G., Griffiths, A. M., Bright, D., Stevens, J. R., & Garcia-Vazquez, E. (2014). Long-term effects of stock transfers: Synergistic introgression of allochthonous genomes in salmonids. *Journal of Fish Biology*, *85*(2), 292–306. https://doi.org/10.1111/jfb.12424

Houston, R. D., Taggart, J. B., Cézard, T., Bekaert, M., Lowe, N. R., Downing, A., Talbot, R., Bishop, S. C., Archibald, A. L., Bron, J. E., Penman, D. J., Davassi, A., Brew, F., Tinch, A. E., Gharbi, K., & Hamilton, A. (2014). Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). *BMC Genomics*, *15*(1), 90. https://doi.org/10.1186/1471-2164-15-90

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Hunter, M. E., Hoban, S. M., Bruford, M. W., Segelbacher, G., & Bernatchez, L. (2018). Next-generation conservation genetics and biodiversity monitoring. *Evolutionary Applications*, *11*(7), 1029–1034. https://doi.org/10.1111/eva.12661

Jepsen, N., Beck, S., Skov, C., & Koed, A. (2001). Behavior of pike (Esox lucius L.) >50 cm in a turbid reservoir and in a clearwater lake. *Ecology of Freshwater Fish*, *10*(1), 26–34. https://doi.org/10.1034/j.1600-0633.2001.100104.x

Jónás, D., Ducrocq, V., & Croiseau, P. (2017). Short communication: The combined use of linkage disequilibrium–based haploblocks and allele frequency–based haplotype selection methods enhances genomic evaluation accuracy in dairy cattle. *Journal of Dairy Science*, *100*(4), 2905–2908. https://doi.org/10.3168/jds.2016-11798

Kayser, M. (2003). A genome scan to detect candidate regions influenced by local natural selection in human populations. *Molecular Biology and Evolution*, *20*(6), 893–900. https://doi.org/10.1093/molbev/msg092

Kerstens, H. H., Crooijmans, R. P., Veenendaal, A., Dibbits, B. W., Chin-A-Woeng, T. F., den Dunnen, J. T., & Groenen, M. A. (2009). Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: Applied to turkey. *BMC Genomics*, *10*(1). https://doi.org/10.1186/1471-2164-10-479

Kim, Y., & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, *167*(3), 1513–1524. https://doi.org/10.1534/genetics.103.025387

Kimura, M. (1955). Random genetic drift in multi-allelic locus. *Evolution*, *9*(4), 419. https://doi.org/10.2307/2405476

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, *61*(4), 893–903. https://doi.org/10.1093/genetics/61.4.893

Kirch, M., Romundset, A., Gilbert, M. T. P., Jones, F. C., & Foote, A. D. (2021). Ancient and modern stickleback genomes reveal the demographic constraints on adaptation. *Current Biology*, *31*(9), 2027-2036.e8. https://doi.org/10.1016/j.cub.2021.02.027

Klein, S., & Aylesworth, T. (1983). *The encyclopedia of North American wildlife.* Facts on File.

Kleinman-Ruiz, D., Martínez-Cruz, B., Soriano, L., Lucena-Perez, M., Cruz, F., Villanueva, B., Fernández, J., & Godoy, J. A. (2017). Novel efficient genome-wide SNP panels for the conservation of the highly endangered Iberian lynx. BMC Genomics, 18(1). https://doi.org/10.1186/s12864-017-3946-5

Kleinman-Ruiz, D., Soriano, L., Casas-Marce, M., Szychta, C., Sánchez, I., Fernández, J., & Godoy, J. A. (2019). Genetic evaluation of the Iberian lynx ex situ conservation programme. Heredity, 123(5), 647–661. https://doi.org/10.1038/s41437-019-0217-z

Klemetsen, A., Amundsen, P.-A., Dempson, J. B., Jonsson, B., Jonsson, N., O'Connell, M. F., & Mortensen, E. (2003). Atlantic salmon Salmo salar L., brown trout Salmo trutta L. and Arctic charr Salvelinus alpinus (L.): A review of aspects of their life histories. *Ecology of Freshwater Fish*, *12*(1), 1–59. https://doi.org/10.1034/j.1600-0633.2003.00010.x

Koelewijn, H. P., Pérez-Haro, M., Jansman, H. A. H., Boerwinkel, M. C., Bovenschen, J., Lammertsma, D. R., Niewold, F. J. J., & Kuiters, A. T. (2010). The reintroduction of the Eurasian otter (*Lutra lutra*) into the Netherlands: Hidden life revealed by noninvasive genetic monitoring. Conservation Genetics, 11(2), 601–614. https://doi.org/10.1007/s10592-010-0051-6

Komarova, & Lavrenchenko. (2022). Approaches to the detection of hybridization events and genetic introgression upon phylogenetic incongruence. *Biology Bulletin Reviews*, *12*(3), 240–253. https://doi.org/10.1134/S2079086422030045

Kotlik, P., & Berrebi, P. (2001). Phylogeography of the barbel (Barbus barbus) assessed by mitochondrial DNA variation. *Molecular Ecology*, *10*(9), 2177–2185. https://doi.org/10.1046/j.0962-1083.2001.01344.x

Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., Pirani, A., Brew, F., Kaiser, P., Hocking, P. M., Fife, M., Salmon, N., Fulton, J., Strom, T. M., Haberer, G., Weigend, S., Preisinger, R., Gholami, M., … Burt, D. W. (2013). Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics*, *14*(1), 59. https://doi.org/10.1186/1471-2164-14-59

Kutschera, V. E., Bidon, T., Hailer, F., Rodi, J. L., Fain, S. R., & Janke, A. (2014). Bears in a forest of gene trees: Phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Molecular Biology and Evolution*, *31*(8), 2004–2017. https://doi.org/10.1093/molbev/msu186

Leaché, A. D., & McGuire, J. A. (2006). Phylogenetic relationships of horned lizards (Phrynosoma) based on nuclear and mitochondrial data: Evidence for a misleading mitochondrial gene tree. *Molecular Phylogenetics and Evolution*, *39*(3), 628–644. https://doi.org/10.1016/j.ympev.2005.12.016

Lehtiniemi, M., Engström-Öst, J., & Viitasalo, M. (2005). Turbidity decreases anti-predator behaviour in pike larvae, Esox lucius. *Environmental Biology of Fishes*, *73*(1), 1–8. https://doi.org/10.1007/s10641-004-5568-4

Levasseur, A., Orlando, L., Bailly, X., Milinkovitch, M. C., Danchin, E. G. J., & Pontarotti, P. (2007). Conceptual bases for quantifying the role of the environment on gene evolution: The participation of positive selection and neutral evolution. *Biological Reviews*, *82*(4), 551–572. https://doi.org/10.1111/j.1469-185x.2007.00024.x

Lewontin, R. (1974). *The genetic basis of evolutionary change* (Vol. 560). Columbia University Press, New York.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175–195. https://doi.org/10.1093/genetics/74.1.175

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A., & Nevo, E. (2002). Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Molecular Ecology*, *11*(12), 2453–2465. https://doi.org/10.1046/j.1365-294x.2002.01643.x

Liu, S., Zeng, Q., Wang, X., & Liu, Z. (2017). SNP array development, genotyping, data analysis, and applications. In *Bioinformatics in Aquaculture* (pp. 308–337). John Wiley & Sons, Ltd. http://dx.doi.org/10.1002/9781118782392.ch18

Lorenzoni, M., Corboli, M., Martin Dörr, A. J., Mearelli, M., & Giovinazzo, G. (2002). The growth of pike (Esox lucius Linnaeus, 1798) in Lake Trasimeno (Umbria, Italy). *Fisheries Research*, *59*(1–2), 239–246. https://doi.org/10.1016/s0165-7836(02)00013-9

Lucarda, A. N., Martini, M., Odore, R., Schiavone, A., & Forneris, G. (2008). Wild trout responses to a stress experience following confinement conditions during the spawning season. *Italian Journal of Animal Science*, *7*(1), 5–18. https://doi.org/10.4081/ijas.2008.5

Lucentini, L., Palomba, A., Gigliarelli, L., Sgaravizzi, G., Lancioni, H., Lanfaloni, L., Natali, M., & Panara, F. (2009). Temporal changes and effective population size of an Italian isolated and supportive-breeding managed northern pike (Esox lucius) population. *Fisheries Research*, *96*(2–3), 139–147. https://doi.org/10.1016/j.fishres.2008.10.007

Lucentini, L., Palomba, A., Lancioni, H., Gigliarelli, L., Natali, M., & Panara, F. (2006). Microsatellite polymorphism in Italian populations of northern pike (Esox lucius L.). *Fisheries Research*, *80*(2–3), 251–262. https://doi.org/10.1016/j.fishres.2006.04.002

Lucentini, L., Puletti, M. E., Ricciolini, C., Gigliarelli, L., Fontaneto, D., Lanfaloni, L., Bilò, F., Natali, M., & Panara, F. (2011). Molecular and phenotypic evidence of a new species of genus Esox (Esocidae, Esociformes, Actinopterygii): The southern pike, Esox flaviae. *PLoS ONE*, *6*(12), e25218. https://doi.org/10.1371/journal.pone.0025218

Lunt, J., & Smee, D. L. (2015). Turbidity interferes with foraging success of visual but not chemosensory predators. *PeerJ*, *3*, e1212. https://doi.org/10.7717/peerj.1212

Marchetto, Zaccara, Muenzel, & Salzburger. (2010). Phylogeography of the Italian vairone (Telestes muticellus, Bonaparte 1837) inferred by microsatellite markers: Evolutionary history of a freshwater fish species with a restricted and fragmented distribution. *BMC Evolutionary Biology*, *10*(1), 1–12. https://doi.org/10.1186/1471-2148-10-111

Marques, D. A., Jones, F. C., Di Palma, F., Kingsley, D. M., & Reimchen, T. E. (2018). Experimental evidence for rapid genomic adaptation to a new niche in an adaptive radiation. *Nature Ecology & Evolution*, *2*(7), 1128–1138. https://doi.org/10.1038/s41559-018-0581-8

Martínez-Páramo, S., Horváth, Á., Labbé, C., Zhang, T., Robles, V., Herráez, P., Suquet, M., Adams, S., Viveiros, A., Tiersch, T. R., & Cabrita, E. (2017). Cryobanking of aquatic species. *Aquaculture*, *472*, 156–177. https://doi.org/10.1016/j.aquaculture.2016.05.042

Maselli, V., Hutton, E. W., Kettner, A. J., Syvitski, J. P. M., & Trincardi, F. (2011). High-frequency sea level and sediment supply fluctuations during Termination I: An integrated sequence-stratigraphy and modeling approach from the Adriatic Sea (Central Mediterranean). *Marine Geology*, *287*(1–4), 54–70. https://doi.org/10.1016/j.margeo.2011.06.012

Mattucci, F., Galaverni, M., Lyons, L. A., Alves, P. C., Randi, E., Velli, E., Pagani, L., & Caniglia, R. (2019). Genomic approaches to identify hybrids and estimate admixture times in European wildcat populations. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-48002-w

McFarlane, S. E., Hunter, D. C., Senn, H. V., Smith, S. L., Holland, R., Huisman, J., & Pemberton, J. M. (2019). Increased genetic marker density reveals high levels of admixture between red deer and introduced Japanese sika in Kintyre, Scotland. *Evolutionary Applications*, *13*(2), 432–441. https://doi.org/10.1111/eva.12880

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the Python in Science Conference*. http://dx.doi.org/10.25080/majora-92bf1922-00a

Medina-Gomez, Felix, Estrada, Peters, Herrera, Kruithof, Duijts, Hofman, Duijn, van, Uitterlinden, Jaddoe, & Rivadeneira. (2015). Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: The Generation R Study. *European Journal of Epidemiology*, *30*(4), 317–330. https://doi.org/10.1007/s10654-015-9998-4

Meldgaard, T., Crivelli, A. J., Jesensek, D., Poizat, G., Rubin, J.-F., & Berrebi, P. (2007). Hybridization mechanisms between the endangered marble trout (Salmo marmoratus) and the brown trout (Salmo trutta) as revealed by in-stream experiments. *Biological Conservation*, *136*(4), 602–611. https://doi.org/10.1016/j.biocon.2007.01.004

Meraner, A., Baric, S., Pelster, B., & Dalla Via, J. (2009). Microsatellite DNA data point to extensive but incomplete admixture in a marble and brown trout hybridisation zone. *Conservation Genetics*, *11*(3), 985–998. https://doi.org/10.1007/s10592-009-9942-9

Meraner, A., Cornetti, L., & Gandolfi, A. (2014). Defining conservation units in a stocking‐induced genetic melting pot: Unraveling native and multiple exotic genetic imprints of recent and historical secondary contact inAdriatic grayling. *Ecology and Evolution*, *4*(8), 1313–1327. https://doi.org/10.1002/ece3.931

Meraner, A., & Gandolfi, A. (2017). Genetics of the Genus Salmo in Italy. In *Brown Trout* (pp. 65–102). John Wiley & Sons, Ltd. http://dx.doi.org/10.1002/9781119268352.ch3

Meraner, A., & Gandolfi, A. (2018). Anwendung der Genetik im aquatischen Artenschutz: Fokus Marmorierte Forelle. *WASSERWIRTSCHAFT*, *108*(2–3), 35–40. https://doi.org/10.1007/s35147-018-0019-x

Meraner, A., Gratton, P., Baraldi, F., & Gandolfi, A. (2012). Nothing but a trace left? Autochthony and conservation status of Northern Adriatic Salmo trutta inferred from PCR multiplexing, mtDNA control region sequencing and microsatellite analysis. *Hydrobiologia*, *702*(1), 201–213. https://doi.org/10.1007/s10750-012-1321-8

Meraner, A., Unfer, G., & Gandolfi, A. (2013). Good news for conservation: Mitochondrial and microsatellite DNA data detect limited genetic signatures of inter-basin fish transfer in Thymallus thymallus(Salmonidae) from the Upper Drava River. *Knowledge and Management of Aquatic Ecosystems*, *409*, 01. https://doi.org/10.1051/kmae/2013046

Meraner, A., Venturi, A., Ficetola, G. F., Rossi, S., Candiotto, A., & Gandolfi, A. (2013). Massive invasion of exotic Barbus barbus and introgressive hybridization with endemic Barbus plebejus in Northern Italy: Where, how and why? *Molecular Ecology*, *22*(21), 5295–5312. https://doi.org/10.1111/mec.12470

Miller, L. M., & Senanan, W. (2003). A Review of Northern Pike Population Genetics Research andIts Implications for Management. *North American Journal of Fisheries Management*, *23*(1), 297–306. https://doi.org/10.1577/1548-8675(2003)023<0297:aronpp>2.0.co;2

Mousseau, T. A., Sinervo, B., & Endler, J. (2000). *Adaptive genetic variation in the wild*. Oxford University Press on Demand.

Muhlfeld, C. C., Kalinowski, S. T., McMahon, T. E., Taper, M. L., Painter, S., Leary, R. F., & Allendorf, F. W. (2009). Hybridization rapidly reduces fitness of a native trout in the wild. *Biology Letters*, *5*(3), 328–331. https://doi.org/10.1098/rsbl.2009.0033

Mullis, K. (1990). The unusual origin of the polymerase chain reaction. *Scientific American*, *262*(4), 56–65.

Nei, M., Maruyama, T., & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution*, *29*(1), 1. https://doi.org/10.2307/2407137

Nelson, M. R., Marnellos, G., Kammerer, S., Hoyal, C. R., Shi, M. M., Cantor, C. R., & Braun, A. (2004). Large-Scale validation of single nucleotide polymorphisms in gene regions. *Genome Research*, *14*(8), 1664–1668. https://doi.org/10.1101/gr.2421604

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, *15*(11), 1566–1575. https://doi.org/10.1101/gr.4252305

Ohta, T., & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, *22*(2), 201–204. https://doi.org/10.1017/s0016672300012994

Orengo, D. J., & Aguadé, M. (2004). Detecting the Footprint of Positive Selection in a European Population of Drosophila melanogasterSequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AJ781836, AJ783306. *Genetics*, *167*(4), 1759–1766. https://doi.org/10.1534/genetics.104.028969

Paetkau, D., Waits, L. P., Clarkson, P. L., Craighead, L., & Strobeck, C. (1997). An empirical evaluation of genetic distance statistics using microsatellite data from bear (ursidae) populations. *Genetics*, *147*(4), 1943–1957. https://doi.org/10.1093/genetics/147.4.1943

Palombo, V., De Zio, E., Salvatore, G., Esposito, S., Iaffaldano, N., & D'Andrea, M. (2021). Genotyping of two Mediterranean trout populations in central-southern Italy for conservation purposes using a rainbow-trout-derived SNP array. *Animals*, *11*(6), 1803. https://doi.org/10.3390/ani11061803

Palti, Y., Gao, G., Liu, S., Kent, M. P., Lien, S., Miller, M. R., Rexroad, C. E., III, & Moen, T. (2014). The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources*, *15*(3), 662–672. https://doi.org/10.1111/1755-0998.12337

Park, S. T., & Kim, J. (2016). Trends in next-generation sequencing and a new era for whole genome sequencing. *International Neurourology Journal*, *20*(Suppl 2), S76-83. https://doi.org/10.5213/inj.1632742.371

Pavlidis, P., Jensen, J. D., Stephan, W., & Stamatakis, A. (2012). A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, *29*(10), 3237–3248. https://doi.org/10.1093/molbev/mss136

Pedreschi, D., Kelly‐Quinn, M., Caffrey, J., O'Grady, M., & Mariani, S. (2013). Genetic structure of pike (Esox lucius) reveals a complex and previously unrecognized colonization history ofIreland. *Journal of Biogeography*, *41*(3), 548–560. https://doi.org/10.1111/jbi.12220

Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. (2012). Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genetics*, *8*(10), e1003011. https://doi.org/10.1371/journal.pgen.1003011

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest radseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2017). *Scaling accurate genetic variant discovery to tens of thousands of samples*. Cold Spring Harbor Laboratory. http://dx.doi.org/10.1101/201178

Povz, M. (1995). Status of freshwater fishes in the Adriatic catchment of Slovenia. *Biological Conservation*, *72*(2), 171–177. https://doi.org/10.1016/0006-3207(94)00079-6

Prakash, Lewontin, & Hubby. (1969). A molecular approach to the study of genic heterozygosity in natural populations iv. Patterns of genic variation in central, marginal and isolated populations of drosophila pseudoobscura. *Genetics*, *61*(4), 841–858. https://doi.org/10.1093/genetics/61.4.841

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, *20*(4), R208–R215. https://doi.org/10.1016/j.cub.2009.11.055

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Pujolar, J. M., Lucarda, A. N., Simonato, M., & Patarnello, T. (2011). Restricted gene flow at the micro- and macro-geographical scale in marble trout based on mtDNA and microsatellite polymorphism. *Frontiers in Zoology*, *8*(1), 7. https://doi.org/10.1186/1742-9994-8-7

Pustovrh, G., Sušnik Bajec, S., & Snoj, A. (2012). A set of SNPs for Salmo trutta and its application in supplementary breeding programs. *Aquaculture*, *370–371*, 102–108. https://doi.org/10.1016/j.aquaculture.2012.10.007

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raat, A. J. (1988). *Synopsis of biological data on the northern pike, Esox lucius Linneaus, 1758*. Food & Agriculture Org.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Ramírez-Soriano, A., Ramos-Onsins, S. E., Rozas, J., Calafell, F., & Navarro, A. (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, *179*(1), 555–567. https://doi.org/10.1534/genetics.107.083006

Reed, D. H., & Frankham, R. (2001). How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. *Evolution*, *55*(6), 1095–1103. https://doi.org/10.1111/j.0014-3820.2001.tb00629.x

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., & Vilo, J. (2016). g:Profiler—a Web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, *44*(W1), W83–W89. https://doi.org/10.1093/nar/gkw199

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). g:Profiler—a Web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, *35*(suppl_2), W193–W200. https://doi.org/10.1093/nar/gkm226

Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, *72*(4), 686–727. https://doi.org/10.1128/mmbr.00011-08

Rubin, C. J., Megens, H.-J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C. B., Archibald, A. L., Fredholm, M., Groenen, M. A. M., & Andersson, L. (2012). Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences*, *109*(48), 19529–19536. https://doi.org/10.1073/pnas.1217149109

Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallböök, F., Besnier, F., Carlborg, Ö., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K., & Andersson, L. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, *464*(7288), 587–591. https://doi.org/10.1038/nature08832

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. https://doi.org/10.1038/nature01140

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive natural selection in the human lineage. *Science*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Saint-Pé, K., Leitwein, M., Tissot, L., Poulet, N., Guinand, B., Berrebi, P., Marselli, G., Lascaux, J.-M., Gagnaire, P.-A., & Blanchet, S. (2019). Development of a large SNPs resource and a low-density SNP array for brown trout (Salmo trutta) population genetics. *BMC Genomics*, *20*(1). https://doi.org/10.1186/s12864-019-5958-9

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Schlötterer. (2004). The evolution of molecular markers — just a matter of fashion? *Nature Reviews Genetics*, *5*(1), 63–69. https://doi.org/10.1038/nrg1249

Schlötterer, C. (2002). A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, *160*(2), 753–763. https://doi.org/10.1093/genetics/160.2.753

Schrider, D. R., & Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, *34*(8), 1863–1877. https://doi.org/10.1093/molbev/msx154

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Seeb, J. E., Seeb, L. W., Oates, D. W., & Utter, F. M. (1987). Genetic Variation and Postglacial Dispersal of Populations of Northern Pike (Esox lucius) in North America. *Canadian Journal of Fisheries and Aquatic Sciences*, *44*(3), 556–561. https://doi.org/10.1139/f87-068

Shi, H., Kichaev, G., & Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, *99*(1), 139–153. https://doi.org/10.1016/j.ajhg.2016.05.013

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, *139*(1), 457–462. https://doi.org/10.1093/genetics/139.1.457

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23–35. https://doi.org/10.1017/s0016672300014634

Sommani, E. (1950). *"Osservazioni sulla sistematica ed ecologia delle trote dell'Italia meridionale."* (Vol. 5, pp. 1–20). Bollettino di Pesca, Idrobiologia e Pescicoltura.

Song, K., Gao, B., Halvarsson, P., Fang, Y., Klaus, S., Jiang, Y.-X., E. Swenson, J., Han, Z.-M., Sun, Y.-H., & Höglund, J. (2022). Conservation genomics of sibling grouse in boreal forests reveals introgression and adaptive population differentiation in genes controlling epigenetic variation. *Zoological Research*, *43*(2), 184–187. https://doi.org/10.24272/j.issn.2095-8137.2021.227

Splendiani, A., Giovannotti, M., Righi, T., Fioravanti, T., Cerioni, P. N., Lorenzoni, M., Carosi, A., La Porta, G., & Barucchi, V. C. (2019). Introgression despite protection: The case of native brown trout in Natura 2000 network in Italy. *Conservation Genetics*, *20*(2), 343–356. https://doi.org/10.1007/s10592-018-1135-y

Splendiani, A., Ruggeri, P., Giovannotti, M., Pesaresi, S., Occhipinti, G., Fioravanti, T., Lorenzoni, M., Nisi Cerioni, P., & Caputo Barucchi, V. (2016). Alien brown trout invasion of the Italian peninsula: The role of geological, climate and anthropogenic factors. *Biological Invasions*, *18*(7), 2029–2044. https://doi.org/10.1007/s10530-016-1149-7

Stefani, F., Galli, P., Zaccara, S., & Crosa, G. (2004). Genetic variability and phylogeography of the cyprinid Telestes muticellus within the Italian peninsula as revealed by mitochondrial DNA. *Journal of Zoological Systematics and Evolutionary Research*, *42*(4), 323–331. https://doi.org/10.1111/j.1439-0469.2004.00272.x

Stephan, W. (2019). Selective sweeps. *Genetics*, *211*(1), 5–13. https://doi.org/10.1534/genetics.118.301319

Sternberg, D. (1992). *Northern Pike and Muskie: Tackle and Techniques for Catching Trophy Pike and Muskies.* Creative Publishing International.

Stronen, Mattucci, Fabbri, Galaverni, Cocchiararo, Nowak, Godinho, Ruiz-González, Kusak, Skrbinšek, Randi, Vlasseva, Mucci, & Caniglia. (2022). A reduced SNP panel to trace gene flow across southern European wolf populations and detect hybridization with other Canis taxa. *Scientific Reports*, *12*(1), 1–14. https://doi.org/10.1038/s41598-022-08132-0

Szatmári, L., Cserkész, T., Laczkó, L., Lanszki, J., Pertoldi, C., Abramov, A. V., Elmeros, M., Ottlecz, B., Hegyeli, Z., & Sramkó, G. (2021). A comparison of microsatellites and genome‐wide SNPs for the detection of admixture brings the first molecular evidence for hybridization between Mustela eversmanii and M. putorius (Mustelidae, Carnivora). *Evolutionary Applications*, *14*(9), 2286–2304. https://doi.org/10.1111/eva.13291

Tang, K., Thornton, K. R., & Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, *5*(7), e171. https://doi.org/10.1371/journal.pbio.0050171

Tortonese, E. (1970). *Osteichthyes (Pesci ossei). Parte prima.[Fauna of Italy, Vol. X. Osteichthyes (Bony fish). Part one.]*. Calderini.

Toscano, B. J., Pulcini, D., Hayden, B., Russo, T., Kelly-Quinn, M., & Mariani, S. (2010). An ecomorphological framework for the coexistence of two cyprinid fish and their hybrids in a novel environment. *Biological Journal of the Linnean Society*, *99*(4), 768–783. https://doi.org/10.1111/j.1095-8312.2010.01383.x

Trask, J. A. S., Malhi, R. S., Kanthaswamy, S., Johnson, J., Garnica, W. T., Malladi, V. S., & Smith, D. G. (2011a). The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (Macaca mulatta). *Primates*, *52*(2), 129–138. https://doi.org/10.1007/s10329-010-0232-4

Trask, J. A. S., Malhi, R. S., Kanthaswamy, S., Johnson, J., Garnica, W. T., Malladi, V. S., & Smith, D. G. (2011b). The effect of SNP discovery method and sample size on estimation of population genetic data for Chinese and Indian rhesus macaques (Macaca mulatta). *Primates*, *52*(2), 129–138. https://doi.org/10.1007/s10329-010-0232-4

Tsvetkov, N., MacPhail, V. J., Colla, S. R., & Zayed, A. (2021). Conservation genomics reveals pesticide and pathogen exposure in the declining bumble bee Bombus terricola. *Molecular Ecology*, *30*(17), 4220–4230. https://doi.org/10.1111/mec.16049

Vähä, J.-P., & Primmer, C. R. (2005). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, *15*(1), 63–72. https://doi.org/10.1111/j.1365-294x.2005.02773.x

Vainikka, A., Hyvärinen, P., Tiainen, J., Lemopoulos, A., Alioravainen, N., Prokkola, J. M., Elvidge, C. K., & Arlinghaus, R. (2021). Fishing-induced versus natural selection in different brown trout (Salmo trutta) strains. *Canadian Journal of Fisheries and Aquatic Sciences*, *78*(11), 1586–1596. https://doi.org/10.1139/cjfas-2020-0313

Väli, Ü., Saag, P., Dombrovski, V., Meyburg, B.-U., Maciorowski, G., Mizera, T., Treinys, R., & Fagerberg, S. (2010). Microsatellites and single nucleotide polymorphisms in avian hybrid identification: A comparative case study. *Journal of Avian Biology*, *41*(1), 34–49. https://doi.org/10.1111/j.1600-048x.2009.04730.x

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

van Wyk, A. M., Dalton, D. L., Hoban, S., Bruford, M. W., Russo, I.-R. M., Birss, C., Grobler, P., van Vuuren, B. J., & Kotzé, A. (2016). Quantitative evaluation of hybridization and the impact on biodiversity conservation. *Ecology and Evolution*, *7*(1), 320–330. https://doi.org/10.1002/ece3.2595

Vatsiou, A. I., Bazin, E., & Gaggiotti, O. E. (2015). Detection of selective sweeps in structured populations: A comparison of recent methods. *Molecular Ecology*, *25*(1), 89–103. https://doi.org/10.1111/mec.13360

Veggiani, A. (1974). *Le ultime vicende geologiche del Ravennate: influenza di insediamenti industriali sul circostante ambiente naturale [Recent geologic history of the Ravenna area: impacts of industry and manufacture infrastructure on the surrounding natural areas]. Studio sulla pineta di S. Vitale di Ravenna.* (pp. 48–58). Compositori.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072

Voight, B. F., & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics*, *1*(3), e32. https://doi.org/10.1371/journal.pgen.0010032

Waldvogel, A. M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., Mock, T., Schmid, K., Schmitt, I., Bataillon, T., Savolainen, O., Bergland, A., Flatt, T., Guillaume, F., & Pfenninger, M. (2020). Evolutionary genomics can improve prediction of species' responses to climate change. *Evolution Letters*, *4*(1), 4–18. https://doi.org/10.1002/EVL3.154

Weber, J. L., & Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Research*, *7*(5), 401–409. https://doi.org/10.1101/gr.7.5.401

Weigand, H., & Leese, F. (2018). Detecting signatures of positive selection in non-model species using genomic data. *Zoological Journal of the Linnean Society*, *184*(2), 528–583. https://doi.org/10.1093/zoolinnean/zly007

Weir, B. S. (1979). Inferences about linkage disequilibrium. *Biometrics*, *35*(1), 235. https://doi.org/10.2307/2529947

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Welcomme, R. L. (1988). *International introductions of inland aquatic species*. Food & Agriculture Org.

Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., & Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *PLoS Genetics*, *3*(6), e90. https://doi.org/10.1371/journal.pgen.0030090

Willoughby, J. R., & Christie, M. R. (2018). Long-term demographic and genetic effects of releasing captive-born individuals into the wild. *Conservation Biology*, *33*(2), 377–388. https://doi.org/10.1111/cobi.13217

Willoughby, J. R., Ivy, J. A., Lacy, R. C., Doyle, J. M., & DeWoody, J. A. (2017). Inbreeding and selection shape genomic diversity in captive populations: Implications for the conservation of endangered species. *PLOS ONE*, *12*(4), e0175996. https://doi.org/10.1371/journal.pone.0175996

Yang, S., Li, X., Li, K., Fan, B., & Tang, Z. (2014). A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. *BMC Genetics*, *15*(1). https://doi.org/10.1186/1471-2156-15-7

Young, W. P., Wheeler, P. A., Coryell, V. H., Keim, P., & Thorgaard, G. H. (1998). A detailed linkage map of rainbow trout produced using doubled haploids. *Genetics*, *148*(2), 839–850. https://doi.org/10.1093/genetics/148.2.839

Zargar, S. M., Raatz, B., Sonah, H., MuslimaNazir, Bhat, J. A., Dar, Z. A., Agrawal, G. K., & Rakwal, R. (2015). Recent advances in molecular marker techniques: Insight into QTL mapping, GWAS and genomic selection in plants. *Journal of Crop Science and Biotechnology*, *18*(5), 293–308. https://doi.org/10.1007/s12892-015-0037-5

Zecherle, L. J., Nichols, H. J., Bar-David, S., Brown, R. P., Hipperson, H., Horsburgh, G. J., & Templeton, A. R. (2021). Subspecies hybridization as a potential conservation tool in species reintroductions. Evolutionary Applications, 14(5), 1216–1224. https://doi.org/10.1111/eva.13191

Zhong, L., Zhu, Y., & Olsen, K. M. (2022). Hard versus soft selective sweeps during domestication and improvement in soybean. *Molecular Ecology*, *31*(11), 3137–3153. https://doi.org/10.1111/mec.16454