UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Biologia Evoluzionistica

TESI DI LAUREA

In Vivo Validation of Bioinformatically Predicted Genetic Load in the Tetraploid Adriatic Sturgeon

Relatore: Prof. Leonardo Congiu
Dipartimento di Biologia

Correlatori: Dott. Stefano Dalle Palle
Dipartimento di Biologia

Prof. Giorgio Bertorelle
Dipartimento di Scienze della vita e biotecnologie
– Università degli studi di Ferrara

Laureanda: Francesca Quitadamo

ANNO ACCADEMICO 2023/2024

Table of Contents

Abstract

The tetraploid Adriatic sturgeon, *Acipenser naccarii*, is an endemic species of the North Adriatic region, which has experienced a dramatic population decline in the past 50 years. It is currently classified as ‘Critically Endangered’ in the IUCN Red List. The species has been targeted by several conservation efforts and reintroduction interventions, but its future still relies on the correct ex-situ management. Next generation sequencing technologies and the latest genomic tools provide valuable resources to assist conservation plans, as they allow us to screen the genome and reveal new insights into the accumulation patterns of mutations in declining populations, and on how the mutation load affects individual fitness. The present work is part of the PRIN project EndemixIT, which aims precisely to apply conservation genomics to five Italian endemic species. Even if whole-genome data can be used to predict the genetic load in endangered populations, the relationship between these predictions and the real fitness of the individuals is still poorly investigated. The present work aims to fill this gap predicting the genetic load and estimating the fitness in the offspring captive stocks of *A. naccarii*. Six pairs of individuals were selected to perform 12 different crosses. The genomes of the parental individuals were screened to obtain a panel of SNPs with alleles predicted to have different levels of negative effects on the fitness. Offspring individuals were genotyped-by-sequencing at the SNPs panel, and their survival and growth rates were recorded. A global negative correlation between predicted load and fitness was found, but different crosses showed different patterns and more detailed analyses of specific genetic variants are required.

1. Introduction

1.1 Preamble

Biodiversity is threatened worldwide due to human footprint. Climate change, overexploitation and habitat degradation are leading to the Earth's sixth mass extinction (Barnosky et al., 2011). According to the IUCN Red List of Threatened Species (International Union for Conservation of Nature (IUCN)., 2023), 45,300 out of the 163,000 species assessed are currently threatened with extinction. Conservation efforts will benefit from the latest genomic technologies (Supple & Shapiro, 2018). In fact, thanks to the enhanced resolution provided by genomic tools, researchers can detect fine-scale genetic variations and evolutionary patterns that were previously undetectable (Allendorf et al., 2012). This allows to explore new questions about the genetic basis of adaptation, the impact of genetic drift, and inbreeding in small populations (Shafer et al., 2015). Additionally, genomics helps assessing the adaptive potential of declining populations to respond to environmental changes, such as climate change and habitat loss (Supple & Shapiro, 2018).

In this context, the EndemixIT project was established (https://endemixit.com/). This conservation genomics project, funded by the Italian Ministry of Education, University and Research (PRIN programs), engages six Italian universities (University of Ferrara, University of Padua, University of Florence, University of Trieste, Polytechnic University of Marche, University of Rome 'Tor Vergata'), and is coordinated by Professor Giorgio Bertorelle from the University of Ferrara. The general aim of the project was to study the genomic variation and patterns in five endangered Italian endemics. Among the more specific objectives, EndemixIT investigated the genome-wide dynamics of accumulation of genetic load in small populations and its effects on individual and population fitness.

Indeed, small and endangered populations are likely to enter the so-called 'extinction vortex': due to reduced population size, genetic drift acts as the major force shaping the composition of the genetic pool in the population, reducing the efficiency of natural selection and possibly leading to the

accumulation of deleterious mutations (the mutation load). This process hampers individual and population fitness, leading to a further decline in population size and, ultimately, to extinction (Frankham, 2005; Gilpin & Soulé, 1986; Lynch et al., 1995).
Therefore, understanding the pattern of accumulation of mutation load can help designing conservation plans aiming to prevent genomic erosion, thus reducing extinction risk.

One of the species addressed by the EndemixIT project is the tetraploid Adriatic sturgeon, *Acipenser naccarii*, which is considered to be at risk of extinction. The species is, in fact, currently classified as ‘Critically Endangered’ by the IUCN Red List (International Union for Conservation of Nature (IUCN)., 2023) and is included in Appendix II of the Convention on International Trade of Endangered Species (CITES). Within EndemixIT, *A. naccarii* was targeted through an *in vivo* study aiming to examine the mutation load segregation pattern and the relationship with individual fitness, performing controlled inbred and outbred crosses.

The work presented here has been carried out within the described framework and was conducted in collaboration with the University of Ferrara, in parallel with another master thesis project (Bordogna, Federica. *Deleterious mutations in the tetraploid Adriatic sturgeon (Acipenser naccarii): from bioinformatic predictions to real phenotypic effects*. 2024. University of Bologna, Master’s thesis).

1.2 *Acipenser naccarii* and its conservation status

The Adriatic sturgeon, *Acipenser naccarii*, is an Italian representative of the taxon of Acipenseriformes, endemic to the northern Adriatic region. Acipenseriformes are a distinctive order of chondrostean ray-finned fishes, inhabiting freshwater and coastal environments across the northern hemisphere. This lineage includes 26 extant species (25 sturgeon species and one paddlefish species) and, placed at the base of the ~30,000 modern teleost species, is noteworthy for its evolutionary significance.

Acipenseriformes are commonly referred to as ‘living fossils’ due to their almost unchanged morphology with respect to fossil records dated approximately 200 Mya (Gardiner, 1984), and in light of the slower rate of molecular evolution observed (Krieger & Fuerst, 2002). Moreover, these species exhibits different levels of ploidy, with taxa having up to 380 chromosomes (Du et al., 2020) due to multiple whole-genome duplication events, occurred independently in diverse branches of this clade (Peng et al., 2007).

Unfortunately, the order of Acipenseriformes has been significantly impacted by human activities: in the past century, overexploitation for caviar and meat, illegal trade, habitat modification, and pollution led to severe declines in these species. Twenty out of 26 species are classified as Endangered or Critically Endangered in the IUCN Red List (IUCN). Lately, the IUCN Sturgeon Specialist Group reported a continued decline in paddlefish and sturgeon populations (Congiu et al., 2023). The Chinese paddlefish *Psephurus gladius*, and the Dabry sturgeon, *Acipenser dabryanus*, both endemic to the Yangtze River, are, respectively, Extinct and Extinct in the Wild. The eight European species, though being targeted by conservation efforts since 1992, are either Endangered or Critically Endangered, according to the latest IUCN update of IUCN Red List (International Union for Conservation of Nature (IUCN)., 2023). North American species present a somewhat better condition than Eurasian ones, resulting from prompter conservation efforts; nevertheless, their situation has also seen a further decline.

Up to 50 years ago, *A. naccarii* was commonly found in the Po River and its tributaries, and in the eastern coast of the Adriatic Sea. But habitat modification, overfishing, and pollution have almost led the species to extinction (Boscari et al., 2015).

This species is anadromous, meaning that it needs to migrate to a freshwater basin to spawn. The construction of dams and barriers to migration is a well-known cause of habitat disturbance for sturgeon species and it is considered one of the main factors contributing to wild populations decline, having caused alteration or even destruction of spawning sites (Williot et al., 2002). For example, since 1960s the Isola Serafini Dam (Piacenza, Italy), in the Po River, has prevented *A. naccarii's* migratory movements, severely reducing spawning substrates (Lohe, 2021). A fish passage has been constructed to restore connectivity but it is possibly not suitable for Adriatic sturgeon, given their large size at maturity; thus the presence of the dam continues to modify Adriatic sturgeon's habitat (Lohe, 2021). An additional threat to the recovery of this species is represented by climate change, since sea level rise is projected to modify salinity in estuarine habitats and increased nutrient load from human activities will cause expansion of hypoxic areas (IPCC, 2023). Warmer and more hypoxic waters have been shown to cause developmental impairment and reduced survival of early life stage in Adriatic sturgeon (Delage et al., 2020).

The captive breeding program started in 1977 prevented the extinction of the species. Initially, 90 wild immature individuals (F0) were transferred to a private aquaculture plant, the Azienda Agricola V.I.P. (Orzinuovi, Brescia, Italy); 50 of them reached maturity and the first successful reproduction in captivity occurred in 1988. Since then, several F1 stocks were generated crossing F0 breeders, and over 500,000 individuals have been released in the wild (Arlati & Poliakova, 2009). All the F1 individuals currently present in aquaculture facilities in Italy and abroad descend from the F0 stock, now reduced to thirteen animals, which represent the last individuals of clear wild origin. Though the genetic variability of F1 stocks' is significantly reduced compared to the stock of wild origin, due to lack of genetic support in the past reproductions (Boscari & Congiu, 2014), this species is showing sporadic signs of recovery. Indeed, its status in the IUCN Red List has recently been updated from 'Extinct in the Wild' to 'Critically Endangered', having found juveniles linked to natural reproduction in the Po basin (Congiu et al., 2021).

1.3 Genetic load and its consequences on fitness

Mutations are the source of genetic variation, which provides the raw material for adaptation and evolution. However, most mutations occurring across the genome are detrimental (Bataillon & Bailey, 2014; Kimura, 1977). Deleterious mutations, except for those with lethal effects, can remain in a population until they are wiped out by selection, and, in the meantime, new mutations appear, thus they are always present and are likely to affect individual fitness, population abundance, and extinction risk (Agrawal & Whitlock, 2012). Therefore, understanding genetic load is relevant for conservation biology, and research directed toward its accumulation and purging dynamics is expanding to inform management practices for threatened and declining populations.

Genetic load is defined in different ways. Put simply, it can be defined as the decrease in the mean fitness of a population relative to the optimal genotype, resulting from deleterious alleles. Genetic load is measured in lethal equivalents which correspond to the sum of negative selection coefficients of the deleterious alleles (Bertorelle et al., 2022), and it encapsulates the combined effects of selection and dominance coefficients of mutations throughout the genome, as a function of their frequencies. The total genetic load can be partitioned into two components: the realized load, which is the fitness reduction in the current population, and the masked load, which represents the potential decrease in the mean fitness in future generations due to the expression of deleterious mutation following demographic events such as inbreeding or population decline (Bertorelle et al., 2022). Theory and empirical data suggest that prolonged bottlenecks, due to the conversion of masked load into realized load, may cause a reduction in the total genetic load, through the purging of highly and mildly deleterious mutations (Caballero et al., 2017), while slightly deleterious mutations tend to accumulate through genetic drift (Dussex et al., 2023). The impact of timescale, the strength of the bottleneck, and the balance between genetic drift and selection (which depends on the distribution of fitness effects of mutations) are crucial to understanding the pattern of accumulation and purging of deleterious mutation, therefore the long-term viability and persistence of the population.

Genetic load can be predicted from whole-genome data. To do so, several tools have been developed, following mainly two types of approaches. The first is a comparative genomic approach, which relies on multiple alignments and assumes that mutations occurring at conserved sites are those with a higher impact on fitness. One of the most used methods based on the evolutionary conservation approach is Genomic Evolutionary Rate Profiling (GERP), which identifies sites under purifying selection, assigning to them higher impact scores (Cooper et al., 2005). This method does not require an annotated reference genome, thus being suitable for non-model organisms, but the reliability of the link between substitution scores and selection strength could be affected by possible changes in the selection coefficient across lineages and by the presence of lineage-specific adaptations.

Other methods rely, instead, on the prediction of fitness effects with respect to protein-coding genes, hence they are based on functional annotations and information from biochemical studies. In this way, mutations are classified as synonymous or non-synonymous, and ultimately categorized into impact classes (low, moderate, or high impact). An example is given by SnpEff software, which supplies further information on the effects of mutations, identifying variants such as missense, frameshift, stop-gain, and stop-loss (Cingolani et al., 2012).

Linking predicted high deleterious variation with population fitness is not straightforward, and though we are now able to interrogate the whole genome to predict the effect of mutant alleles, direct insights about how bioinformatically inferred mutation load impacts fitness in non-model organisms are still limited.

Quantitative genetic studies have enhanced our comprehension of the relationship between inbreeding load (which is the same as masked load) and direct fitness estimates, in populations that experienced demographic shifts and different levels of inbreeding. These load estimates are obtained from the slope of the linear regression of phenotypic values used as proxies for the fitness on pedigree-derived inbreeding coefficients (Nietlisbach et al., 2019; Van Oosterhout et al., 2007; C. W. Fox et al., 2008).

Instead, genetic load estimates inferred from sequence data have been rarely tested in their ability to predict real fitness consequences, and this relationship has been validated mostly for known human disease-causing variants or in model organisms (Kono & Al., 2018).

Doekes et al. (2021) compared, in a meta-analysis on livestock, pedigree-based with SNP-based measures of inbreeding, finding a consistent trend of declining fitness with increasing inbreeding; while Stoffel et al., (2021) showed, through empirical data and simulations, that inbreeding depression increases with Runs of homozygosity’s length, and that the latter is associated with a higher amount of deleterious mutations, thus linking the predicted mutation load with fitness with an indirect inference.

Estimating the genetic load from genome sequences s a relevant issue in conservation biology: the load data can be used, for example, in captive breeding programs to prevent gene pool deterioration, selecting individuals with the lowest load as breeders. With appropriate breeding strategies, this can be achieved while minimizing the cost of genetic variability reduction (Van Oosterhout, 2020). Moreover, the use of this type of data can guide genetic rescue programs, preventing the reintroduction of deleterious mutations in the wild (Van Oosterhout, 2020).

1.4 Polyploidy: effects on fitness and challenges in genotyping

The genome of the Adriatic sturgeon is generally considered functionally tetraploid, with a total of 240 chromosomes (Fontana et al., 1999). Polyploidy arises as a consequence of whole-genome duplication (WGD) events. The first event of WGD arose in the sturgeons’ common ancestor, which had a chromosome number of 2n = 60. Subsequently, the Pacific and Atlantic clade underwent secondary WGD events, which led to a total of 240 chromosomes (Peng et al., 2007). Currently, a debate between two distinct positions exists regarding the number of chromosomes associated with each ploidy level. While the first defines the nominal ploidy with respect to the number of WGD events, the second is based on functional aspects and defines the ploidy level in view of the number of gene copies observed to be active through cytogenic analysis (Fontana et al., 2007). In this light, the first position argues that species with 120 chromosomes should be considered tetraploid, deriving from the first WGD event in the common ancestor, thus species with 240 chromosomes would be octoploid. The second point of view associates, instead, a state of diploidy and tetraploidy to species with 120 and 240 chromosomes, respectively. These two positions are not mutually exclusive and are both correct (Fontana et al., 2008).

Poliploidy is agreed to have several downstream effects, such as changes in gene expression and cell size, alterations in gene interaction networks and epigenetics, and modified responses to environmental stress (D. T. Fox et al., 2020).

Genome doubling has, consequently, an impact on the way deleterious mutations affect the fitness of the individual. One of the reasons is the buffering effect derived from gene redundancy (Comai, 2005). Additionally, some authors claim that polyploids should suffer less of inbreeding depression than diploids, due to the reduced probability to form full homozygous genotypes (Clo et al., 2022). On the other hand, the masking of mutant alleles observed in polyploids would allow mutations to persist in the population and reach a higher frequency, which, over time, would overcome the benefits of masking, and reduce the equilibrium fitness by a higher degree with respect to diploids (Otto, 2007).

1.4.1 Genotyping in polyploids

Genotyping in polyploids is particularly challenging, given the existence of different degrees of heterozygosity. In the case of biallelic SNPs, which is what dealt with in this study, the tetraploid status implies up to five genotypic classes at each locus: 0/0/0/0 (*nulliplex*), 0/0/0/1 (*simplex*), 0/0/1/1 (*duplex*), 0/1/1/1 (*triplex*), 1/1/1/1 (*quadruplex*). According to Matias et al. (2019), *nulliplex* and *quadruplex* are, in theory, more likely to be unambiguously inferred, while a higher degree of uncertainty exists for the heterozygous states. Specifically, tetraploid genotyping tends to suffer more from bias against *duplex* genotype, being erroneously inferred as *simplex* or *triplex* (Matias et al., 2019). To prevent biases in the assessment of allele dosage, studies on autotetraploid crops recommended increasing sequencing depth up to 50x-80x (Uitdewilligen et al., 2013; Bastien et al., 2018), but this implies higher costs of genotyping. This issue could be circumvented adopting a Bayesian approach, which would possibly allow to have no missing data since, even when the read depth is zero, the most probable genotype can be predicted from the priors. Moreover, using continuous genotypes (i.e. genotypes expressed as continuous values rather than discrete genotype classes) helps preventing genotypic class misclassification even with low sequencing depths; for example, a study of genomic prediction in the autotetraploid blueberry demonstrated that, using this approach, a predictive accuracy similar to the one reached with higher depths can be achieved under low-depth scenarios (6x-12x) (de Bem Oliveira et al., 2020).

Softwares like polyRAD (Clark et al., 2019), specifically designed for polyploids organisms, incorporate informative priors, allowing an accurate genotype calling with low depth of coverage and, additionally, permit to take in account specific issues of polyploid genomes. Moreover, the program allows to export both the most probable genotype as well as continuous numerical genotypes that incorporate the relative probabilities associated to all possible allele copy numbers.

2. Aim of the thesis

Deleterious variation has a central role in hampering small population viability and determining extinction risk. Supporting small populations persistence, especially in the case in which they are destinate to remain small and highly inbred, implies that efforts should be put towards mutation load minimization (Kyriazis et al., 2020). Moreover, genomics-informed conservation practices can diminish load and at the same time minimize genome-wide diversity loss, hence are expected to become a significant instrument to promote long-term viability of captive populations (Speak et al., 2024; Oosterhout, 2020).

The present work has been carried out within the framework of the PRIN EndemixIT project (https://endemixit.com/), and has two main purposes: to understand the relationship between bioinformatic predictions and the real fitness, and, consequently, to obtain useful information for the correct management and conservation of the Adriatic sturgeon. To accomplish this, controlled crosses have been performed in the ex-situ managed *Acipenser naccarii*, and the genomes of the selected breeders have been completely sequenced. Then, a genotyping-by-sequencing (GBS) has been conducted in the offspring and eventually the correlation between observed individual fitness and bioinformatic predictions has been analysed.

3. Materials and Methods

3.1 Samples selection and fitness assessment

Twelve different crosses were performed from six pairs of first-generation (F1) full sibs, selected from the stock kept in captivity at the aquaculture plant ‘Storione Ticino’, obtaining six inbred and six outbred families. About 500 eggs from each cross were sampled, and hatched larvae were kept in a controlled environment for two months.

Standard length and weight were measured as fitness proxies. Due to high mortality in the first days, two families were reduced to a few individuals, thus fitness assessments were only possible for ten out of the twelve initial groups. The results shown here are relative to the latest measurements, taken in July 2021.

Moreover, all dead larvae had been sampled and stored in ethanol, allowing for the discrimination of two additional differentially performing categories, that hereafter are referred to as Larvae, consisting of those individuals who died before the yolk sac absorption, and Juveniles, which include the ones who appeared best performing (in terms of growth rate) at the end of the observations.

At last, a total of 384 offspring individuals, from 10 different families, were selected for genotyping.

3.2 Loci identification

The genomes of the twelve breeders have been completely sequenced (Illumina NovaSeq 6000 system at the University of Florence), producing paired-end reads, with an average coverage of 34.9X (SD = 9.6) per sample. The obtained reads were mapped with Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009) to the closely related sterlet sturgeon, *Acipenser ruthenus*, genome (Du et al., 2020). A *de novo* assembly of the *Acipenser naccarii* genome is also part of the EndemixIT project, but it is currently unavailable.

Variant calling has subsequently been performed, with the GATK HaplotypeCaller tool (Ryan et al., 2018). A hard filtering has been carried out, according to GATK Best Practices (Van der Auwera et al., 2013), depending on the bases' quality and excluding variants mapping in repeated regions. Then, indels were discarded, and only biallelic SNPs were retained. Only loci with a Genotype Quality (GQ) greater than 10 were selected, and the impact of the variants was predicted with the SnpEff program (Cingolani et al., 2012), so that each SNP has been classified as belonging to one of the following categories of effects: High-impact variants, which include start-codon loss and stop-codon gain mutations; Missense variants; Synonymous variants, which serve as control.

1000 loci were selected to perform a genotyping-by-sequencing in the offspring, according to several criteria: first, the locus had to be present in every parental individual; then, we maintained only loci that could form homozygous genotypes for the mutated allele in the offspring, and, for this purpose, only loci with at least two mutated alleles in at least one parental individual were kept. Loci were also divided into frequency classes, based on the Allele Frequency (AF), which is computed as the ratio between the count of the mutated allele in the sample (Allele Count, AC) and the total number of alleles (or the number of all the possible positions, AN). The obtained frequencies were divided into three ranges: 0 - 0.3, 0.3 - 0.7, 0.7 -

1. All the High-impact mutations that fell into the higher frequency range were discarded; this was done to prevent the possibility of having SNPs miscalled as High, given that the presence of fixed, highly deleterious variants with no visible consequence on fitness would have been doubtful. For the same reason, loci in complete homozygosity for the alternative allele in the breeders were also excluded. Additionally, the selected SNP had not to be in linkage, thus they were filtered to map in a window of 100000 bp; this was an attempt to avoid redundancy of the information carried from the single SNP.

After these filtering steps, almost all of the High mutations were maintained for genotyping, the missense and synonymous mutations were instead randomly taken. A total of 400 High, 400 Missense, and 200 Synonymous SNPs were in this way retained.

3.3 Allele dosage estimation

For these loci, the chosen offspring individuals were genotyped-by-sequencing at ‘Floodlight Genomics LLC’ (Knoxville, TN, USA). The reads thus obtained were mapped against the reference genome (*Acipenser ruthenus*). At last, 798 SNPs were successfully amplified and sequenced, with an average read depth of 43.6X (DS=18.3). The variant calling has been performed with the GATK tool.

At this stage, the offspring genotyping was carried out, following two different approaches.

First, the genotypes were inferred with the GATK HaplotypeCaller, as was already done for the breeders. This approach led to a high amount of missing data.

We therefore proceeded with a Bayesian approach, using the polyRAD genotype calling tool (Clark et al., 2019). With this program, designed specifically for polyploid organisms, it is possible to estimate the posterior probability for each possible allele dosage class, using Bayes’ theorem (Box 1). For the present work, the prior was obtained from parental read depth of the two alleles (which roughly represents the parental genotype), using the

"PipelineMapping2Parents" pipeline, with default parameter values. Segregation was assumed to be tetrasomic.

$$P(G|D) = \frac{L(D|G)P(G)}{\sum_{i=1}^{k} L(D|G_i)P(G_i)}$$

- *P(G|D)* = posterior probability of the genotype
- $L(D|G)P(G)$ = likelihood of the observed distribution of allelic read depth (D) if the given genotype G were the true genotype
- P(G) = prior probability of the genotype
- *k* = possible genotypes

Box 1. Bayes' theorem applied to genotype calling.

3.4 Post hoc filtering

Before proceeding with the downstream analysis, loci with excessively high depth (3 standard deviations above the mean) were removed, to take into account read misplacements errors. Then, we imposed various filters based on the Minimum Likelihood Ratio (MLR) for determining parental genotypes with confidence, and on the read depth (DP). The different filters' combinations were inspected in order not to lead to an excessive amount of missing data and to any bias towards specific genotype classes (Table S1-S2). Only results relative to the most conservative filter are shown (Minimum Likelihood Ratio = 10, DP > 10); the ratio of the depth threshold is based on the observation that, when comparing GATK's and polyRAD's genotype predictions, loci with excessively low read depth showed conflicting results.

			Original dataset	Filtered dataset
tot	Loci	N°	798	727
		fraction	1	0,83
High	Loci	N°	318	293
		fraction	1	0,86
Missense	Loci	N°	329	301
		fraction	1	0,82
Synonymous	Loci	N°	150	133
		fraction	1	0,81

Table 1. Dataset size before and after the filtering steps, reporting the number of loci present in at least one individual. The original dataset column refers to the data obtained from the vcf file produced after the variant calling of the genotyped-by-sequencing loci and used as input for PolyRAD genotype calling. The filtered dataset was obtained running the PolyRAD pipeline and imposing the post-hoc filtering (details in the text). The “fraction” rows show the proportion of loci retained, for each impact class, after filtering, relative to the original dataset size.

3.5 Genetic Load proxies

At the individual level, the mean number of alternative alleles per locus, for every impact class (High (H), Missense (M), Synonymous (S)), has been calculated as follows:

$$MEAN\_ALT_k = \frac{1}{n_k} \sum_{i=1}^{n_k} N_{altk} \qquad (1)$$

with:

- N_{altk} = number of alternative alleles at each locus, derived from the most probable genotype (i.e. with the higher posterior probability), for every *k* impact class (H, M or S)
- n_k = total number of loci available for each *k* impact class in the individual under consideration

Subsequently, R_{HS} and R_{MS} indexes were defined as the mean number of alternative alleles at each H/M locus, standardized for the mean number of alleles at each S locus. This type of standardization has been implemented to identify the effects of deleterious mutation independently from the level of synonymous, and therefore likely neutral, variation. Indeed, the total number of Synonymous alternative alleles was found to be correlated with the number of mutated alleles at High-impact (Spearman's ρ = 0.565, p < 2.2e-16) and Missense loci (Spearman's ρ = 0.615, p < 2.2e-16).

$$R_{HS} = \frac{\frac{1}{n\_H}\sum_{i=1}^{n\_H} N_{altH}}{\frac{1}{n\_S}\sum_{i=1}^{n\_S} N_{altS}} \qquad (2)$$

$$R_{MS} = \frac{\frac{1}{n\_M}\sum_{i=1}^{n\_M} N_{altM}}{\frac{1}{n\_S}\sum_{i=1}^{n\_S} N_{altS}} \qquad (3)$$

The High[Missense]-Synonymous indexes were also weighted for the posterior probabilities associated with each possible genotype at the given locus.

$$weighted\ R_{H[M]S} = \frac{MW_{ALT_{H[M]}}}{MW_{ALT_S}} \qquad (4)$$

with:

- MW_{ALT_k} = mean number of alternative alleles at locus, weighted for the posterior probabilities associated to every possible genotype = $\frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{i=1}^{j} N_{jn_k}p_{j_{n_k}}$ (5)

- n_k = total number of loci for each *k* impact class

- N_{jn_k}= number of alternative alleles associated with the *j* genotype, at locus n_k

- $p_{j_{n_k}}$= posterior probability of the *j* genotype

The term $\sum_{i=1}^{j} N_{jn_k}p_{j_{n_k}}$ in Formula (5) represents a calculation of the continuous genotype, obtained weighting every possible allele copy number at the given locus for the relative posterior probability.

3.6 Statistical analysis

All analyses were performed using RStudio, R version 4.3.2 (2023-10-31 ucrt). First, the data were explored to inspect patterns, correlations, and outliers. The individual F5M5_J13 was excluded from the analysis due to the low number of loci retained after the post-hoc filtering steps.

With the purpose of examining the correlation between the computed genetic load indexes and the fitness indicators, we fitted linear regression models (Fitness ~ Load) using the ‘lm’ function from the *stats* R package. For this part of the analysis, the dataset consisted only of Juvenile individuals. We compared models based on genetic load proxies calculated from discrete genotypes ($R_{H[M]S}$) with models using proxies obtained from continuous genotypes ($weighted\, R_{H[M]S}$) to inspect differences in the explanatory power of the two approaches. Parametric assumptions of linearity, normality of residuals, and homoscedasticity were verified using normal Q-Q plots, residual vs. fitted plots, the Shapiro-Wilk normality test on residuals, and the Breusch-Pagan test (*stats* R package, *lmtest* package (Zeileis & Hothorn, 2002)).

To test whether differences in individual performances were largely influenced by the family factor, the interaction between the two explanatory variables (Load Estimate and Family) was checked and found not significant. Thus, only the additive effect was inspected. We therefore computed a linear mixed-effects model using the ‘lmer’ function from the *lme4* package (Douglas et al., 2015), with Family as a random factor (Fitness ~ Load + (1 | Family)), to account for the fact that individuals from the same family do not represent independent observations. For the same purpose, we fitted a linear regression model (‘lm’) summarizing the mean values in fitness indicators and load proxies for each family.

Then, we examined differences in mutation load between the two differentially performing groups, Larvae and Juveniles. A Student's t-test was computed, and whenever the explanatory variable had a frequency distribution significantly deviating from a normal distribution, we used the

Mann-Whitney U test from the *stats* R package. The latter was also computed when parametric assumptions were met, to provide comparable results.

Again, the effect of the Family as random factor was inspected through a linear mixed effect model (Load ~ Group + (1 | Family).

4. Results and Discussion

4.1 Effect of the mutation load on individual Weight and Length

A variety of models were fitted to inspect the existing correlation between bioinformatically derived mutation load estimates and fitness indicators.

Among the parameters measured to assess fitness, body weight and standard length were chosen, and the relationship with mutation load has been explored in parallel. Both body weight and standard length, are, by themselves, incomplete to estimate fitness: the weight of an individual strongly depends on its foraging ability, therefore, the presence of other individuals in the same tanks can influence it, imposing competition constraints. The individual length, instead, is not necessarily a predictor of an animal's health, as demonstrated by the presence of lean and possibly unfit individuals with high values of length. These issues have been overcome by selecting individuals with the highest weight, hence it is reasonable to assume that, among the selected ones, the longest individuals are also the fittest.

All the linear regression models computed showed a consistent, highly significative pattern of decreasing fitness estimates as the genetic load increased (Figure 1).

This pattern is in accordance with the putative effect of mutations: when looking at the models' results (Table 2), the effect of the genetic load calculated on High-impact loci appears more evident as compared to Missense loci, in terms of slope, p-value and R-squared. Additionally, the correlation between estimated genetic load and fitness proxies seems more robust when the models include load estimates derived from continuous genotype ($weighted\ R_{H[M]S}$) as explanatory variables (Figure 1C-D-E-F, Table 2).

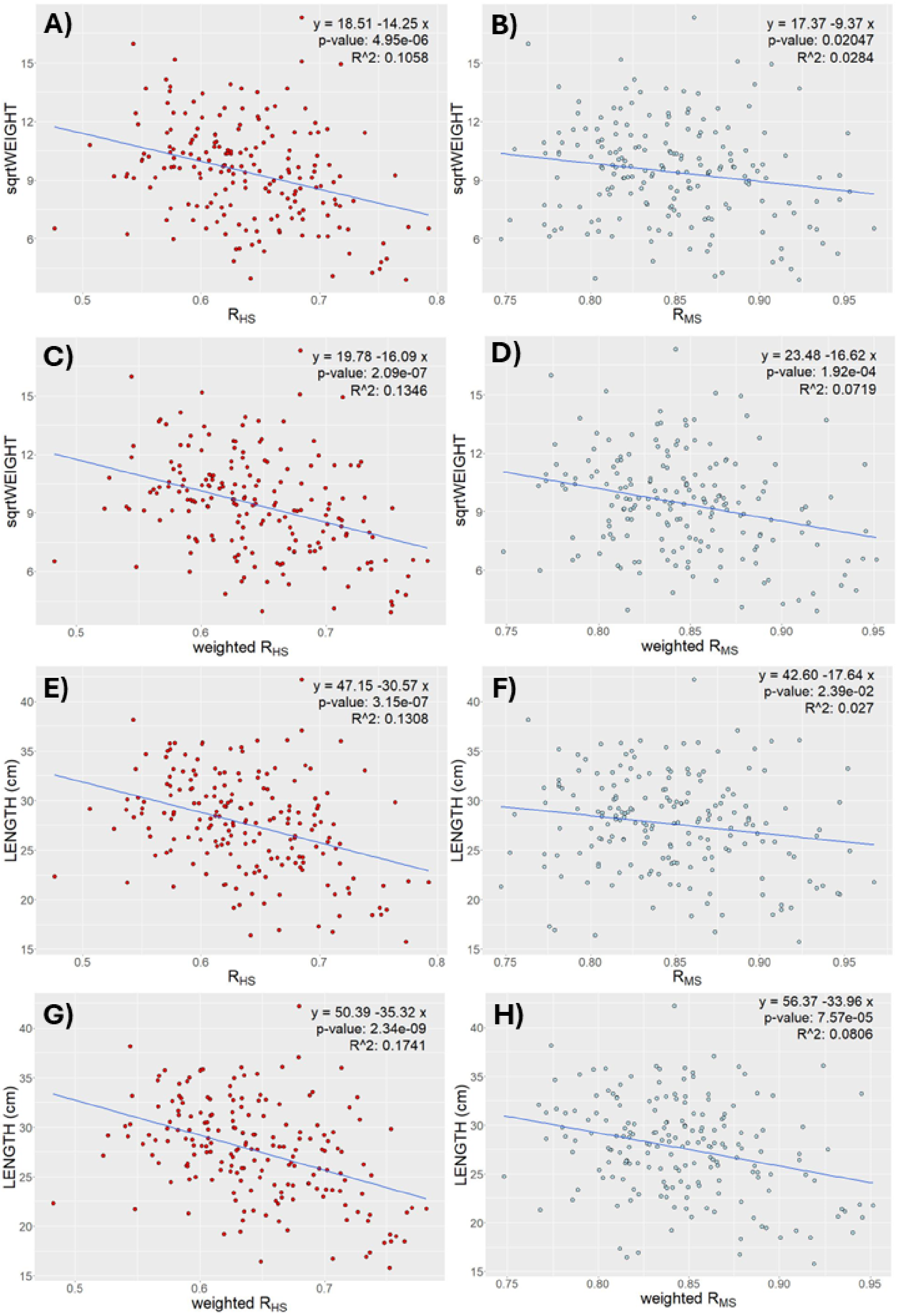


Figure 1. Linear regression models displaying the relationship between genetic load, estimated as $R_{H[M]S}$ (A, B, E, F) and $weighted\ R_{H[M]S}$ (C, D, G, H), and fitness parameters (individual square-root weight (A-D) and individual length (E-H)). The blue line represents the fitted values.

Model name	term	estimate	std.error	t value	p-value	R^2
sqrtWEIGHT ~ R_HS	(Intercept)	18,513	1,940	9,542	7,86E-18	0,106
	R_HS	-14,245	3,029	-4,704	4,95E-06	
sqrtWEIGHT ~ R_MS	(Intercept)	17,369	3,404	5,103	8,17E-07	0,028
	R_MS	-9,374	4,010	-2,338	0,020471	
sqrtWEIGHT ~ weighted R_HS	(Intercept)	19,784	1,929	10,256	7,28E-20	0,135
	weighted R_HS	-16,094	2,985	-5,392	2,09E-07	
sqrtWEIGHT ~ weighted R_MS	(Intercept)	23,483	3,699	6,348	1,61E-09	0,072
	weighted R_MS	-16,625	4,369	-3,805	0,000192	
LENGTH ~ R_HS	(Intercept)	47,154	3,691	12,775	2,84E-27	0,131
	R_HS	-30,570	5,762	-5,306	3,15E-07	
LENGTH ~ R_MS	(Intercept)	42,596	6,573	6,481	7,88E-10	0,027
	R_MS	-17,636	7,744	-2,277	0,023902	
LENGTH ~ weighted R_HS	(Intercept)	50,387	3,637	13,856	1,69E-30	0,174
	weighted R_HS	-35,323	5,627	-6,277	2,34E-09	
LENGTH ~ weighted R_MS	(Intercept)	56,373	7,105	7,935	1,88E-13	0,081
	weighted R_MS	-33,965	8,391	-4,048	7,57E-05	

Table 2. Linear regression model results, showing the correlation between the genetic load estimates and the fitness proxies.

4.2 Effect of the mutation load on the condition

To test the impact of mutation load on survival, we inspected the differences in the mean number of alternative alleles per locus, for each impact class, between Larvae and Juveniles (Figure 2). The comparison was also performed with the weighted mean number of mutated alleles per locus, which takes into account the uncertainty in the allele dosage estimation (Figure 3).

The difference between the genetic load expressed as mean number of mutated alleles between Larvae and Juveniles appears to be significant for High-impact loci (p < 0,05), when comparing the mean values derived from discrete genotypes (Figure 2A). This suggests that individuals who died before the yolk sac absorption carried a higher number of strongly deleterious mutation. This difference does not hold, instead, for moderate-impact (Missense, Figure 2B) and for synonymous substitutions (Figure 2C), in accordance with our working hypothesis. However, the deleterious mutation load does not seem to be significantly higher in Larvae when the comparison is based on the mean number of alternative alleles weighted for the posterior probabilities (Figure 3), in contrast with what we previously observed (Figure 1, Table 2). Nevertheless, the p-value is close to significance only for High-impact mutations (Figure 3A), and the differences between the two groups appear to decrease with the predicted deleteriousness of variants (Figure 3).

The weakness of the signal of mutations' impact on the developmental condition can be attributed to the fact that individuals with the highest load (and a higher number of loci in complete homozygosity) for strongly deleterious variants had possibly died before hatching. Unfortunately, we do not dispose of DNA samples from unhatched eggs, hence this hypothesis cannot be directly tested. However, some additional insights could be gained in future analyses inspecting the correlation between the parental load and data on hatching success.

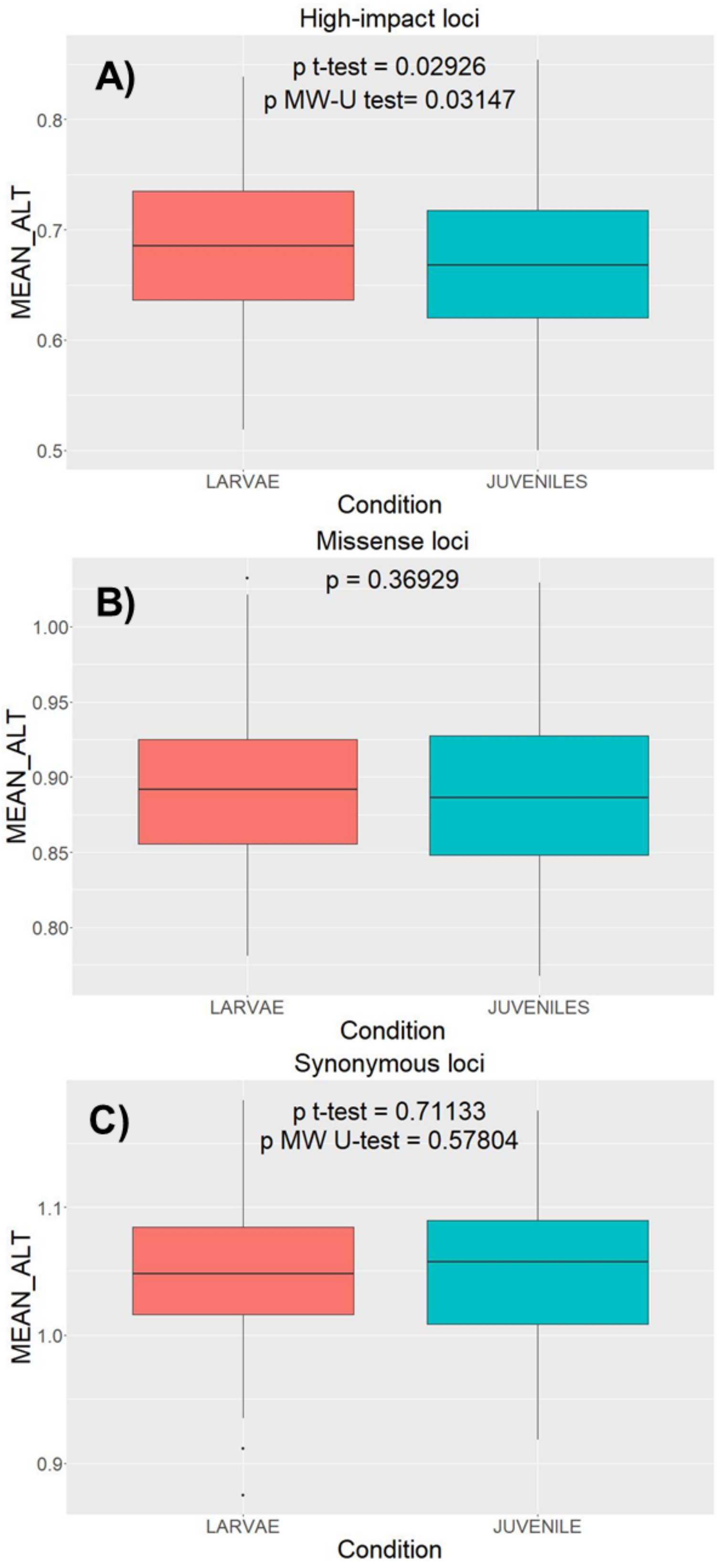


Figure 2. Effect of the mean number of alternative alleles per locus (MEAN_ALT) on the condition (Larva vs Juvenile), for each of the three impact classes: High (A), Missense (B), Synonymous (C).

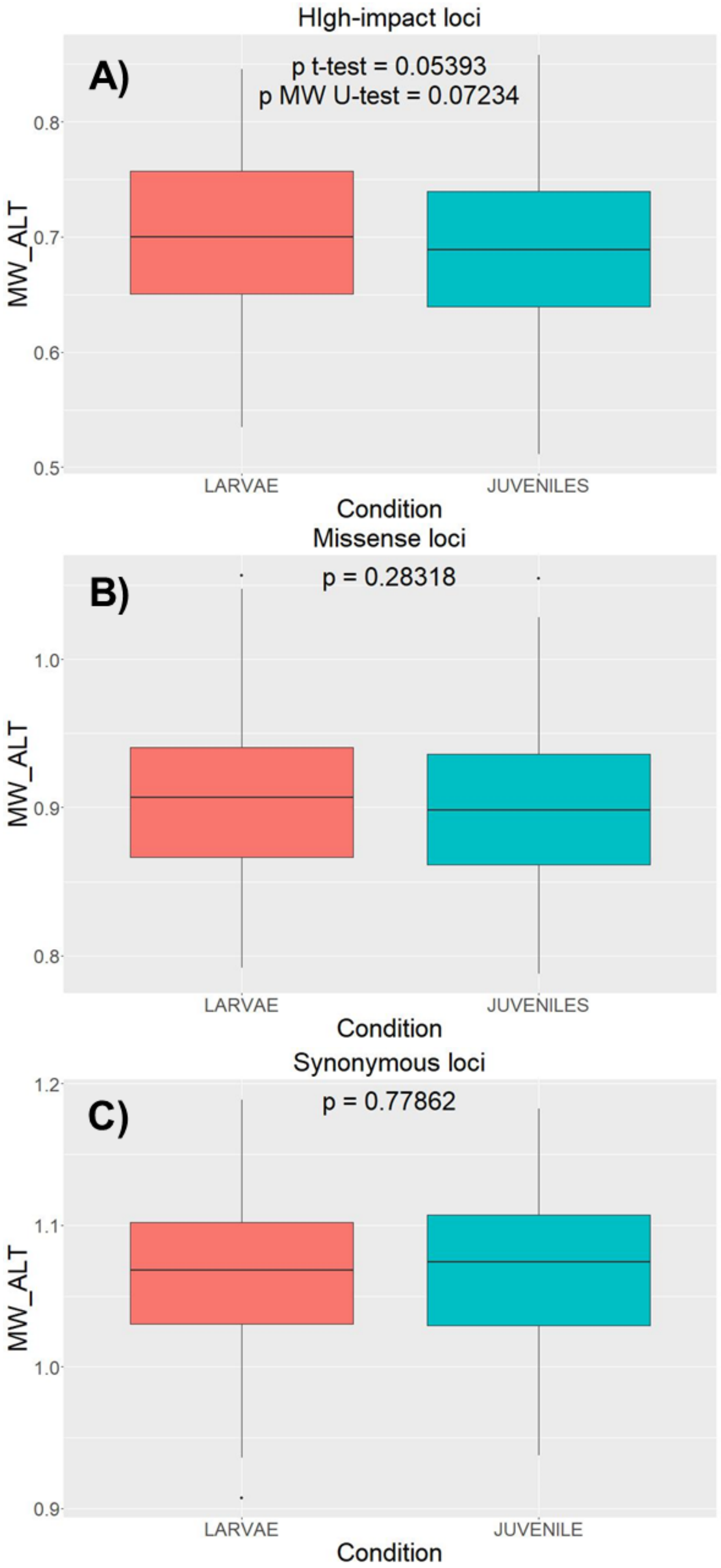


Figure 3. Effect of the mean weighted number of alternative alleles per locus (MW_ALT) on the condition (Larva vs Juvenile), for each of the three impact classes: High (A), Missense (B), Synonymous (C).

MEAN_ALT			Larvae	Juveniles
High	mean		0.6857025	0.6700119
	p-value	Mann Whitney U	0.02999307	
		Student's t	0.02812058	
Missense	mean		0.8947363	0.8894891
	p-value	Mann Whitney U	0.4034761	
		Student's t		
Synonymous	mean		1.048110	1.049274
	p-value	Mann Whitney U	0.6413722	
		Student's t	0.8349719	
MW_ALT			Larvae	Juveniles
High	mean		0.7024806	0.6880219
	p-value	Mann Whitney U	0.07084244	
		Student's t	0.05309853	
Missense	mean		0.9094865	0.9032461
	p-value	Mann Whitney U	0.3084791	
		Student's t		
Synonymous	mean		1.068054	1.067199
	p-value	Mann Whitney U	0.8486746	
		Student's t		

Table 3. Results of the models inspecting the difference in mutation load between Larvae and Juveniles for each impact class of loci.

4.3 Family Effect

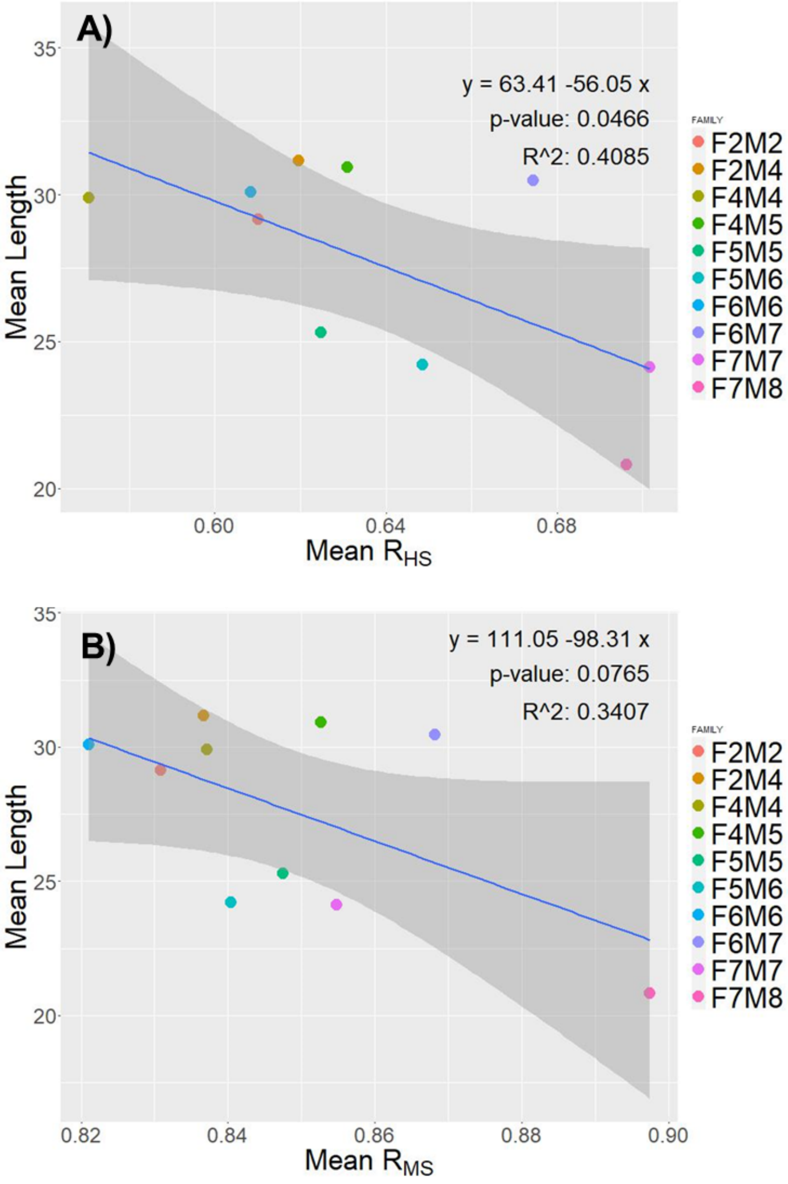
Our results suggested that a considerable amount of variation observed in fitness performances is explained by the estimated deleterious mutation load carried by individuals (Figure 1, Table 2). However, the effect of the family factor needed to be considered, since individuals from the same family do not represent independent observations.

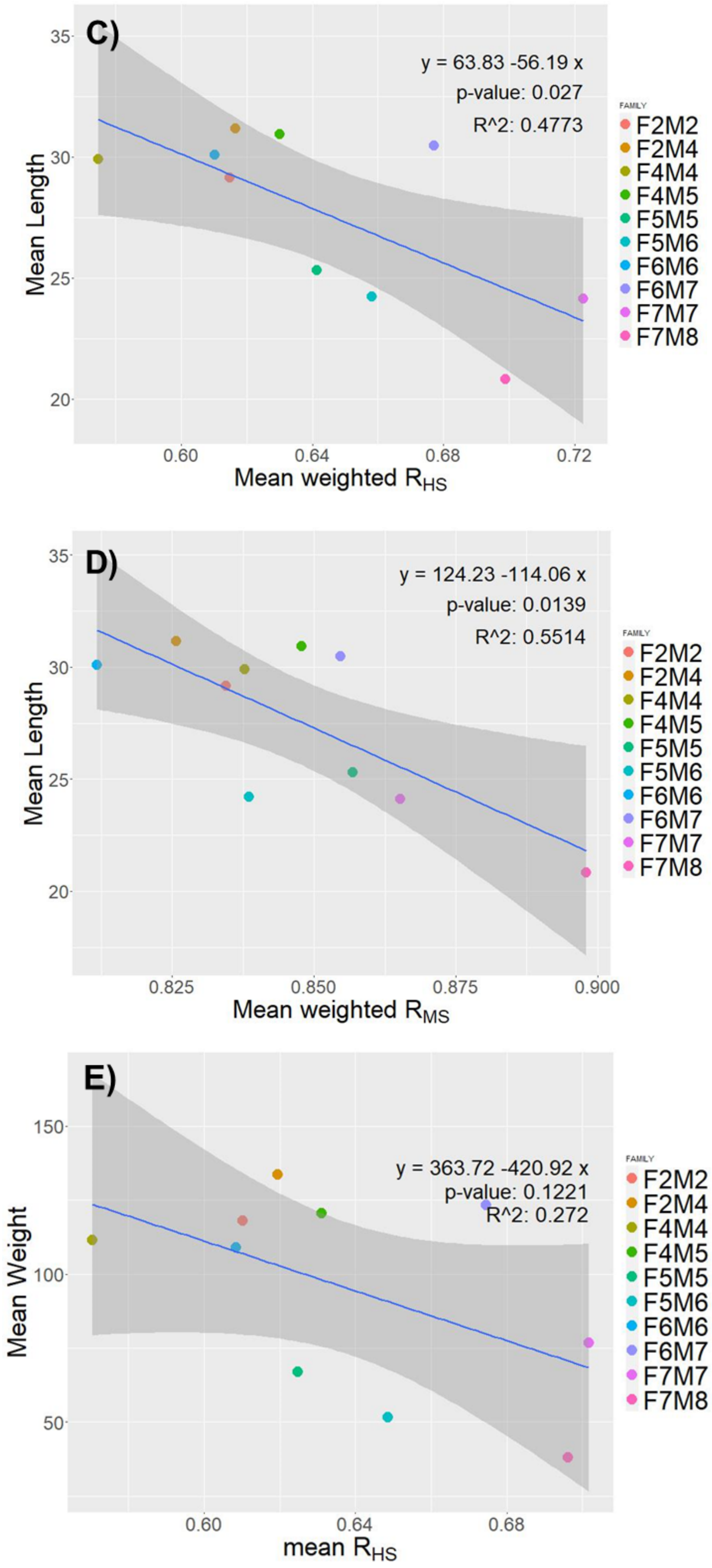
Linear mixed effects models were explored, at first testing the interaction effect between family and genetic load, which was found non-significant. We consequently proceeded inspecting a mixed model including the additive effect of family as a random factor. Results show that, even when taking family into account, an indication of significance of the correlation between deleterious genetic load and fitness exists, although only for models based on $weighted\ R_{HS}$ (Table 4).

Model name	**term**	**estimate**	**std.error**	**t value**	**p-value**
sqrtWEIGHT ~ R_HS + (1 | FAMILY)	(Intercept)	12,983	2,059	6,305	
	R_HS	-5,589	3,105	-1,800	0,072
	sd__(Intercept)	1,728			
	sd__Observation	1,820			
sqrtWEIGHT ~ R_MS + (1 | FAMILY)	(Intercept)	9,867	2,850	3,462	
	R_MS	-0,531	3,289	-0,161	0,872
	sd__(Intercept)	1,838			
	sd__Observation	1,831			
sqrtWEIGHT ~ weighted R_HS + (1 | FAMILY)	(Intercept)	13,743	2,214	6,209	
	weighted R_HS	-6,720	3,332	-2,017	0,044
	sd__(Intercept)	1,685			
	sd__Observation	1,818			
sqrtWEIGHT ~ weighted R_MS + (1 | FAMILY)	(Intercept)	11,403	3,329	3,425	
	weighted R_MS	-2,348	3,875	-0,606	0,545
	sd__(Intercept)	1,809			
	sd__Observation	1,830			
LENGTH ~ R_HS + (1 | FAMILY)	(Intercept)	34,332	3,975	8,638	
	R_HS	-10,492	5,999	-1,749	0,080
	sd__(Intercept)	3,293			
	sd__Observation	3,519			
LENGTH ~ R_MS + (1 | FAMILY)	(Intercept)	26,992	5,505	4,903	
	R_MS	0,763	6,351	0,120	0,904
	sd__(Intercept)	3,563			
	sd__Observation	3,535			
LENGTH ~ weighted R_HS + (1 | FAMILY)	(Intercept)	36,546	4,261	8,578	
	weighted R_HS	-13,837	6,422	-2,155	0,031
	sd__(Intercept)	3,161			
	sd__Observation	3,512			
LENGTH ~ weighted R_MS + (1 | FAMILY)	(Intercept)	30,305	6,433	4,711	
	weighted R_MS	-3,152	7,487	-0,421	0,674
	sd__(Intercept)	3,501			
	sd__Observation	3,536			

Table 4. Linear mixed effects models exploring the correlation between fitness proxies and genetic load at High and Missense loci, with Family as random factor.

Intra-family correlations were also explored but found non-significant in almost every case. In general, there seems to be a considerable variability within families, but observations within each group are insufficient to reveal any pattern (n < 20). For this reason, we summarized information from each family (Figure 4) and found indications of persistence of the previously observed pattern of fitness reduction with increasing genetic load, especially for the Length variable (Figure 4A-D), while the pattern is less clear for the Weight variable (Figure 4E-H).




C)
y = 63.83 -56.19 x
p-value: 0.027
R^2: 0.4773
Mean Length
35
30
25
20
0.60
0.64
0.68
0.72
Mean weighted R_{HS}
FAMILY
F2M2
F2M4
F4M4
F4M5
F5M5
F5M6
F6M6
F6M7
F7M7
F7M8
D)
y = 124.23 -114.06 x
p-value: 0.0139
R^2: 0.5514
Mean Length
35
30
25
20
0.825
0.850
0.875
0.900
Mean weighted R_{MS}
FAMILY
F2M2
F2M4
F4M4
F4M5
F5M5
F5M6
F6M6
F6M7
F7M7
F7M8
E)
y = 363.72 -420.92 x
p-value: 0.1221
R^2: 0.272
Mean Weight
150
100
50
0.60
0.64
0.68
mean R_{HS}
FAMILY
F2M2
F2M4
F4M4
F4M5
F5M5
F5M6
F6M6
F6M7
F7M7
F7M8

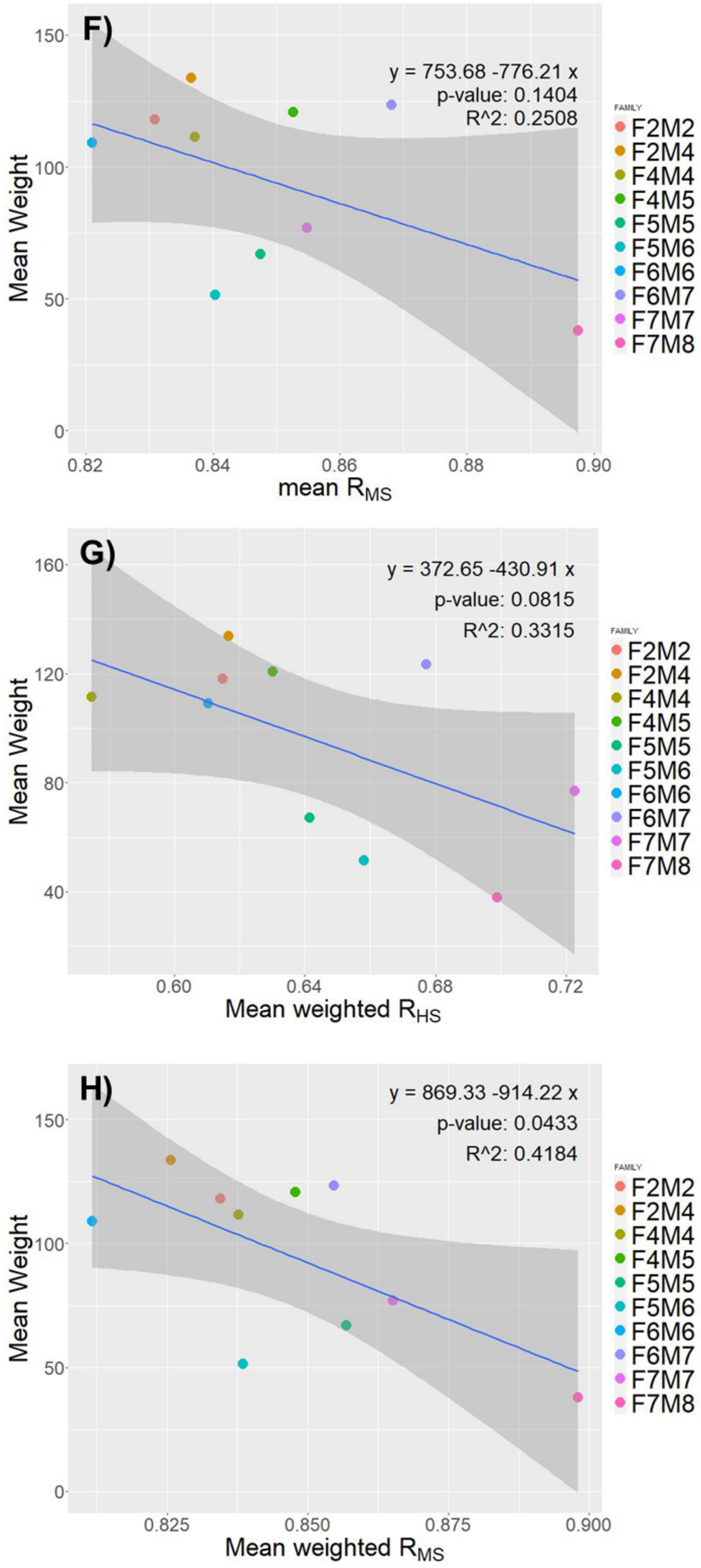

F)
y = 753.68 -776.21 x
p-value: 0.1404
R^2: 0.2508
Mean Weight
150
100
50
0
0.82
0.84
0.86
0.88
0.90
mean R_{MS}
FAMILY
F2M2
F2M4
F4M4
F4M5
F5M5
F5M6
F6M6
F6M7
F7M7
F7M8
G)
y = 372.65 -430.91 x
p-value: 0.0815
R^2: 0.3315
Mean Weight
160
120
80
40
0.60
0.64
0.68
0.72
Mean weighted R_{HS}
FAMILY
F2M2
F2M4
F4M4
F4M5
F5M5
F5M6
F6M6
F6M7
F7M7
F7M8
H)
y = 869.33 -914.22 x
p-value: 0.0433
R^2: 0.4184
Mean Weight
150
100
50
0
0.825
0.850
0.875
0.900
Mean weighted R_{MS}
FAMILY
F2M2
F2M4
F4M4
F4M5
F5M5
F5M6
F6M6
F6M7
F7M7
F7M8

Figure4. Linear models summarizing the variation in fitness estimates with the increasing mutation load among the different families. The blue line represents the fitted values, while the grey band indicates the 95% confidence interval.

Family effect consists in the non-genetic source of variation in the offspring condition, which can be from either parents. In particular, maternal effect could affect egg quality and the early stage of development, through the provisioning of nutrients, hormones and cytoplasm to the egg (Green, 2008). In this case, we cannot exclude some maternal effect, for example in the quality of female 7 (F7); nevertheless, F7M7 and F7M8 families are also the ones who display the highest genetic load at deleterious loci (Figure 4).

Linear mixed effect models were computed also to test the differences in the mutation load between Larvae and Juveniles, with Family as a random factor. When comparing the mean number of alternative alleles per locus, the difference appears significant for High-impact loci (p = 4.692e-05), and not for Missense (p = 0.06427) and Synonymous loci (p = 0.6755) (Figure 5). Consistent results were obtained testing the difference in the mean weighted number of mutated alleles per locus (p = 0.0001775 for High-impact loci; p = 0.02987 for Missense loci; p = 0.9529 for Synonymous loci; results not shown).

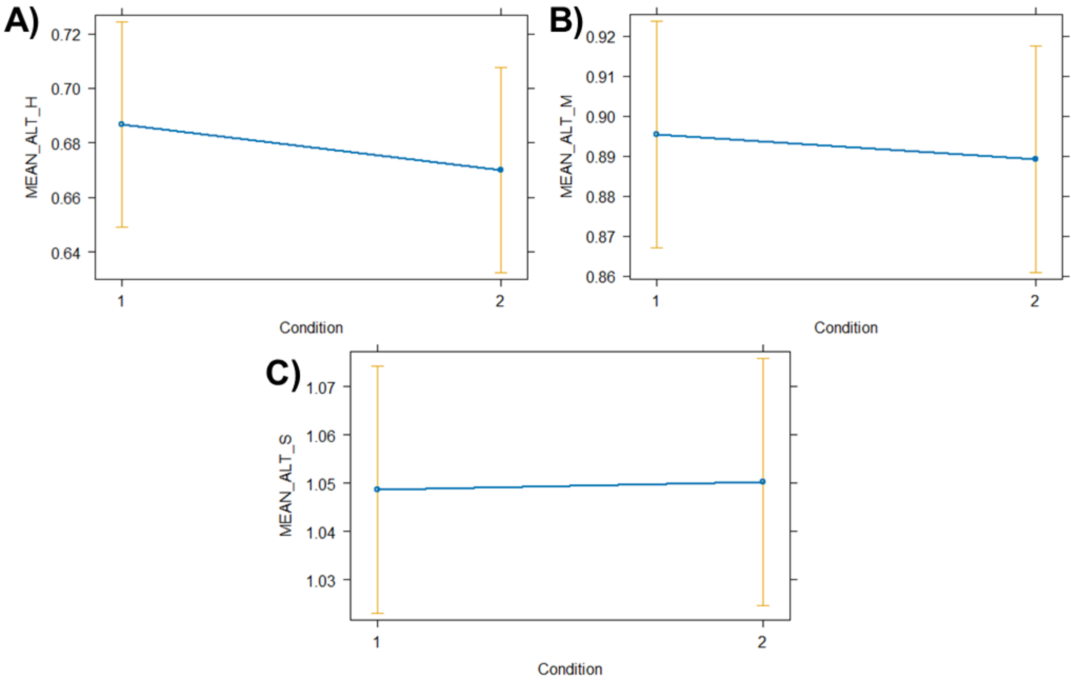


Figure 5. Effect plots for the linear mixed effects models inspecting the difference in the mean number of alternative alleles per locus (MEAN_ALT), for each of the three impact classes: High-impact loci (A), Missense loci (B), Synonymous loci (C). The two levels of the "Condition" variable refer, respectively, to Larvae and Juveniles.

4.4 Considerations on the use of continuous genotypes

In our analysis we compared the explanatory power of genetic load estimates based on discrete genotype calling ($R_{H[M]S}$) with the one of estimates based on continuous genotype calling ($weighted\ R_{H[M]S}$). When inspecting the relationship between individual weight and individual length with genetic load, models including $weighted\ R_{H[M]S}$ as explanatory variable appeared to show a more robust correlation (Figure 1, Table 2). Likewise, when introducing the Family factor, only models based on weighted genetic load proxies show significance for High-impact loci (Table 4).

One of the aspects that makes polyploid genomes more complex than diploid ones is the presence of genotypes with higher allele dosage and the larger number of genotypic classes. Allelic dosage in polyploid organisms is likely to have an impact on phenotype through the additive effect of multiple copies of the same allele and though complex interactions between alleles (de Bem Oliveira et al., 2019). Accordingly, it is crucial to include reliable information on allele dosage.

The assignment of genotypic classes based on NGS data in polyploids is particularly challenging and often suffers from misclassification. The uncertainty in estimating the correct number of allele copies at each locus is further increased by complex inheritance patterns and genome multiplicity (Njuguna et al., 2023). The use of approaches based on continuous genotypes, which reflect the relative probabilities of all possible allele copy numbers, rather than discrete values, can prevent misclassification thus reducing genotype calling errors (Clark et al., 2019). Indeed, this type pf approach have been proved successful in improving accuracy and predictive ability in genomic selection and genome wide association studies, compared to discrete genotype classifications (Njuguna et al., 2023; de Bem Oliveira et al., 2019; Grandke et al., 2016). These findings, along with our results, suggest that, when dealing with higher ploidy levels, taking into account the uncertainty of allele dosage estimation and incorporating the probabilities associated with each possible genotypic class into continuous values, can provide reliable genotype calling and insights on the relationships between genotypic and phenotypic information, even working with low depth of coverage.

4.5 Assumptions and additional remarks

An accurate genotype calling in polyploids needs to take into account the complexity of chromosomes' behaviour during meiosis. Indeed, polyRAD genotype calling tool (Clark et al., 2019) has been designed to model different inheritance modes. In autotetraploids, like the Adriatic sturgeon, the presence of four homologous chromosomes leads to tetrasomic inheritance (meaning that all possible allelic combinations can be observed in gametes); however, different degrees of preferential pairing are possible and tetrasomic inheritance may shift to disomic for some chromosomes (Stift et al., 2008). In the case of the Adriatic sturgeon, mixed inheritance patterns have been revealed, but the majority of markers examined displayed tetrasomic inheritance (Dalle Palle et al., 2022), making it reasonable to assume a tetrasomic segregation pattern in order to simplify the analysis.

A further consideration must be made about the fact that the impact of the mutations here analysed has been inferred from the annotated genome of the closely related *Acipenser ruthenus* (Du et al., 2020). Therefore, we cannot exclude that an analysis based on an annotated genome of *Acipenser naccarii* would lead to slightly different predictions. In this context, more reliable predictions of the effect of the identified mutations will be achieved once the *de novo* assembly of the Adriatic sturgeon genome will be finalized.

An additional layer of complexity when trying to uncover the dynamics that link deleterious variation with fitness impairment in sturgeon species, arise from the intricate nature of the genome of these organisms. Polyploidy and gene duplication are known to upset gene-interaction networks, ultimately altering transcriptional patterns and resulting in some genes' expression levels increase or decrease (Wertheim et al., 2013). In this light, some elucidations might come from transcriptomic analysis.

As previously mentioned, another master's thesis work, besides the present one, was conducted under the EndemixIT project. The work from Bordogna represents a validation of the results presented here, given that similar questions about the relationship between fitness indicators and mutation load estimated were investigated. Differently from what has been done here, Bordogna did not test the Bayesian approach for genotype calling but managed data processed with the widely used Genome Analysis Toolkit (GATK) (McKenna et al., 2010); however, the findings of the two studies were consistent for almost all the issues addressed.

5. Conclusions and future perspectives

This work has been carried out with the aim of increasing the knowledge on the impact of predicted deleterious variants on the viability in non-model, threatened animals. The availability at the ‘Storione Ticino’ aquaculture plant of captive stocks of the critically endangered Adriatic sturgeon, which possess a tetraploid, highly duplicated and complex genome, offered the opportunity to better understand this relationship, which has also practical implications for the ex-situ management of this species.

Our findings support the hypothesis that the global predicted number of deleterious mutations can at least partially explain the difference between individual fitness, and the observed correlation appears to weaken with the putative impact of variants. Detected strongly deleterious mutations (High-impact) seem to significantly impair development, with Juvenile individuals possessing a lower mutation load as compared to individuals who did not survive until the end of the experiment (Larvae). Additionally, the consistent pattern of negative correlation between load estimates and fitness indicators uncovered for Juvenile individuals, suggest that, among them, the best performing ones may carry a diminished load.

We tested the potential of a Bayesian approach for genotype calling to deal with low-depth sequencing data, which allows to infer reliable genotypes while reducing the amount of missing data. This would possibly respond to the need of increasing the number of genotyped individuals in a cost-effective way. Moreover, the use of continuous genotypes, which incorporate the relative probability of every possible allele copy number, has shown to prevent genotype misclassification and increase the predictive power when linking genotypic to phenotypic information.

At last, future analyses can be developed to uncover the impact of specific genetic variants on fitness. This can be accomplished inspecting the differences in the number of alternative alleles at each locus between Larvae and Juveniles. This type of investigation aiming at identifying loci with a significant impact on fitness can then be refined through a comprehensive Gene Ontology (GO) analysis of the identified loci. This approach will provide deeper insights into the biological functions and pathways associated with these loci, offering a more detailed understanding of their potential roles in shaping fitness-related traits.

In conclusion, as some authors suggest (Kyriazis et al., 2020), when populations are destined to remain small and highly inbred, conservation efforts should concentrate on the minimization of strongly deleterious variation. On this basis, we emphasize the relevance of improving our predictive capacity of the effect of identified deleterious mutations. This is crucial especially for species which persistence strongly depends on ex-situ management, since it would help preventing gene pool deterioration and the reintroduction of unfavourable mutations.

Acknowledgements

I want to thank all the people that made the completion of this work possible with the support they provided me throughout the whole internship period.

First, my supervisor Leonardo Congiu, who allowed me to take part to the project. Then, my tutor Stefano Dalle Palle, who transmitted to me the knowledge I needed to initiate and carry on the project and supported me while conducting the work. I also want to thank my co-tutor Giorgio Bertorelle and the whole genomics equipe from the University of Ferrara for their advice and support in the analyses: Maëva Gabrielli, Andrea Benazzo, Victor Muñoz, and Silvia Fuselli.

Finally, I thank Federica Bordogna, who carried out this project at my side, and helped me whenever I was in need.

References

Agrawal, A. F., & Whitlock, M. C. (2012). Mutation load: The fitness of individuals in populations where deleterious alleles are abundant. *Annual Review of Ecology, Evolution, and Systematics*, *43*, 115–135. https://doi.org/10.1146/annurev-ecolsys-110411-160257

Allendorf, F. W., Luikart, G. H., & Aitken, S. N. (2012). *Conservation and the genetics of populations.* John Wiley & Sons.

Arlati, G., & Poliakova, L. (2009). Restoration of Adriatic sturgeon (Acipenser naccarii) in Italy: situation and perspectives. *Biology, Conservation and Sustainable Development of Sturgeons*, 237–245.

Barnosky, Anthony, D., & Al., E. (2011). Has the Earth's sixth mass extinction already arrived?. *Nature*, *471*(7336), 51–57.

Bastien, M., Boudhrioua, C., Fortin, G., & Belzile, F. (2018). Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. *Genome*, *61*(6), 449–456. https://doi.org/10.1139/gen-2017-0236

Bataillon, T., & Bailey, S. F. (2014). *Effects of new mutations on fitness : insights from models and data*. *1320*, 76–92. https://doi.org/10.1111/nyas.12460

Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H. E., & van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, *23*(8), 492–503. https://doi.org/10.1038/s41576-022-00448-x

Boscari, E., & Congiu, L. (2014). The need for genetic support in restocking activities and ex situ conservation programmes: The case of the Adriatic sturgeon (Acipenser naccarii Bonaparte, 1836) in the Ticino River Park. *Journal of Applied Ichthyology*, *30*(6), 1416–1422. https://doi.org/10.1111/jai.12545

Boscari, E., Vidotto, M., Martini, D., Papetti, C., Ogden, R., & Congiu, L. (2015). Microsatellites from the genome and the transcriptome of the tetraploid adriatic sturgeon, acipenser naccarii (Bonaparte, 1836) and cross-species applicability to the diploid beluga sturgeon, huso huso (linnaeus, 1758). *Journal of Applied Ichthyology*, *31*(6), 977–983. https://doi.org/10.1111/jai.12906

Caballero, A., Bravo, I., & Wang, J. (2017). Inbreeding load and purging: implications for the short-term survival and the conservation management of small populations. *Heredity*, *118*(2), 177–185. https://doi.org/10.1038/hdy.2016.80

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Clark, L. V., Lipka, A. E., & Sacks, E. J. (2019). polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3: Genes, Genomes, Genetics*, *9*(3), 663–673. https://doi.org/10.1534/g3.118.200913

Clo, J., Kolá, F., & Clo, J. (2022). *Inbreeding depression in polyploid species : a meta-analysis*.

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

Congiu, L., Boscari, E., Pagani, S., Gazzola, M., & Bronzi, P. (2021). Resumption of natural reproduction of the Adriatic sturgeon in the River Po. *Oryx*, *55*(6), 816. https://doi.org/DOI: 10.1017/S0030605321001150

Congiu, L., Gessner, J., & Ludwig, A. (2023). IUCN Red List reassessment reveals further decline of sturgeons and paddlefishes. *Oryx*, *57*(1), 9–10. https://doi.org/10.1017/S0030605322001260

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7), 901–913.

Dalle Palle, S., Boscari, E., Bordignon, S. G., Muñoz-Mora, V. H., Bertorelle, G., & Congiu, L. (2022). Different Chromosome Segregation Patterns Coexist in the Tetraploid Adriatic Sturgeon Acipenser naccarii. *Diversity*, *14*(9). https://doi.org/10.3390/d14090745

de Bem Oliveira, I., Amadeu, R. R., Ferrão, L. F. V., & Muñoz, P. R. (2020). Optimizing whole-genomic prediction for autotetraploid blueberry breeding. *Heredity*, *125*(6), 437–448. https://doi.org/10.1038/s41437-020-00357-x

de Bem Oliveira, I., Resende, M. F. R., Ferrão, L. F. V., Amadeu, R. R., Endelman, J. B., Kirst, M., Coelho, A. S. G., & Munoz, P. R. (2019). Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3: Genes, Genomes, Genetics*, *9*(4), 1189–1198. https://doi.org/10.1534/g3.119.400059

Delage, N., Couturier, B., Jatteau, P., Larcher, T., Ledevin, M., Goubin, H., Cachot, J., & Rochard, E. (2020). Oxythermal window drastically constraints the survival and development of European sturgeon early life phases. *Environmental Science and Pollution Research*, *27*, 3651–3660.

Doekes, H. P., Bijma, P., & Windig, J. J. (2021). How depressing is inbreeding? A meta-analysis of 30 years of research on the effects of inbreeding in livestock. *Genes*, *12*(6). https://doi.org/10.3390/genes12060926

Douglas, B., Maechler, M., Ben, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Du, K., Stöck, M., Kneitz, S., Klopp, C., Woltering, J. M., Adolfi, M. C., Feron, R., Prokopov, D., Makunin, A., Kichigin, I., Schmidt, C., Fischer, P., Kuhl, H., Wuertz, S., Gessner, J., Kloas, W., Cabau, C., Iampietro, C., Parrinello, H., … Journot, L. (2020). Mechanisms of segmental rediploidization. *Nature Ecology & Evolution*, *4*(June). https://doi.org/10.1038/s41559-020-1166-x

Dussex, N., Morales, H. E., Grossen, C., Dalén, L., & van Oosterhout, C. (2023). Purging and accumulation of genetic load in conservation. *Trends in Ecology and Evolution*, *38*(10), 961–969. https://doi.org/10.1016/j.tree.2023.05.008

Fontana, F., Congiu, L., Mudrak, V. A., Quattro, J. M., Smith, T. I., Ware, K., &

Doroshov, S. I. (2008). Evidence of hexaploid karyotype in shortnose sturgeon. *Genome*, *51*(2), 113–119.

Fontana, F., Lanfredi, M., Chicca, M., Congiu, L., Tagliavini, J., & Rossi, R. (1999). Fluorescent in Situ Hybridization with rDNA Probes on Chromosomes of Acipenser ruthenus and Acipenser naccarii (Osteichthyes, Acipenseriformes). *Genome*, *42*(5), 1008–1012.

Fontana, F., Zane, L., Pepe, A., & Congiu, L. (2007). Polyploidy in Acipenseriformes: cytogenetic and molecular approaches. *Fish Cytogenetics*, *385*(403).

Fox, C. W., Scheibly, K. L., & Reed, D. H. (2008). Experimental evolution of the genetic load and its implications for the genetic basis of inbreeding depression. *Evolution*, *62*(9), 2236–2249.

Fox, D. T., Soltis, D. E., Soltis, P. S., Ashman, T. L., & Van de Peer, Y. (2020). Polyploidy: A Biological Force From Cells to Ecosystems. *Trends in Cell Biology*, *30*(9), 688–694. https://doi.org/10.1016/j.tcb.2020.06.006

Frankham, R. (2005). *Genetics and extinction*. *126*, 131–140. https://doi.org/10.1016/j.biocon.2005.05.002

Gardiner, B. G. (1984). Sturgeons as Living Fossils. In N. Eldredge & S. M. (Eds. . Stanley (Eds.), *Living Fossils* (pp. 148–152). Springer-Verlag.

Gilpin, M. E., & Soulé, M. E. (1986). Minimum Viable Populations: Processes of Species Extinction. In M. E. Soulé (Ed.), *Conservation Biology: The Science of Scarcity and Diversity* (pp. 19–34). Sinauer.

Grandke, F., Singh, P., Heuven, H. C. M., de Haan, J. R., & Metzler, D. (2016). Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: A comparative study in hexaploid chrysanthemum. *BMC Genomics*, *17*(1), 1–9. https://doi.org/10.1186/s12864-016-2926-5

Green, B. S. A. in marine biology. (2008). Maternal effects in fish populations. *Advances in Marine Biology*, *54*, 1–105.

International Union for Conservation of Nature (IUCN). (2023). *IUCN Red List of Threatened Species. Version 2023-3.* https://www.iucnredlist.org.

IPCC. (2023). Section 4: Near-Term Responses in a Changing Climate. *Climate Change 2023: Synthesis Report*, 42–66. https://doi.org/10.59327/IPCC/AR6-9789291691647

Kimura, M. (1977). Preponderance ofsynonymous changes as evidence for the neutral theory ofmolecular evolution. *Nature*, *267*(5608), 275–762.

Kono, T. J. Y., & Al., E. (2018). Comparative genomics approaches accurately predict deleterious variants in plants. *G3*, *8*(10), 3321–3329.

Krieger, J., & Fuerst, P. A. (2002). *Evidence for a Slowed Rate of Molecular Evolution in the Order Acipenseriformes*. 891–897.

Kyriazis, C. C., Wayne, R. K., & Lohmueller, K. E. (2020). *Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression*. 33–47. https://doi.org/10.1002/evl3.209

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-

Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lohe, A. (2021). *Adriatic Sturgeon (Acipenser naccarii), European sturgeon (Acipenser sturio), Chinese sturgeon (Acipenser sinensis), Sakhalin sturgeon (Acipenser mikadoi), Kaluga sturgeon (Huso dauricus) 5-Year Review: Summary and Evaluation*. 46.

Lynch, M., Conery, I. J., & Burger, R. (1995). *Mutation accumulation and the extinction of small populations*. *146*(4), 489–518.

Matias, F. I., Xavier Meireles, K. G., Nagamatsu, S. T., Lima Barrios, S. C., Borges do Valle, C., Carazzolle, M. F., Fritsche-Neto, R., & Endelman, J. B. (2019). Expected Genotype Quality and Diploidized Marker Data from Genotyping-by-Sequencing of Urochloa spp. Tetraploids . *The Plant Genome*, *12*(3), 1–9. https://doi.org/10.3835/plantgenome2019.01.0002

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

Nietlisbach, P., Muff, S., Reid, J. M., Whitlock, M. C., & Keller, L. F. (2019). Nonequivalent lethal equivalents: Models and inbreeding metrics for unbiased estimation of inbreeding load. *Evolutionary Applications*, *12*(2), 266–279. https://doi.org/10.1111/eva.12713

Njuguna, J. N., Clark, L. V., Lipka, A. E., Anzoua, K. G., Bagmet, L., Chebukin, P., Dwiyanti, M. S., Dzyubenko, E., Dzyubenko, N., Ghimire, B. K., Jin, X., Johnson, D. A., Kjeldsen, J. B., Nagano, H., de Bem Oliveira, I., Peng, J., Petersen, K. K., Sabitov, A., Seong, E. S., … Sacks, E. J. (2023). Impact of genotype-calling methodologies on genome-wide association and genomic prediction in polyploids. *Plant Genome*, *16*(4), 1–17. https://doi.org/10.1002/tpg2.20401

Otto, S. P. (2007). The Evolutionary Consequences of Polyploidy. *Cell*, *131*(3), 452–462. https://doi.org/10.1016/j.cell.2007.10.022

Peng, Z., Ludwig, A., Wang, D., Diogo, R., Wei, Q., & He, S. (2007). *Age and biogeography of major clades in sturgeons and paddle W shes (Pisces : Acipenseriformes)*. *42*, 854–862. https://doi.org/10.1016/j.ympev.2006.09.008

Ryan, P., Valentin, R.-R., Mark, A. D., Tim, J. F., Mauricio, O. C., Geraldine, A. V. der A., Davird, E. K., Laura, D. G., Ami, L.-M., David, R., Khalid, S., Joel, T., Sheila, C., Chris, W., Monkol, L., Stacey, G., Mark, J. D., Ben, N., Daniel, G. M., & Eric, B. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxIV*, 1–22.

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergstro, L., Meester, L. De, Bruford, M. W., Bra, I., Colling, G., Dale, L., Ekblom, R., Fawcett, K. D., Fior, S., Hajibabaei, M., Hill, J. A., Hoezel, A. R., Ho, J., Jensen, E. L., Norman, A. J., Ogden, R., … Zielin, P. (2015). *Genomics and the challenging translation into conservation practice*. *30*(2). https://doi.org/10.1016/j.tree.2014.11.009

Speak, S. A., Birley, T., Bortoluzzi, C., Clark, M. D., Percival-Alwyn, L., Morales, H. E., & van Oosterhout, C. (2024). Genomics-informed captive breeding can reduce inbreeding depression and the genetic load in zoo populations. *Molecular Ecology Resources*, *April*, 1–10. https://doi.org/10.1111/1755-0998.13967

Stift, M., Berenos, C., Kuperus, P., & Van Tienderen, P. H. (2008). Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: A general procedure applied to rorippa (yellow cress) microsatellite data. *Genetics*, *179*(4), 2113–2123. https://doi.org/10.1534/genetics.107.085027

Stoffel, M. A., Johnston, S. E., Pilkington, J. G., & Pemberton, J. M. (2021). Mutation load decreases with haplotype age in wild Soay sheep. *Evolution Letters*, *5*(3), 187–195. https://doi.org/10.1002/evl3.229

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biology*, *19*(1), 1–12. https://doi.org/10.1186/s13059-018-1520-3

Uitdewilligen, J. G. A. M. L., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J. A., Visser, R. G. F., & van Eck, H. J. (2013). A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. *PLOS ONE*, *8*(5), e62355. https://doi.org/10.1371/journal.pone.0062355

Van Oosterhout, C. (2020). Mutation load is the spectre of species. *Nature Ecology & Evolution*, 16–18. https://doi.org/10.1038/s41559-020-1204-8

Van Oosterhout, C., Smith, A. M., Hänfling, B., Ramnarine, I. W., Mohammed, R. S., & Cable, J. (2007). The guppy as a conservation model: Implications of parasitism and inbreeding for reintroduction success. *Conservation Biology*, *21*(6), 1573–1583. https://doi.org/10.1111/j.1523-1739.2007.00809.x

Wertheim, B., Beukeboom, L. W., & Van De Zande, L. (2013). Polyploidy in animals: Effects of gene expression on sex determination, evolution and ecology. *Cytogenetic and Genome Research*, *140*(2–4), 256–269. https://doi.org/10.1159/000351998

Williot, P., Arlati, G., Chebanov, M., Gulyas, T., Kasimov, R., Kirschbaum, F., Patriche, N., PAVLOVSKAYA8, L. P., & LUDMILLA POLIAKOVA2, MOHAMMAD POURKAZEMI9, YULYIA. KIM10, P. Z. and I. M. Z. (2002). *Status and Management of Eurasian Sturgeon : An Overview*. 483–506.

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, *2*(3), 7–10. https://cran.r-project.org/doc/Rnews/

Appendix

		MinLikeRatio = 1		MinLikeRatio = 2 NoHighDP DP > 5		MinLikeRatio = 10 NoHighDP DP > 5		MinLikeRatio = 10 NoHighDP DP > 10	
		N°	fraction	N°	fraction	N°	fraction	N°	fraction
tot	GT0	123629	1,00	110929	0,90	110360	0,89	106418	0,86
	GT1	104408	1,00	97992	0,94	95871	0,92	94699	0,91
	GT2	48350	1,00	45602	0,94	44264	0,92	43420	0,90
	GT3	10147	1,00	9801	0,97	9642	0,95	9405	0,93
	GT4	698	1,00	651	0,93	646	0,93	569	0,82
High	GT0	57379	1,00	52240	0,91	51990	0,91	50350	0,88
	GT1	4272	1,00	40896	0,96	40028	0,94	39643	0,93
	GT2	13141	1,00	12755	0,97	12320	0,94	12204	0,93
	GT3	2275	1,00	2230	0,98	2216	0,97	2186	0,96
	GT4	68	1,00	64	0,94	64	0,94	59	0,87
Missense	GT0	48315	1,00	42796	0,89	42601	0,88	40944	0,85
	GT1	42963	1,00	39739	0,92	38945	0,91	38362	0,89
	GT2	22261	1,00	20473	0,92	19845	0,89	19360	0,87
	GT3	4990	1,00	4794	0,96	4739	0,95	4606	0,92
	GT4	511	1,00	472	0,92	468	0,92	401	0,78
Synonymous	GT0	17935	1,00	15893	0,89	15769	0,88	15124	0,84
	GT1	18724	1,00	17357	0,93	16898	0,90	16694	0,89
	GT2	12948	1,00	12374	0,96	12099	0,93	11856	0,92
	GT3	2882	1,00	2777	0,96	2687	0,93	2613	0,91
	GT4	119	1,00	115	0,97	114	0,96	109	0,92

Table S1. Genotype count for each impact category (High, Missense, Synonymous), over all samples, for each filter imposed. GT0-4 refers to the 5 genotypic classes (respectively: *nulliplex*, *simplex*, *duplex*, *triplex* and *quadruplex*). The number of genotypes for each impact class (N°) is reported, as well as the fraction of genotypes for each genotypic class maintained from the filter (relative to the less conservative filter).

		MinLikeRatio = 1		MinLikeRatio = 2 NoHighDP DP > 5		MinLikeRatio = 10 NoHighDP DP > 5		MinLikeRatio = 10 NoHighDP DP > 10	
		N°	fraction	N°	fraction	N°	fraction	N°	fraction
tot	**tot**	287232	1	264975	1	260783	1	254511	1
	GT0	123629	0,4304	110929	0,4186	110360	0,4232	106418	0,4181
	GT1	104408	0,3635	97992	0,3698	95871	0,3676	94699	0,3721
	GT2	48350	0,1683	45602	0,1721	44264	0,1697	43420	0,1706
	GT3	10147	0,0353	9801	0,037	9642	0,037	9405	0,037
	GT4	698	0,0024	651	0,0025	646	0,0025	569	0,0022
High	**tot**	115584	1	108185	1	106618	1	104442	1
	GT0	57379	0,4964	52240	0,4829	51990	0,4876	50350	0,4821
	GT1	42721	0,3696	40896	0,378	40028	0,3754	39643	0,3796
	GT2	13141	0,1137	12755	0,1179	12320	0,1156	12204	0,1168
	GT3	2275	0,0197	2230	0,0206	2216	0,0208	2186	0,0209
	GT4	68	0,0006	64	0,0006	64	0,0006	59	0,0006
Missense	**tot**	119040	1	108274	1	106598	1	103673	1
	GT0	48315	0,4059	42796	0,3953	42601	0,3996	40944	0,3949
	GT1	42963	0,3609	39739	0,367	38945	0,3653	38362	0,37
	GT2	22261	0,187	20473	0,1891	19845	0,1862	19360	0,1867
	GT3	4990	0,0419	4794	0,0443	4739	0,0445	4606	0,0444
	GT4	511	0,0043	472	0,0044	468	0,0044	401	0,0039
Synonymous	**tot**	52608	1	48516	1	47567	1	46396	1
	GT0	17935	0,3409	15893	0,3276	15769	0,3315	15124	0,326
	GT1	18724	0,3559	17357	0,3578	16898	0,3552	16694	0,3598
	GT2	12948	0,2461	12374	0,255	12099	0,2544	11856	0,2555
	GT3	2882	0,0548	2777	0,0572	2687	0,0565	2613	0,0563
	GT4	119	0,0023	115	0,0024	114	0,0024	109	0,0023

Table S2. Genotype count for each impact category (High, Missense, Synonymous), over all samples, for each filter imposed. For each impact category, the distribution of the inferred genotypes among the different genotypic classes (GT0-4) is reported. The fraction is relative to the total number of genotypes inferred for every impact category and has been computed to reveal the presence of biases in the filter imposed towards any of the genotypic classes.