Universitat Autònoma de Barcelona

Detecting signals of polygenic variability in domestication and in breeding

Ioanna-Theoni Vourlaki

DOCTORAL THESIS

Detecting signals of polygenic variability in domestication and in breeding

Author

Ioanna-Theoni Vourlaki

Supervisors

Dr. Sebastián E. Ramos-Onsins
Dr. Miguel Pérez-Enciso

Tutor

Dr. Mario Cáceres Aguilar

Universitat Autònoma de Barcelona

Departament de Genética I de Microbiologia
Facultat de Biociéncies
Universitat Autónoma de Barcelona

2022

CRAG CENTRE DE RECERCA EN AGRIGENÒMICA

MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD

This PhD thesis has been funded by the grant from the Ministry of Economy and Science (MINECO, Spain) by the MINECO grant BES-2017-08 11 139 funded by MCIN/AEI/10.12039/501100011033 and by “ESF Investing in your future”.

Γηράσκω αεί διδασκόμενος

Σόλων

Contents

List of Figures

List of Tables

Acronyms

MKT	McDonald and Kreitman Test
MKTa	Asymptotic MKT
SFS	Site Frequency Spectrum
uSFS	Unfolded Site Frequency Spectrum
ML	Maximum Likelihood
ML	Machine Learning
LD	Linkage Disequilibrium
PRF	Poisson Random Field
MA	Mutation Accumulation
DFE	Distribution of Fitness Effect
BGS	Background Selection
ABC	Approximate Bayesian Computation
α	Proportion of adaptive substitutions
s	Selection coefficient
S	Scaled selection coefficient
LRTs	Likelihood ration tests
GP	Genomic Prediction
GS	Genomic Selection
TEs	Transposable Elements
TIPs	Transposable Insertion Polymorphisms
SVs	Structural Variations
SNPs	Single Nucleotide Polymorphisms
IBS	Identical by State
IBD	Identical by Descent
QTL	Quantitative Trait Loci
DL	Deep Learning
ANN	Artificial Neural network
MLP	Multilayer Perceptron
CNN	Convolutional Neural Network
GREML	Restricted Maximum Likelihood Estimation
SSVS	Stochastic Search Variable Selection
SVM:	Support Vector Machine
RF	Random Forest
LS	Least-square estimation
GBLUP	Genomic Best Linear Unbiased Predictor
GRM	Genomic Relationship Matrix
1D	One Dimension

2D	Two Dimensions
LRTs	Likelihood Ratio Tests
AIC	Akaike Information Criterion
RKHS	Reproducible Kernel Hilbert Space
BRR	Bayesian Ridge Regression
RR	Ridge Regression
EN	Elastic Net
GWAS	Genome Wide Association Studies
IND	Indica
ARO	Aromatic
ADM	Admixed
JAP	Japonica
AUS	Aus/Boro
INS	Insertions
DEL	Deletions
DUP	Tandem Duplications
CNVs	Copy Number Variants
INV	Inversions
MITEs	Miniature Inverted-Repeat Transposable Elements
DTX	DNA TEs with terminal inverted repeats
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
SINE	Short Interspersed Nuclear Element
RIX	Non-LTR Retrotransposons
RLX	LTR Retrotransposons
PCA	Principal Component Analysis
PCs	Principal Components

Abstract

Most complex traits of interest are controlled by many genes of small effects which experience only subtle changes in their frequency, making it hard to detect a specific derived pattern in the genome. Therefore, the genetic architecture underlying the phenotypic variation of most complex traits is still to be revealed. This thesis aims to understand polygenic effects from a population genetics (inference) and a quantitative genetics (prediction) perspective. We reason that observing how patterns of variability are formed under different selective and demographic conditions, such as domestication, may reveal patterns of polygenic adaptation signals in the genome of species. In addition, association between phenotypic traits and causative variants should not be restricted to Single Nucleotide Polymorphisms (SNPs). Transposable Insertion polymorphisms (TIPs) and Structural Variations (SVs) could also explain an important fraction of the variability.

Firstly, the thesis focuses on detecting a genome-wide polygenic signal of domestication process through the analysis of full Distribution of Fitness Effects (DFE). We study the joint DFE using the 2-dimensional site frequency spectrum (2D-SFS) between populations in two ways: (i) we describe and compare the patterns of genetic diversity between the wild and domestic populations under ten domestication scenarios derived from forward simulations, and (ii) we propose a new joint DFE model designed to quantify a signal of domestication. We successfully retrieved this signal in the presence of shared polymorphisms. Finally, we highlight the strengths and limitations of current population genetic models in detecting a polygenic signal of domestication under different genetic and demographic architectures.

Secondly, we investigate whether TIPs can increase the effectiveness of Genomic Prediction (GP) of traits when compared to using only SNPs. We used eleven traits of agronomic importance originated by five different rice population groups (Aus/Boro, Indica, Aromatic, Japonica and Admixed), 738 accessions in total. In a within group scenario, we predicted performance of improved Indica varieties using the rest varieties. In an across group scenario, all Aromatic and Admixed accessions were predicted using the rest of populations. Our analysis showed that TIPs can explain an important fraction of total genetic variance and also improve the genomic prediction of complex traits.

The third purpose of this thesis is to add SVs to explore its capacity to predict complex agronomic traits in rice. SVs such as deletions, inversions and duplication can be found in a high proportion in the plant genomes. As in TIPs, we found that SVs can explain an important fraction of genetic variation in the traits of interest. Also, our results suggested that Deep Learning (DL) models outperform in 50% of the studied cases. Finally, DL seems to improve prediction ability of continuous traits compared to Bayesian models when training and test dataset are distantly related.

Chapter 1

Introduction

1.1 Natural Selection

Evolution means that species change over time. Natural selection is the mechanism that can describe the way species change. The theory of evolution through natural selection was the first scientific theory that put together evidence of change through time as well as mechanism for how it happens. The idea that traits are inherited and passed down from parents to offspring has been around since the ancient Greek philosopher's time. In the middle 1700s Carolus Linnaeus suggested a taxonomic naming system which grouped like species together implying that there was an evolutionary connection between species within the same group. In the late 1700s, the first theories that species changed over time, arise. Theories proposed by scientists like the Compte de Buffon and Erasmus Darwin, suggested that species changed over time but neither man could explain how or why the changed.

Natural selection or else "survival of the fittest" as is called was suggested by Charles Darwin in his book "*On the origin of the species"*. Darwin suggested that individuals with the traits most suitable to their environments lived long enough to reproduce and passed down those desirable traits to their offspring. If an individual had less than favorable traits, they would die and not pass on those traits. Over the time, only the "fittest" of the individuals survived. Eventually, after enough time passed, these small adaptations would add up to create new species. Due to limited environmental resources, not all organisms survive. Over time, the population has adapted to its environment through the process of natural selection boosting the most favorable traits in the population. According to Charles Darwin, all species descended from only a few lifeforms that had been modified over time. This descent with modification as he called it forms the backbone of his Theory of Evolution which points out how new certain species evolve.

1.1.1 From natural to artificial selection

Darwin considered the process by which animals and plants are domesticated (artificial selection) as a useful analogy for the mechanism by which populations are adapted in the wild (natural selection). Domestication occupies the introductory chapter of "*On the Origin of Species"* (1859) but also it is thoroughly analyzed in his book "*The Variation of Animals and Plants under Domestication"* published in 1868. In the latter book, Darwin considered two types of artificial selection, in addition to natural selection, the methodical and unconscious selection. All three types of selection share a mechanism of non-random difference in the reproductive success among individuals on the basis of heritable traits. The difference among the three processes is the reason why some individuals will reproduce while others not (Gregory 2009). While natural selection is not controlled by humans and is affected by the natural environment, in unconscious artificial selection the humans may choose which individuals will contribute more to the next generation but without necessarily account for a long-term effect. Instead, in methodical selection, humans select individuals

for breeding in order to maintain and enhance the traits of interest. Darwin himself stated that his discovery of natural selection came through his studies of artificial selection. Particularly, he mentioned in his private autobiography that his recognition of artificial selection as the main process in domestication triggered him to conceive the idea of natural selection (Darwin 1958). Darwin dedicated many pages to discuss about the domesticated pigeons in the *Origin*, fascinated by the extraordinary variety of form produced by the methodical selection applied by breeders. Although Darwin's analogy became a target of critics, Darwin made heavy use of this in arguing for the historical reality of common descent and the efficacy of natural selection.

1.1.2 Domestication and selection

The emergence of human civilization as we know today was achieved thanks to the domestication of plants and animals such as wheat, lentils, dogs, pigs, chickens (Purugganan and Fuller 2009; Driscoll et al. 2009; Larson and Burger 2013; Amills et al. 2017; Stetter et al. 2018; Avni et al. 2017; Redding 2015; Zeder 2012; Dayan 1994). Therefore, the domestication of such as plants and animals by *Homo sapiens* is one of the most crucial developments in the history of humans (Purugganan 2022). The transition of human societies from hunting and gathering to the cultivation of plants and herbing of animals started 11000 years ago, leaded to the domestication of crops and livestock. The definition of what domestication is, can be quite challenging, or it might be used to describe mistakenly interspecies relationships. Zeder (2012) gives a definition of domestication describing it as the mutualistic long-term relationship between humans and other species based on which both sides can increase their fitness. Specifically, a biologically centered definition of domestication can be described it as a coevolutionary process in which one species, the domesticator, constructs an environment where it actively manages both the survival and reproduction of another species, called the domesticate, in order to provide the former with resources. As a result, an increase in fitness can be observed for the organisms with this mutualistic relationship leading to the evolution of traits that ensures the stable association of domesticator and domesticate across generations. The pace of domestication is controlled by the strength of the selection applied by the domesticator and the genetic and ecological characteristics of the target domesticate (Purugganan 2022).

1.1.3 Pathways of domestication

Even though human evolution is strongly correlated to the domestication process, the genomic and evolutionary processes that go along with domestication is not fully understood. Selective pressures dictated by unintentional and deliberate human actions, possibly made the process relatively faster than natural selection in evolutionary time scale. Animal domestication was developed under different pathways including a direct human selection or an unconscious one. Zeder (2012) describes three different pathways followed by animal domestication, the commensal pathway, the prey pathway, and the directed pathway. Figure 1.1 shows the domesticated pathways followed by different species.

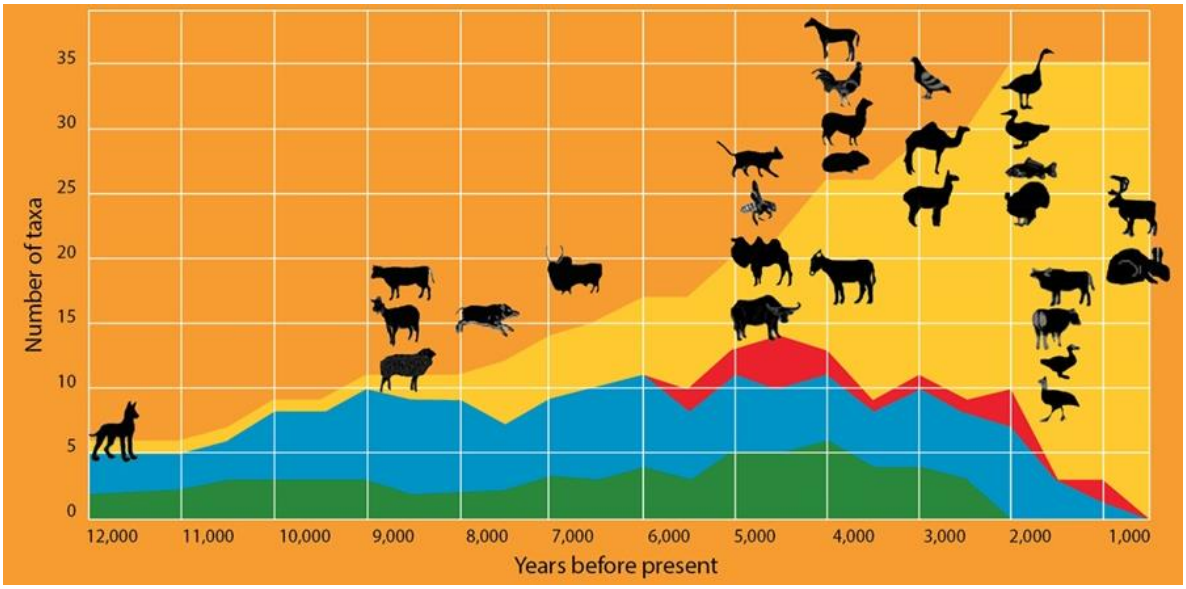

Figure 1.1: The first large carnivore domesticated was the dog. There are three possible pathways for the domestication of wild animals. The first one, the "commensal pathway" (green), occurred when humans saw a benefit from living alongside an animal. The second one called the "prey pathway" (blue), happened when humans bred animals in captivity for their meat. Directed domestication (red) happened when humans raised animals from the wild to take benefits for them. Post-domestication (yellow) took places for many species, since humans were able to recognize these traits of animals that could increase their benefits. Figure is form Larson and Fuller 2014, https://www.discovermagazine.com/planet-earth/the-origins-of-dogs)

1.1.4 Genetic variation in domestication

Domestication instances are related to bottlenecks and as a consequence a decrease of the effect of natural selection is predicted since a small number of individuals from the wild population become domesticated (Wright et al. 2005). Even in the human-modified environments that control the domesticated traits, the genetic patterns were modified as well by the effect of gene flow between the wild and domestic populations. The concept of introgressive capture, introduced by Larsson and Fuller (2014), highlighted the importance of gene flow between domestic and wild populations to spread domesticated animals across the geography, in front of models considering independent domestication events. Domestic animals differentiated from the wild relatives by differential effects of selection, reduced gene flow and drift. Nowadays, breeders apply truncation selection which is a standard method in selective breeding to select of the top percentage of individuals for the desired traits (Granleese et al. 2019). Desired traits can be related to fashion as coat color or have an economic impact as milk, wool, and egg-laying. Breeders then rank the animals based on their phenotypic value on some of these traits and the top percentage is reproduced (Crow and Kimura 1979). Moyers et al. (2018), described the cost of domestication as an increase in the number of deleterious mutations that are segregating at higher frequencies. As a result, the effect of selection and genetic gain can be reduced in a breeding program. It is still unclear which evolutionary processes contribute most to increase the cost of domestication across species. Andersson and Purugganan (2022) reviewed the progress of the last 35 years in discovering the genetic variation that underlies the phenotypic variation in crops and domesticated animals. A number of genes associated

with the domestication process have been identified in plants, while in animal domestication many traits seem to be polygenic without a specific gene involved.

1.2 Genetic variation and Population Genetics

Theodosius Dobzhansky defined evolution as the change over time in the genetic composition of a population in his pioneer work entitled "*Genetics and the origin of Species*" published in 1937. The genetic composition of a population is modified along generations due to alterations on the genomes that each individual carries. Population genetics studies the genetic composition of natural populations, and the evolutionary forces cause the genetic variation within and between populations.

1.2.1 Genotype and allele frequencies and the Site Frequency Spectrum (SFS)

Population genetics uses mathematical models of gene frequency dynamics to detect patterns that can explain the genetic variance in actual populations. These changes are due to evolutionary forces such as mutation, drift, migration, selection, and recombination (Charlesworth 2010). The most common type of genetic variation is single nucleotide polymorphisms frequently called SNPs. Each SNP represents a difference in a single DNA building block called nucleotide. Population genetics uses as basic unit the allele and genotype frequencies to describe the genetic structure instead of using just the genotype counts. Allele frequencies play a vital role in population genetics as well as genotype frequencies. The allele frequency spectrum or the site frequency spectrum (SFS) as it sometimes called, is the distribution of the allele frequencies of a given set of loci in a population or sample (Fisher 1930; Kimura 1964; Evans 2007; Hartl and Clark 2007). SFS can be calculated separately for synonymous and nonsynonymous sites.

1.2.2 Wright-Fisher model

Randomness in a natural population is introduced because of two reasons, Mendel's law of segregation and demographic stochasticity (Gillespie 1994). The first one is caused when a parent produces a gamete being randomly selected by one of the two homologous alleles. The second one is due to the different number of offspring that an individual can leave to the next generation. Even though these two reasons cause the genetic drift in a population, computer simulations follow simplest model which easier can be biological interpreted and mathematical analyzed. The Wright-Fisher model was the first one investigating the impact of genetic drift in this relatively simple way. Wright-Fisher model names after the early pioneers of theoretical population genetics, Sewall Wright and Ronald A. Fisher, describes the sampling of alleles in a population assuming non-overlapping generation times and genetic drift.

1.2.3 Diffusion approximations

While Wright-Fisher model considers discrete generations, diffusion theory models time and the frequencies of alleles as continuous variables assuming a large population size. Thus, diffusion

theory estimates a continuous probability distribution of allele frequencies over time. We expect that the estimations of both approaches, Markov chain and diffusion theory, will be similar considering a genetic drift. Wright (1931, 1945) laid the foundations of use diffusion process to model the dynamics of population generics. His work was completed by Kimura (1955, 1964) who defined fixation probabilities using forward and backward equations (Gillespie 1989). Other groups significantly contributed to estimation of the accuracy of diffusion approximations (Moran 1962; Watterson 1962; Ewens 1965). What make diffusion theory more versatile than Markov chain is that it can incorporate not only genetic drift but other evolutionary forces too (Evans et al. 2007) making it a central tool for modern population genetics.

Like Markov chains, diffusion theory can predict the probability distribution of frequency. Hence, both predictions are similar to the outcome of neutrality and genetic drift. Also, diffusion theory considers large populations resulting in a continuous probability distribution compared to discrete prediction over generations outputs form the Markov chain estimations. The probability of fixation of neutral alleles at frequency *x* in diffusion theory is $x/2N$, where *N* is the population size of a diploid species. Kimura (1964, 1968) gave the definition of the probability of fixation of a selective allele, one of the most important equations in population genetics:

$$P_{fixation} = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}} \qquad \textbf{(1.1)}$$

Where p is the initial frequency of the mutation, s is the selection coefficient (in co-dominance) and N is the population size of a diploid species. Particularly, the probability of fixation of a mutation strongly depends on the biological effect of this mutation on the individual and on the population size (Ohta 1973). The scaled selection coefficient Ns determines the dynamic of forces such as drift and selection. When the scaled selection coefficient is in the interval $-1 < Ns < 1$ the probability of fixation approaches that of neutrality. In case of $|Ns| > 1$, probability and time of fixation are mostly driven by selection. Whereas if $|Ns| > 10$, mutations are considered strongly dictated by the power of selection. Thus, the scaled selection coefficient is a term of high importance.

Following diffusion theory and the Wright-Fisher model, we will describe a fundamental equation used in modeling of evolutionary dynamics in population genetics. The equation suggested by Wright (1938) defines the population stationary frequency distribution. Specifically, the stationary frequency distribution describes the density probability of a mutation i at frequency $x + dx$, allowing the calculation of the frequency spectrum at different distributions of selection coefficients. Evans et al. (2007) explored further the potential of the equation including non-equilibrium populations. The equation for the stationary distribution is the following:

$$\varphi(x) = \frac{1}{x(1-x)} \frac{e^{4Ns} - e^{4Ns(1-x)}}{e^{4Ns} - 1} \qquad \textbf{(1.2)}$$

1.2.4 The neutral theory of molecular evolution

The neutral theory argues that the effect of the vast majority of molecular polymorphisms within species and substitution between species is neutral and dictated by the rate of genetic drift

and mutation. The role of genetic drift in molecular evolution has been controversial since the 60s, when it was introduced, mostly because the divergence of the species seems driven by random drift rather than by natural selection. However, neutral theory does not deny the role of natural selection, it argues that the most genetic variation has no effect on the fitness though. The theory which also called, the mutation-drift balance hypothesis, was proposed firstly by Kimura in 1968 while in 1971 Kimura and Tomoko Ohta analyzed fully the aspects of the theory in population genetics in their paper, "*Protein polymorphism as a phase of molecular evolution*". Particularly, Neutral theory suggests that the new mutations are either neutral or strongly deleterious (mutations that change the protein function negatively affecting the fitness of the individuals). However, these deleterious mutations don't contribute to the polymorphism or substitution rate since they are rapidly eliminated by selection. In Neutral theory advantageous mutations (also called beneficial or adaptive because they increase the fitness of the individuals) exist but are rare enough reaching fixation state quickly. As a result, since neither advantageous mutations contribute to the bulk of polymorphism or divergence, they can be ignored. The basic concepts of Neutral theory are following (Casillas and Barbadilla 2017):

- Deleterious mutations are quickly purged by the population while the adaptive mutations rapidly reach fixation. As a result, most genetic variation within species is neutral.
- Polymorphisms that are segregating in a population eventually get lost or fixed rather than balanced by selection.
- The level of polymorphism in a diploid population called θ, is defined as a function of two variables, the neutral mutation rate (μ_0) and the effective population size (N_e): $\theta = 4N_e\mu_0$. Heterozygosity increases with the increase of the population size.
- Neutral mutations reach fixation in a population at constant rate (K) independently of population size. K is defined as the product of the neutral mutation rate and the proportion of neutral mutation (f), as $K = f\mu_0$.

The strength of genetic drift is inversely proportional to the population size. Requirements of Wright-Fisher model are hardly satisfied in natural populations. However, we can assume an idealized population size (N_e) under which a Wright-Fisher population shows the same amount of genetic diversity as at the actual population. Under neutrality, one of the most important and simple equation of molecular evolution defines that the rate at which allelic alterations reach fixation in a species is equal to the mutation rate (μ_0). Thus, mutations are fixed in each generation in a population with a rate K. This simple equation encompasses the molecular evolution defining the rate at which species diverge over their evolutionary time.

1.2.5 Nearly neutral theory and the Distribution of Fitness Effect (DFE)

In 1973 Kimura´s neutral theory was reformulated by Tomoko Ohta (Ohta 1973), who introduced a new class of mutation, the nearly neutral mutations. When a new mutation enters the population is classified as neutral, deleterious or beneficial regarding its effect on the fitness of the individual carrying it. That means that the mutation is neutral when has no effect on the fitness, deleterious when disrupts important protein functions leading even in the death of the individual and beneficial when increases the fitness of the individual. However, the distribution of all the fitness

effects (DFE) of mutations in a certain species is rather continuous than discrete, as Otha proposed incorporating two new types of mutations, slightly deleterious and slightly beneficial (Ohta 1973, Ohta and Gillespie 1996). The theory predicts that nearly neutral mutations are mostly removed by natural population when the population size is large but there is a proportion of them which behaves as effectively neutral and randomly fixed or lost in small populations. Particularly, mutations with fitness effects much smaller in magnitude than $\ll {}^{1}/_{N_e}$ spanning in the range $-1 < N_e s < 1$, are considered effectively neutral. The fate of these mutations is dictated by the action of genetic drift. Mutations with selection coefficients s (fitness effects) on the order of $\approx {}^{1}/_{N_e}$ are nearly neutral. They are slightly deleterious if $s < 0$ or slight beneficial when $s > 0$. Both types of mutations have small effect on fitness and their fate is defined by a combination of natural selection and genetic drift. Finally, mutations with fitness effects $> {}^{1}/_{N_e}$ are strongly deleterious with $s < 0$ or strongly beneficial with $s > 0$. Their fate is determined by natural selection (Figure 1.2).

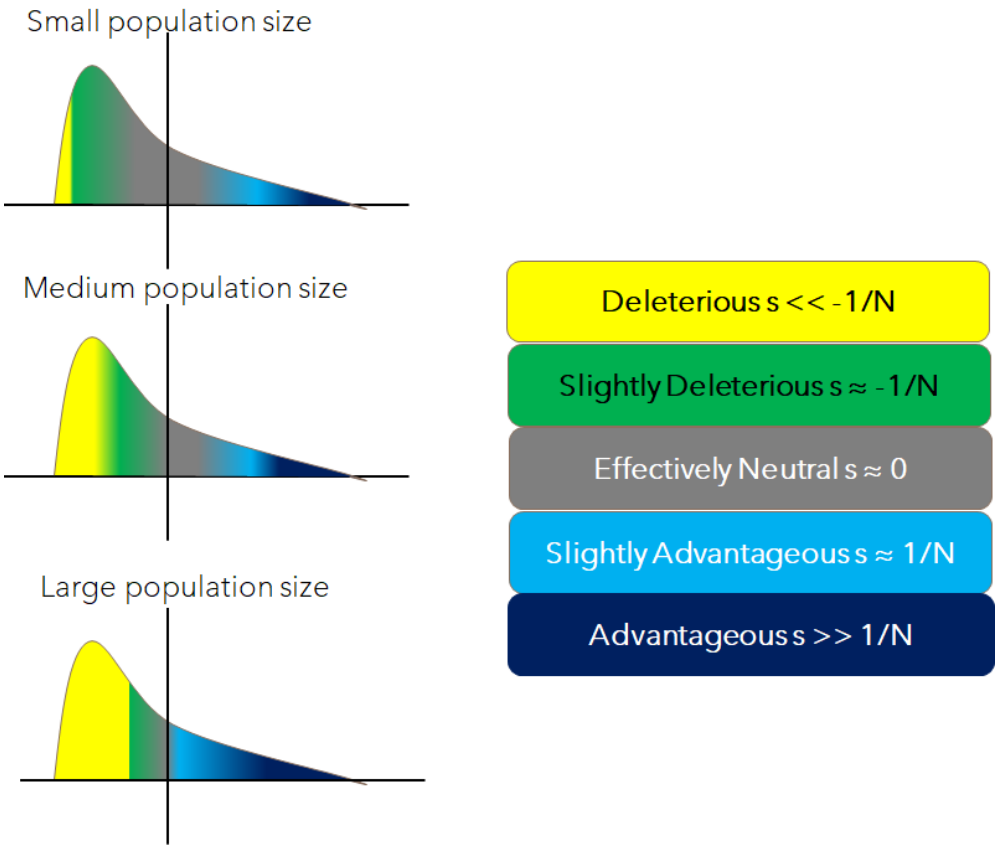

Figure 1.2: Hypothetical Distribution of fitness effects for populations of different sizes.

As we can imagine, when N_e is small then the range between $-1 < N_e s < 1$ is larger than in a large population meaning that there are more effectively neutral mutations. On the other hand, when N_e is large the mutations are subject to the action of natural selection. Consequently, a mutation can behave as effectively in one species when N_e is small while it can be under the action of natural selection in another species when N_e is large. The higher the value of N_e the lower the influence of genetic drift on new mutations. Then, natural selection dominates over deleterious mutations leading them in elimination while increases the frequency of those that are beneficial.

Considering a species following the nearly neutral theory we expect that the DFE is a continuum distribution of all the mutational effect ranging from those that are strongly deleterious, slightly deleterious, effectively neutral to those that are slightly beneficial and highly beneficial (Figure 1.2). The knowledge of the shape and strength of DFE can address several important questions around the evolution of the species and the patterns of genetic diversity. We can obtain information about the quantity of mutations segregate in a species and the evolutionary forces acting on them. Furthermore, by comparing DFE produced by different species, we can learn more about their divergence and their different genetic processes. The extent to which the DFE varies across species has yet to be revealed along with the different biological factors. DFE has been a fundamental tool in population genetics defining the proportion of new mutations that are advantageous, neutral, or deleterious. Most of the mutations are either deleterious or neutral while the contribution of the strongly deleterious and beneficial mutations in the standing variation is small. The strongly deleterious are eliminated by natural selection while the beneficial mutations reach fixation rapidly. (Eyre-Walker and Keightley 2007).

For modelling the DFE, several mathematical distributions have been used with two parameters. Particularly, the negative part (deleterious mutations) of the DFE is usually modelled by a normal, lognormal, gamma or beta-shaped distribution. For the positive part (advantageous mutations) an exponential distribution is usually used. Overall, the DFE can be modelled by a bimodal distribution. The site frequency spectrum (SFS) is widely used to infer the DFE of new mutation in a population.

1.2.6 Inference of DFE and adaptive substitutions

Expected SFS

Models to infer DFE use Maximum Likelihood (ML) estimations and assume a Poisson Random Field (PRF) framework (Sawyer and Hartl 1992; Sethupathy and Hannenhalli 2008). PRF framework is widely used for multiple statistical approaches to estimate the proportion of adaptive substitutions, including ML estimations of the DFE. The same framework can be applied to Bayesian models to infer the population-scaled selection coefficients (Sawyer et al. 2003). ML models perform likelihood estimates based on the expected levels of fixation and polymorphisms considering different evolutionary models. The first models of DFE modelled the deleterious part of the DFE using a Gamma distribution (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2009). However, the more sophisticated ML models today go a step further modeling not only the deleterious mutations but also the advantageous using an exponential distribution function (Galtier 2016; Tataru et al. 2017). Thus, the state-of-the-art methods in population genetics are based on the PRF framework presented in Galtier (2016) and Tataru et al. (2017) to perform an ML estimation of the DFE. From PRF theory, the expected counts of synonymous mutations (P_S) given a frequency i is estimated as:

$$E\left[P_{S[i]}\right] = \frac{4N_e\mu L_S}{i} \qquad (1.3)$$

Where L_S is the total number of synonymous sites, μ mutation rate per site per generation.
The expected counts of nonsynonymous mutations are defined as follows:

$$E\left[P_{N[i]}\right] = 4N_e\mu L_N \int_0^1 B(i,n,x)\, H(s,x)dx \tag{1.4}$$

Where L_N is the total number of nonsynonymous sites, whereas

$$B(i,n,x) = \binom{n}{i} x^i (1-x)^{n-i} \tag{1.5}$$

is the binomial probability of observing i derived alleles in a sample of size n when the true allele frequency is x, and H(s,x) is defined by Eq. 1.2.

To obtain the expected polymorphic count given the underlying DFE of new mutations, Eq. 1.4 should be integrated over the full DFE (given a distribution). From Eyre-Walker and Keightley (2009), the underlying DFE for new deleterious mutations is defined by a gamma distribution by the expression:

$$\varphi(s;\alpha,b) = \alpha^b s^{b-1} \frac{e^{-\alpha s}}{\Gamma(b)} \tag{1.6}$$

Where α and b are scale and shape parameters from the Gamma distribution. Then the expected polymorphic count given a particular DFE is defined as:

$$E\left[P_{N[i]}\right] = 4N_e\mu L_N \int_{-\infty}^{\infty}\int_0^1 B(i,n,x)\, H(s,x)\, \varphi(s;\alpha,\beta)dxds \tag{1.7}$$

Full DFE and divergence counts

Apart from considering a Gamma distribution for modeling the deleterious mutations, other methods have suggested different distributions as well, which we will review in the next sections. Unlike methods to infer a strictly deleterious DFE, Tataru et al. (2017) incorporates a full DFE that includes both deleterious and beneficial mutations. The divergence expression can be obtained by calculating the Eq. 1.7 to the limit of the frequency x to 1 and multiplying it by the time of divergence. Galtier (2016) and Tataru et al. (2017) defined the expected number of synonymous (D_S) and nonsynonymous (D_N) substitutions as (see also Murga 2022):

$$D_S = 4N\mu t L_S \tag{1.8}$$

$$\mathrm{E}[D_N] = 4N_e\mu t L_N \int_{-\infty}^{\infty} \frac{4N_e s}{1-\mathrm{e}^{(-4N_e s)}}\, \varphi(s;\mathrm{a},\mathrm{b})\, \mathrm{ds} \tag{1.9}$$

Note that if $s = 0$ then Eq. 1.7 and 1.9 for expected polymorphic and divergence sites respectively become similar to Eq. 1.3 and 1.8.

Inferring α

The proportion of adaptive substitutions α can be calculated from the observed divergence counts as follows:

$$\boldsymbol{\alpha = \left(d_N - d_N^{nadv}\right)/d_N} \quad \textbf{(1.10)}$$

where d_N^{andv} is the number of non-adaptive substitutions and is estimated as:

$$\boldsymbol{d_N^{nadv} = \frac{4L_N N_e \mu \int_{-\infty}^{s_{adv}} \frac{4N_e s}{1 - e^{(-4N_e s)}} \varphi(s) ds}{L_N}} \quad \textbf{(1.11)}$$

Thus, α is estimated subtracting the observed non-adaptive substitutions and neutral substitutions from the total observed divergence counts at selected sites. s_{adv} is defined by Galtier (2016) as the population selection coefficient threshold below which nonsynonymous substitutions are not considered adaptive. This expression is similar to the ones shown at Tataru et al. (2017) and Eyre-Walker and Keightley (2009) where the $s_{adv} = 0$. However, Galtier (2016) assumes positive values for s_{adv} to subtract nearly neutral positive fixation too. Note that while the above approach relies on the calculation of α based on the assumption that the ingroup and outgroup share the same scaled DFE. Yet, if the estimated full DFE is available from polymorphism data, then α can be estimated by replacing the observed divergence counts (d_N) with the expected ($\mathrm{E}[D_N]$) counts (Tataru et al. 2017).

In addition, demography, ascertainment bias, nonrandom sampling and linkage can affect the shape of the SFS resulting in bias under a Wright-Fisher constant population. To account for such distortions, different methods incorporate nuisance parameters Eyre-Walker et al. (2006). The nuisance parameter r can modify the frequency of the SFS individually. Hence, r modifies the effective mutation rate at each frequency i, considering the relative mutations rate at i with respect to the mutation rate at singletons Eyre-Walker et al. (2006). This approach accounts for such distortions that affect both neutral and selected sites.

1.2.7 Methods to infer the DFE

1.2.7.1 Estimates of DFE in experimental studies

The DFE represents the distribution of all selection coefficients (s) of random mutations in the genome. Thus, the DFE is an important quantity in evolutionary genetics because it determines how selection affects genetic variation (Eyre-Walker and Keightley 2007). Under experimental studies DFE can be directly estimated. Particularly, the DFE is estimated by directly measuring the fitness from a collection of single-step mutants or indirectly from observed changes in the population fitness in mutation accumulation (MA) experiments (Eyre-Walker and Keightley 2007; Bataillon and Bailey 2014). The first approach has been successfully applied in small mutational target regions in a number of viral, bacterial and yeast systems, examining the full spectrum of selection coefficients. These methods usually estimated a gamma, unimodal or bimodal shaped DFE (Fowler et al. 2010; Hietpas

et al. 2011; Boutcher et al. 2014; Bataillon and Bailey 2014). While the first approach directly identifying the mutations involved the second infers the DFE from fitness trajectories of a collection of population over time in MA experiments (Kim et al 2017). They modelled DFE by a gamma distribution and estimated the parameters that best fit to the observed changes in the mean and variance of fitness among populations (Halligan and Keightley 2009). The vast majority of these studies suggest a high proportion of new mutations are strongly deleterious pointing out that the DFE shape is less leptokurtic than an exponential distribution (Halligan and Keightley 2009). However, the true underlying DFE shape can be more complex than the gamma distribution or the methods MA show a bias towards mutations with large fitness effects (Eyre-Walker and Keightley 2007).

1.2.7.2 Estimates of DFE from genetic variation data

1.2.7.2.1 Folded and Unfolded SFS

A second category of methods to infer the DFE, models information from the SFS for two classes of nucleotide changes, the neutral ones (generally synonymous variability) and the amino acid changing ones under selection (nonsynonymous variability) found in natural populations. The SFS of a sample of chromosomes describes in how many segregating sites each allele appears. That is, in one site we can have one copy of this allele, two copies and so on. Accordingly, to the class of nucleotide changes, the SFS can describe the distribution of synonymous or nonsynonymous alleles across the sample. Furthermore, if we know the ancestral/derived allele at a polymorphic site we can obtain the unfolded SFS (uSFS) with $n-1$ classes for a sample of size n. Otherwise, the SFS is folded with $n/2$ or $n+1/2$ classes depending on whether n is even or odd. Unfolded SFS are always a better source of information than the folded one, the identification of ancestral state is challenging though leading to bias estimation of DFE (Keightley and Jackson 2018). For the estimation of the parameters, a maximum likelihood method is applied. The most common distribution used to model the DFE of deleterious mutations, is a gamma distribution with a mean $N_e s$ and a shape parameter. Many studies have used this approach to various species such as humans (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Boyko et al 2008; Li et al 2010), *D. melanogaster* (Keightley and Eyre-Walker 2007; Kousathanas and Keightley 2013), yeast, gorillas, and mice (Koufopanou et al. 2015; McManus et al. 2015; Halligan et al. 2013). Findings of DFE in humans from genetic variation data have been proved very useful to different genetic processes such as genetic load, ancient human introgression of Neanderthal alleles into humans, estimation of strength of selection acting on disease (Henn et al. 2016; Harris and Nielsen 2016; Uricchio et al. 2016; Moon and Akey 2016).

1.2.7.2.2 Assumptions

It is assumed that the nonsynonymous sites will be affected by selection. In the presence of slightly deleterious mutations, we could expect an excess of rare variants compared with the neutral SFS under a mutation drift equilibrium. However, selection is not the only evolutionary force affecting the SFS, it is affected by demographic changes too.

Thus, a similar excess of rare variants would result from a population expansion or not too recent bottleneck. The distinguishing between effects produced by selection and demography is

challenging yet necessary. For this scope, it is assumed that synonymous sites are evolving under neutrality and can be used to infer parameters related to demography. Then, these demographic parameters are used to correct for it when estimating the effect of selection from the nonsynonymous SFS. This is the central idea upon which all these methods have been built to infer the DFE from the SFS. Methods differ in the way the correct for the effects of demography, whether they use divergence data, consider beneficial mutations, assume a correlation of DFEs across populations on the same species. Another important assumption considered by the DFE models is the independence of the sites. However, this assumption is violated in real data by the effect of linkage selection.

1.2.7.2.3 Violation of assumptions: Linkage selection

1.2.7.2.3.1 Selective sweep

Genetic drift is not the only source of randomness in the dynamics of alleles (Coop 2020). Random genetic backgrounds with different finesses can alter the frequency of alleles. When a beneficial allele arises via a single mutation arises on a particular genetic background. As beneficial allele becomes established in the population and increases in frequency rapidly, having escaped the loss by genetic drift, other alleles that happened to be present on the haplotype that the mutation arose on will do the same. These alleles are usually neutral or at least not too deleterious and are getting to "hitchhiking" along (Smith and Haigh 1974). The hitchhiking has as effect the reduction of diversity around the beneficial alleles because neutral variants are swept along the beneficial alleles. The process was named selective sweep. An example of the effect of selective sweep has been identified in the genetic basis of melanism in the peppered moth (Biston betulatia). Van't Hof et al. (2011) found that the adaptation of the moth to industrial pollution (Cook et al. 2013) was achieved by the insertion of Transposable Element (TE) into a pigmentation gene and its sweep to fixation. As a result, a decrease of diversity in region around TE was observed.

When a novel selection pressure switches on, multiple mutations at the same gene may start to sweep such that no one of these alleles sweeps to fixation. This results to soften the impact of selection on genomic diversity and so are called "soft sweeps". Another way that the impact of a sweep can be softened is if our allele was segregating in the population for some time before it become beneficial. This type of variation is called standing variation. These standing variants can have recombined onto various haplotype backgrounds such that when selection pressures switch, the selected allele sweeps up in frequency on multiple different haplotypes (Hermisson and Pennings 2017).

1.2.7.2.3.2 Background selection

While populations experience a constant influx of deleterious mutations at functional loci, selection purge them from the population preventing deleterious substitutions and maintaining function at these loci. This balance between mutation and selection results in a constant level of deleterious variation in the population. As a constant selection against a deleterious mutation purges it from the site it removes with it any neutral alleles that were also on this haplotype. This constant

removal of linked alleles from the population reduces the diversity in the surrounding regions of the functional loci (Charlesworth et al. 1993, Hudson and Kaplan 1995, Nordborg et al. 1996). The effect is known as background selection (BGS).

1.2.7.3 The DFE of deleterious mutations

Many studies that infer DFE contrasting 1D-SFS of synonymous and nonsynonymous mutations, suggest that DFE has a strongly leptokurtic distribution in contrast to the observations made by MA-based estimates (Eyre-Walker et al. 2006, Keightley and Eyre-Walker 2007, Boyko et al. 2008, Li et al. 2010). A large proportion of nearly neutral mutations and strongly deleterious mutations reported by many site-directed mutagenesis studies (Batallon and Bailey 2014, Boucher et al. 2014). Huber et al. 2017 found that humans have more strongly deleterious mutations than D. *Melanogaster* and that species complexity is positive correlated to the fraction of new deleterious mutations. Studies in humans have estimated the parameters of a gamma distribution (Eyre-Walker et al. 2006, Boyko et al. 2008). They found approximately 56-61% of new nonsynonymous mutations to have moderately to strongly deleterious effect ($|s| \geq 10^{-3}$), 15-16% to have weakly deleterious effect and about 24-28% to have nearly neutral effect. In another study in humans, Li et al. (2010) found that best fit to their data a mixture distribution consisting of a neutral point mass and gamma distribution. They found that only 1% of new mutations have $|s| > 10^{-4}$ (compared to 57% reported by Boyko et al. 2008) and 78% of new mutations fall in the $10^{-4} \leq |s| \leq 10^{-3}$ range (compared to 15% in Boyko et al. 2008). The different estimated proportions of moderately vs strongly mutations in humans indicating that the estimation of DFE in humans is still elusive and not accurate. Note that the number of individuals used in a study of DFE can be key factor for the estimation of proportions related to the selection coefficient classes. Thus, the estimated DFE parameters can be in contrast in various studies (Boyko et al. 2008, Li et al. 2010). Last advanced approaches exploiting of all the genome information have led to more accurate and robust estimates of DFE (Tataru et al. 2017, Gutenkunst et al. 2009, Kim et al. 2017, Huber et al. 2017, Huang et al. 2021).

Finally, several mathematical models have been suggested to infer the DFE yet is still unclear what is the best type of distribution that may best fit the data. The DFE of *D. melanogaster* species seem to be better described by a lognormal DFE while of Mus musculus castaneous by a bimodal DFE (Kousathanas and Keightley 2013). Galtier and Rousselle (2020) found that a Gamma + lethal model can best describe DFE while the mean deleterious effects of nonsynonymous mutations is shared across species.

1.2.7.4 The DFE of deleterious plus advantageous mutations

While most studies have described the DFE of new deleterious mutations, the DFE of new beneficial mutations is yet to be investigated fully across species. Most of the studies mentioned assume that beneficial mutations contribute negligibly to polymorphism and are not modelled. The reasoning behind this assumption is that strongly beneficial mutations are fixed rapidly and consequently they don't contribute to the polymorphism (Smith and Eyre-Walker 2002; Keighley and Eyre-Walker 2007). Tataru et al. (2017,2019) showed that weakly selected deleterious and beneficial mutations can contribute to both polymorphism and divergence data. In addition, they suggested a

model (polyDFE) to infer the full DFE and the proportion of adaptive substitutions (a) using polymorphism data. Results indicated that not counting for the contribution of beneficial mutations to polymorphism can lead to biased estimation of the DFE and a. Using the same model Castellano et al. (2019) inferred the full DFE of new amino acids mutations across great apes. They found that the shape of deleterious DFE is constant across the set of closely related species while the confirmed that N_e plays an essential role in the strength of negative selection. However, the strength of negative selection across species varies more than expected given the differences in N_e. While Castellano et al. focused on new slightly beneficial mutations (still segregating), Zhen et al. (2018) used divergence data to detect beneficial mutations that reached fixation many generations ago. They found that strongly beneficial mutations contribute significantly more to divergence than to polymorphism. In addition, results indicated that when counting for the population size of the outgroup, the proportion of beneficial mutations in humans is higher than in mice and flies. The fraction of new beneficial mutations seems to be approximately 14% with the vast majority of them to have a small effect on fitness. In *D. Melanogaster*, the proportion of new beneficial mutations is smaller than in humans, approximately 1.5% (Huber et al. 2017). Galtier (2016) highlighted the importance of using beneficial mutations in shaping the SFS comparing various DFE models to 44 different datasets. Even though polymorphic data are widely used in the inference of DFE, dissimilar results have been presented by various studies indicating that DFE can be biased. Consequently, assumptions regarding the DFE shape, selection coefficient and populations sizes must be reviewed (Booker 2020; Zhen et al. 2021).

1.2.7.5 DFE Software

Models such as DoDFE and DFE-alpha (Eyre-Walker et al. 2006; Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009), estimate demography using a Wright-Fisher transition matrix. In each new mutation arises in a site, a scaled selection coefficient is assigned, the effective population size is constant among loci. The SFS jointly estimates demographic parameters while DFE is drawn from an underlying distribution fitted to the data. DFE-alpha is slow due to computational complexity when accounting for more complex demographic models. A different class of methods infer the DFE by using Poisson Ransom Field (PRF) approach (Sawyer and Hartl 1992; Hartl et al. 1994, Williamson et al. 2007; Boyko et al. 2008; Tataru et al. 2017). As we mentioned before, polyDFE infers the DFE from an unfolded SFS using only polymorphisms (SFS counts). It can model a full DFE, assuming a combination of different distributions as gamma and exponential to model mutations with negative and beneficial effects. Grapes (Galtier 2016), DFE-alpha and DoDFE don't account on error in SFS while polyDFE model an independent rate of error in the data. A set of nuisance parameters are used to correct for demography. An extension of polyDFE (Tataru and Bataillon 2019) can be used to fit several SFS datasets simultaneously. This approach can provide evidence for differences in DFE among genomic regions or species. Kim et al. (2017) suggested a new software to infer the DFE of new mutations under the PRF using the SFS. The method was an extension of dadi packaged (Gutenkunst et al. 2009).

Table 1.1: Table based on Moutinho et al. 2020.

References	Input Data	$N_e s$ distribution (DFE)	Beneficial DFE	Joint DFE	Method
Bierne and Eyre-Walker 2004	polymorphism levels (P_N, P_S); divergence data	Gamma; Beta	No	No	DoFE
Eyre-Walker et al. 2006 and Eyre-Walker and Keightley 2009 Stoletzki and Eyre-Walker 2011	folded SFS; divergence data	Gamma	No	No	
Keightley and Eyre-Walker 2007 and Eyre-Walker and Keightley 2009	folded SFS; divergence data	Gamma	No	No	DFE-alpha
Schneider et al. 2011	unfolded SFS; divergence data	Gamma	No	No	
Galtier 2016	unfolded/folded SFS; divergence data	Gamma; GammaExponential; Displaced Gamma; FGMBesselK; SclaledBeta	Yes	No	Grapes
Tataru et al. 2017 and Tataru and Bataillon 2019	unfolded SFS; divergence data (optional)	Gamma; Exponential; GammaExponential; Displaced Gamma; K bins	Yes	No	polyDFE
Gutenkunst et al. 2009 and Kim et al. 2017	unfolded/folded SFS	Gamma, Lognormal, Exponential, beta, normal,	Yes	Yes	dadi
Uricchio et al. 2019b	unfolded SFS; divergence data	Gamma; Continuous	Yes	No	ABC*-MK

*ABC corresponds to the approximate Bayesian computation

The latter uses diffusion theory to compute the expected SFS for a set of demographic and selective parameters. The extension offered a computational improvement of dadi allowing the model to precompute the SFS for models involving more than a single selection coefficient. They inferred demography and selection from segregating sites in a maximum likelihood framework.

Firstly, they estimated a demographic model from synonymous sites and then conditionally to the estimated demographic parameters the DFE of nonsynonymous mutations was estimated. All the previously mentioned methods use a 1D-SFS to fit the data and estimate the DFE. Recently, another method based on Kim et al. (2017) was suggested using the 2D-SFS (Huang et al. 2021), to infer a jointly DFE between species that have undergone an environmental change. Tataru and Bataillon (2019) and Huang et al. (2021) have been discussed and used in Chapter 3. Table 1.1 shows a summarize of DFE models based on Moutinho et al. (2020).

1.2.7.6 Quantifying the fraction of adaptive substitutions

Population geneticists usually look at evidence of positive selection in the genome to identify adaptive variants and quantify the impact of selection in the genome. It is expected that during the process of fixation of adaptive variants, selection leaves signatures in the genome such as a reduction in the genetic diversity, a skew toward rare derived alleles and an increase in the linkage disequilibrium (Nielsen 2005; Franssen et al. 2015; Garud et al. 2015). Background selection also reduces the level of genetic variation in the region by eliminating chromosomes carrying strongly deleterious mutations (Charlesworth et al. 1993; Casillas and Barbadilla, 2017).

1.2.7.6.1 D_N/D_s *ratio*

The strength and direction of selection can be calculated by contrasting the nonsynonymous (D_N) and synonymous divergence (D_N) in a given gene (Miyata et al. 1990; Yang and Nielsen 2002, Eyre-Walker et al. 2006). Assuming that mutation rates at synonymous and nonsynonymous sites are constant and equal and that silent substitutions are neutral, then one expects that the ratio D_N/D_S (noted as ω) equals 1 under neutrality. Otherwise, if $D_N/D_S > 1$ genes are under positive selection since the advantageous mutations have been frequent among nonsynonymous sites and spread faster in the population than neutral mutations. Finally, if $D_N/D_S < 1$, genes are under negative selection with deleterious mutations have been removed by the population. The statistic is more efficient when genes are under strong positive selection otherwise the value tends to be lower than 1 since the most nonsynonymous mutations are expected to be deleterious (Yang and Bielawski 2000; Yang and Nielsen 2002; Eyre-Walker 2006).

1.2.7.6.2 McDonald and Kreitman test

The McDonald and Kreitman (MK, 1991) test is one of the most widely used methods in population genetics to identify protein coding sequences under positive selection combined both between-species divergence (D) and within-species polymorphism sites (P). Particularly, it compares the number of polymorphisms to the number of substitutions for a locus in two classes of sites, synonymous (which are assumed to evolve neutrally) and nonsynonymous (which are potentially under selection). The number of nonsynonymous and synonymous substitutions is denoted as D_N and D_s respectively while the number of nonsynonymous and synonymous polymorphisms is defined as

P_N and P_S. If all mutations are either strongly deleterious or neutral, then $^{D_N}/_{D_S}$ is expected to be equal to $^{P_N}/_{P_S}$. Conversely, if $^{D_N}/_{D_S}$ is higher than $^{P_N}/_{P_S}$ is taken as signature of positive selection since then adaptive mutations rapidly reach fixation and thus contribute more to divergence than to polymorphism compared to neutral mutations. If $^{D_N}/_{D_S}$ smaller than $^{P_N}/_{P_S}$ balancing selection is on action in the region (McDonald and Kreitman 1991; Eyre-Walker 2006).

1.2.7.6.3 Proportion of adaptive substitution (α)

An extension of MK test to estimate the proportion of adaptive substitution is given considering that adaptive mutations contribute substantially more to divergence than to polymorphism. Thus, proportion of adaptive evolution is defined as $\alpha = 1 - {}^{D_s P_N}/_{D_N P_s}$ (Charlesworth 1994; Smith and Eyre-Walker 2002). Estimates of α for single genes tend to have large sampling variances since the numbers of polymorphic sites and nonsynonymous substitutions are very small for most genes taken individually. Hence, a solution can be the pooling of data across many genes summing counts of polymorphisms and divergence in each category or by calculating the average across genes (Stoletzki and Eyre-Waler 2011; Fay et al. 2001; Smith and Eyre-Walker 2002). Another approach suggested by Sawyer and Hatl (1992) used a PRF model to define the expected counts of D_N, D_S, P_N, P_S assuming that the processes of mutation, selection and genetic drift acting independently and simultaneously ay multiple sites. This approach was fundamental for the development of Bayesian models relating the scales selection coefficient and counts of polymorphism and divergence (Moutinho et al. 2020). Particularly, Bayesian models would assume a fixed-effect model when the scaled selection coefficient (γ) is constant across sites (Bustamante et al. 2002) or would be of random effects if γ of each new mutations is coming from a single underlying normal distribution (Sawyer et al. 2003).

However, a limitation of these approaches is that estimates of α can be biased since they do not account for the segregation of slightly deleterious mutations (Smith and Eyre-Walker 2002). Also, while the most methods assume that sites evolve independently, there evidence that selection at linked sites controls for patterns of polymorphisms (Barton 1995; Andolfatto 2007). Moreover, the frequency of a given allele can be shaped by genetic draft, a process of recurrent selective sweeps at closely linked positions (Gillespie 2000). Messer and Petrov (2013) developed an improved extension, the asymptotic MK test, to correct for the underestimation of α due to slightly deleterious mutations accounting for the effects of background selection and genetic draft. An extension of this method suggested by Uricchio et al. (2019b), investigating the impact of background selection on the rate of adaptation using an approximate Bayesian computation method (ABC). These methods are less sensitive to the demography of population (Moutinho et al. 2020). Other robust estimates of α that are not sensitive in the presence of slightly deleterious mutations, are derived from the DFE and accounting for demography (Keightley and Eyre-Walker 2007; Galtier 2016; Tataru et al. 2017). As it is already mentioned, polyDFE estimates α using only polymorphism counts to infer the negative DFE and the expected number of non-adaptive nonsynonymous substitutions is contrasted with the observed number of nonsynonymous substitutions to estimate α.

1.3 The analysis of complex traits: Polygenic adaptation

1.3.1 From sweeps to shifts

The genetic architecture of a complex trait can be varied, and the trait can be controlled either through a few loci with strong effects or via many loci with small effects (Jain and Stephan 2017a; b; Orr and Coyne 1992; Johnson and Barton 2005). Consequently, different patterns of genetic diversity can be observed around the selected loci depending on these two models of selection (Figure 1.3, Stephan and John 2020).

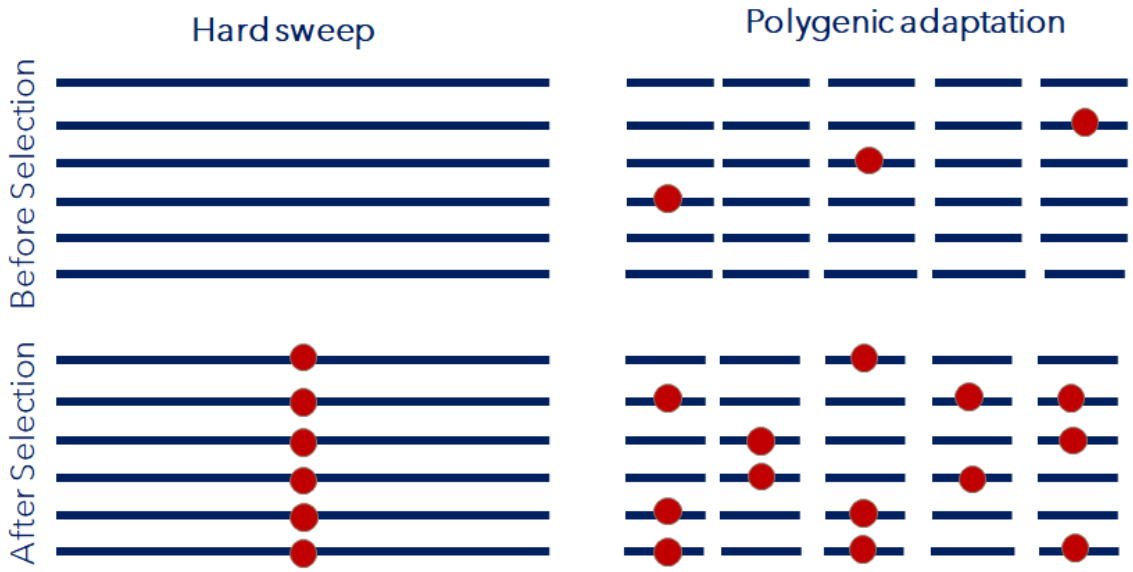

Figure 1.3: Figure depicts the genomic footprints observed in regions under a hard sweep and a polygenic adaptation process.

Usually looking at the genome, we try to find evidence of positive selection through the existence of hard selective sweeps. In standard selective sweep models, a single new mutation sweeps through a population to fixation, purging variation from a region of linkage around the selected site (Figure 1.3, Smith at al. 1974). Recent models focus on partial sweeps and soft sweeps as well. A hard sweep pattern in the genome shows that a strong selection is acting on the selective loci affecting few of them (Pritchard et al. 2010; Hermisson et al. 2005). Patterns like these have been observed in domesticated traits such as IGF2 gene region associated with lean domestic pigs (Andersson 2012; van Laere et al. 2003) and in the thyroid-stimulating hormone receptor (TSHR) in domestic chickens (Rubin et al. 2010). Domesticated traits controlled by a few loci like the one reported above, can be described by the standard Mendelian genetic architecture. On the contrary of selective sweeps, polygenic adaptation describes a process under which a population adapts to a new environment through small changes in allele frequencies at hundreds or thousands of loci (Figure 1.3, Pritchard et al. 2010). Most of the traits in humans but in other species too, are highly polygenic, affected by standing genetic variation at many loci. When a population experiences a change in its phenotypic optimum, the population adapts to its new environment via small direction shifts in allele

frequencies spread across all the variants that affect the trait. Because of these small and subtle changes, polygenic adaptation is hard to be detected. Boyle et al. (2017) proposed an extended version of the polygenic model which was similar to the infinitesimal model developed by Ronald Fisher, the omnigenic model. According to the model thousands of individuals genes with biological relevance to a trait and their regulators, contribute at least slightly to the phenotype through the expression in relevant cells. However, polygenic adaptation signals can be overestimated because of the effect of population stratification on the genome as observed in height (Sohail et al. 2019).

1.3.2 Simulating the polygenic adaptation

Numerous studies in population genetics, model and simulate the polygenic adaptation process (Stephan 2016; de Vladar and Barton 2014). Jain and Stephan (2015) studied the response of polygenic trait in an infinite large population under the action of stabilizing selection and mutation assuming that the allelic effects control the trait vary between the loci. Their results could be interested when considering a rapid adaptation since studies have shown that indeed adaptation can occur very rapidly (e.g., Reznick 2009, Vignieri et al 2010), as they did in a later study in 2017 (Jain and Stephan 2017a). In the latter study, they showed that fast polygenic adaptation can be caused by two very different mechanisms, strong positive directional selection at a few loci of large effects or subtle shifts of alleles at many loci of small effects. Combinations of these two mechanisms could lead to rapid adaptation. Finally, the study highlights the need of new powerful statistical methods to detect the signatures of polygenic adaptation in the genome. The short-term response of a quantitative trait after a sudden environmental change of the phenotypic optimum is related to the effect sizes and to the scaled mutation rate (Jain and Stephan 2017b). Polygenic adaptation in response to selection on quantitative trait was investigated under a highly polygenic model that includes both directional and stabilizing selection in a population of finite size that experiences random genetic drift (Stephan and John 2020). Adaptation of populations to new environments is often accompanied by population size bottlenecks. Because bottlenecks reduce the genetic variance, they cause large deviation of the trait mean from the fitness optimum. Also, the effect of genetic drift seems to reduce the signals of polygenic adaptation in the genome (Stephan and John 2020). Höllinger et al. (2019) studied the effect of selection and genetic drift in the two adaptive modes, hard sweep, and polygenic adaptation by measuring the sweeps and small allele frequencies shifts in a finite population size. Forward simulations used to describe the average behavior of selected and neutral mutations during the adaptation of a quantitative trait to a single sudden shift in the optimal trait value (Thornton 2019). The study showed that the new optimum trait in a population of finite size is reached before selective sweeps are completed in polygenic models with many involved loci. The impact of the demographic conditions investigated further by Stetter et al. (2018) using forward simulations for several different traits varying in the effect size distribution of new mutations, the strength of stabilizing selection and the contribution of the genomic background. They found that selective sweeps occur even for traits under relatively weak selection and where the genetic background explains most of the variation. Also, the study showed that population bottlenecks and expansion affect the genetic variation along with the relative importance of sweeps from standing variation and the speed with which adaptation can occur. As many of the mentioned studies showed,

demographic conditions can affect the capacity of detecting signatures of polygenic adaptation in the genome causing deviation of the fitness optimum.

As previous work has shown, the detection of polygenic adaptation is harder that in classic selective sweeps (Pritchard et al. 2010). However, some studies successfully detected polygenic adaptive signals. Firstly, an empirical framework and a model-based rejection sampling approach was developed by Uricchio et al. (2019a) for detecting polygenic selection and mutational bias that can be applied to genome-wide association studies (GWAS) data for a single population. Berg and Coop (2014) found that combining GWAS with population genetic modeling can be proved a powerful method for detecting signals of polygenic adaptation. Distinct patterns of selective sweep and polygenic models were found by generating computer simulations for each model based on the experimental evolution framework (Barghi and Schlotterer 2020). A polygenic adaptive response to temperature changes was detected in *D. melanogaster* that could be explained by functional redundancy and quantitative traits (Barghi et al. 2019). Selection acting in maize domesticated traits seems to eliminate large-effect genetic variants while the small-effect polygenic variants are responsible for most of the standing variation (Xue et al. 2016). In humans, evidence of polygenic adaptation was detected at the pathway or gene set level instead of analyzing single independent genes. Most pathways globally enriched for signals of positive selection are whether directly or indirectly involved in immune response (Daub et al. 2013). Initial domestication process in animals seem to act mostly on complex behavioral traits such as taming (Trut et al. 2009) while other studies have shown a large number of genes involve in domesticated traits (Jasinska and Freimer 2009).

Chapter 3 of this thesis focuses on detecting signals of domestication process simulating it under the two different adaptive mechanisms, a few loci with large effects and many loci with small effects. The study uses forward simulations under different selective and demographic parameters to define which conditions could describe better a domestication process.

1.4 The analysis of complex traits: Prediction

Quantitative genetics focuses on the phenotype and aims at modelling the genetic basis underlying phenotypic variation in a population. The main assumption of quantitative genetics is that many genes influence a trait while non-genetic factors may also be important. Therefore, to understand how genetic variation contributes to phenotypic variation is a basic question in genetics. During the beginning of the 20th century, there was a debate between supporters of Mendelian inheritance which believed on discrete, monogenic phenotypes and of biometricians, who argued that Mendelian genetics could not explain the continuous distribution of variation observed for many traits in humans and other species. RA Fisher gave an end to this fierce debate showing that if many genes affect a trait, then the random sampling of alleles at each gene produces a continuous, normally distributed phenotype in the population (Fisher 1918). As the number of genes grows very large, the contribution of each gene becomes correspondingly smaller, leading in the limit to Fisher´s famous "infinitesimal model" (Barton et al. 2016; Boyle et al. 2017). Fisher (1918) together with Haldane (1932) and Wright (1921), laid the theoretical basis of quantitative genetics which was established around the 1920 by the work of these three pioneers. They proposed statistical methods, such as the analysis of variance and path coefficients, to partition the variation and describe the resemblance between relatives. These methods have remained at the center of the field since then and allow

predictions of quantities such as response to artificial and natural selection. Some of the useful parameters are the breeding value (A), which is the expected performance of offspring, and the heritability ($h^2 = {}^{V_A}/_{V_P}$, the ratio of additive genetic variance to the overall phenotypic variance). Quantitative genetics has various applications. Firstly, can be used to help us explain and understand the phenotypic evolution in natural populations and between them, as well as the selective breeding of domestic animals and crops. Secondly can be used in methods of animal and plant improvement, such as genomic prediction, and for alleviation of complex disease focusing on detecting genes associated to specific diseases (Hill 2010).

1.4.1 The statistical foundation of quantitative genetics

1.4.1.1 Genetic variance components

The partition of variation to different causes is fundamental for a trait. The amount of variation is measured and expressed as the variance. The main components of the variance are the genotypic variance (V_G), that is the variance of genotypic values, and the environmental variance (V_E), that is the variance of environmental deviations. Then, the total variance will be the sum of the separate components and is defined as the phenotypic variance (V_P). The genetic variance can be further divided into additive variance (V_A), dominance variance (V_D) and interaction variance (V_I). The model of partition variance components firstly proposed Fisher (1918), Cockerham (1954), Kempthorne (1954) and later popularized by Falconer and Mackay (1996) and Lynch and Walsh (1998). An important question about what determines a phenotype is the role of heredity versus the environment. Then, the relative importance of a specific source of variation is the variance caused by that source versus the total phenotypic variance. The relative importance in determining phenotypic variation is named heritability. Table 1.2 shows the different categories of variance components.

Table 1.2: Variance components

Variance Component	Symbol	Corresponding variance
Phenotypic	V_P	Phenotypic value
Genotypic	V_G	Genotypic value
Additive	V_A	Breeding value
Dominance	V_D	Dominance deviation
Interaction	V_I	Interaction deviation
Environmental	E	Environmental deviation
Narrow sense heritability	$h^2 = {}^{V_A}/_{V_P}$	Phenotypic value due to genes transmitted from the parents
Broad sense heritability	$H^2 = {}^{V_G}/_{V_P}$	Phenotypic value due to genotypes

1.4.1.2 Linearity

To assess the relationship between genotypes and phenotypes, linear models are used to fit the data. Usually, a linear regression of phenotypes for our individuals on their genotypes at a particular SNP_l is defined as follows:

$$\boldsymbol{y \sim \mu + \alpha_l G_l} \qquad \textbf{(1.12)}$$

Where y is a vector of phenotypes of a set of individuals, G_l is the vector of genotypes at locus l taking the values 0, 1 or 2 depending on whether the individual is homozygote, heterozygote, or the alternate homozygote at the locus of interest and μ is the phenotypic mean. The slope of this regression line (α_l) is interpreted as the average effect of substituting a copy of allele 2 for a copy of allele 1 (Coop 2020). A basic assumption is that the regression is linear.

1.4.1.3 The infinitesimal model

We can predict the phenotypic values for the first generation under truncation selection using the breeder´s equation, Response = $h^2 \times$ S, where S is the selection differential (Lush 1937). The breeder equation predicts evolutionary change in a trait of interest. The process of selection either natural or artificial, introduces into the prediction through S. The selection differential is a measure of association between trait values and fitness (Falconer and Mackay, 1996). Hence, S is negative when lower values of a trait increase fitness; positive when selection favors higher traits values. In the case of truncation selection, where a fixed proportion of the population is selected and reproduced, S is equal to the difference in mean traits values between the selected individuals and the entire population. However, we know that selection changes the gene frequencies and hence the genetic variance (Hill 2010), making the prediction of response without knowing of the individual gene effects and frequencies difficult. Bulmer (1980) suggested a formalized Fisher´s infinitesimal model to provide a practical but some unrealistic biological resolution; the model assumes infinitely many unlinked genes with infinitesimally small additive effect so that the selection produces negligible changes in gene frequencies and variance at each locus. Consequently, the selection response in successive generation can be predicted by estimated population parameters such as heritability and phenotypic variance (Hill 2010).

1.4.2 Genomic prediction and analysis of polygenic traits

1.4.2.1 The incorporation of markers to determine genetic potential of livestock and plants era

The investigation of DNA during the late 1970´s and early 1980´s was fundamental for the discovery of several polymorphism marker types in the genome. One of the first works describing the multiple uses of the new polymorphism published by Soller and Beckman (1983). Surprisingly, their vision of using markers was not much different than how DNA is used today in the genetic improvement of livestock and plants. Specifically, they assumed that markers would be beneficial in constructing more precise genetic relationships, followed by parentage determination and the identification of quantitative trait loci (QTL). Although the newly discovered markers looked very promising, the high cost of genotyping animals and plants at that time prevented the early

widespread use of this technology. However, the Human genome project in 2001 (The International SNP MAP Working Group, 2001) allowed the discovery of numerous SNPs, which genotyping could be automatized. They have become the main markers used to analyze variation as they can found throughout the entire genome (Schork et al. 2000). Also, to genotype SNPs is cheap and easy in an automated high-throughput manner (Lourenco et al. 2017).

Marker genotyping contributes to the detection of genes that affect traits of importance. The idea behind this task is that a SNP found to be associated with a phenotypic trait is a proxy for a nearby gene or causative variant (a SNP that directly affects the trait). Since many SNPs are present in the genome, at least one SNP would be linked to a causative variant increasing the chance of finding genes that contribute to the genetic variation of the trait of interest. New genetic tests or profiles of DNA were developed trying to find which of these tests were associated to genetic variation of traits. One method that became popular was the marker assisted selection (MAS). MAS is an indirect selection method, where a trait of interest is selected based on its association to a marker. The process has been extensively used in plant breeding. For example, individuals with disease resistance are selected if a marker allele is identified that is linked with disease resistance rather than the level of disease resistance. This could result in great genetic improvement with the selection of parents that fulfilled the desired marker profile. However, many quantitative and complex traits of interest are often controlled by many small-effect genes and influenced by environmental factors which have been difficult to take advantage of in practical breeding (Lande and Thompson 1990). These genes of small effects are difficult to map and even if mapping is successful often multiple quantitative trait loci are involved which are usually difficult to use simultaneously in breeding (Robertsen et al. 2019). Therefore, MAS when defined as the use of mapped genes in breeding has had limited success in improving quantitative traits (Heffner et al. 2009). Worthy to note that using MAS, important genes or loci were detected but not always the same as it would be expected in replicated studies; that is the most of these QTL had small effects on the traits (Meuwissen et al. 2016). Andersson (2001) showed that the number of QTL associated with a phenotype depends on the threshold effect size used. To sum up, for any given polygenic trait there is only a small number of genes contribute more than 1% of the genetic variation (Lourenco et al. 2017).

1.4.2.2 The genomic selection era

In 1989, Fernando and Grossman incorporated marker information into Best Linear Unbiased Prediction (BLUP). The method resulted in a large amount of genetic gain in breeding programs which traditionally used pedigree information to define the covariance between relatives. The covariance matrix using markers is named the genomic relationship matrix (GLM). An extension of this idea was proposed by Meuwissen et al. (2001), introducing what we know today as genome-wide selection or genomic selection (GS). The paper suggested that using SNPs results in an increase of genetic gain especially for traits with low heritability as well as animals can be selected early in life prior to performance or progeny testing. Therefore, GS promised to overcome the limitations of MAS of quantitative traits. The goal of GS is to determine the genetic potential of an individual instead of identifying the specific QTL. However genetic markers are not only used for determining the genetic value of individuals so that they can be selected for breeding purposes. They are widely used for the estimation of heritability and genetic variance components as well as for the prediction of genetic

merits such as phenotypic values especially in animal and plant industries (Genomic Prediction). Genomic prediction (GP) is established to select new lines and crosses based on genomic data without the need for laborious phenotypic by making accurate prediction of phenotypic values using statistical methods. A schematic workflow of GP is depicted in Figure 1.4.

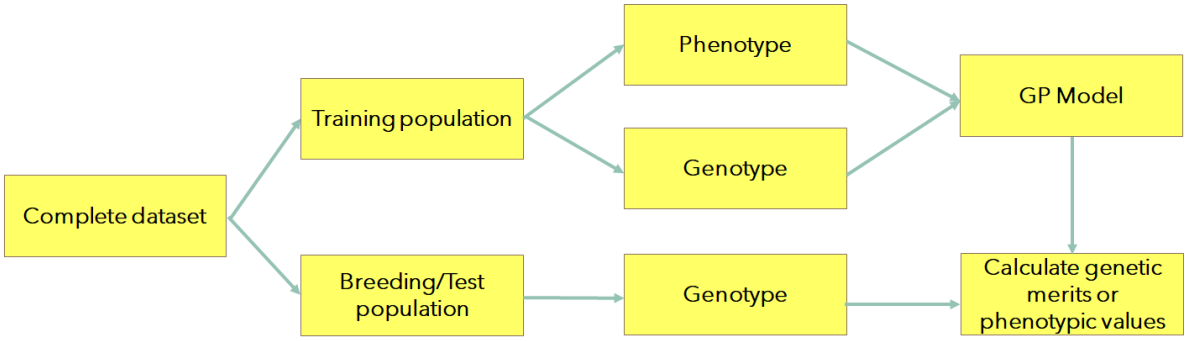

Figure 1.4: Overview of GS with cross validation using a training population to predict the phenotypic or genetic values of lines in the test-population. Particularly, the process starts with the generation of training population, that is individuals having both genotypic and phenotypic information. This information is used to build a model, where the phenotype is used as response and genotype as a predictor. Then, the information from the developed model is used to estimate the breeding or phenotypic values of breeding populations, i.e, individuals having only genotypic information.

1.4.2.3 Importance of linkage disequilibrium and marker density

The importance of incorporating SNPs in the GP studies relies on the fact that they may be linked to QTL or genes through linkage disequilibrium (LD). The LD measures the non-random association of alleles across loci and is based on expected versus observed allele frequencies. This association can represent the physical distance between loci that is strong LD means two loci are close. As a result, a locus can be used as a proxy for the others. Assuming a dense SNP panel, there is a high chance that QTL will be in LD with at least one SNP. Then if QTL A is linked to SNP B, depending on the strength of this linkage, once SNP B is observed it will imply QTL A was inherited together. As a result, an indirect association between SNP and trait can be observed (Figure 1.5). Meuwissen et al. (2001) showed that prediction accuracies are increasing with an increase in marker density and in training population size. It can be argued that at least one marker should be in LD with each QTL to capture all the genetic variation in a population. Especially when unrelated lines are used, as LD between markers may vary between the training and the test population. Also, high marker density seems to be more critical in prediction of distant relatives (Robertsen et al. 2019; Norman et al. 2018).

1.4.2.4 Genetic factors affecting the predictive ability

The accuracy of GP methods in crop breeding can be affected by several genetic factors as marker density, linkage disequilibrium (LD) between markers and QTL, sample size, the relationship between the training population and test population, population structure, heritability, and genetic architecture of the traits of interest (Xu et al. 2021). The predictive ability increases as marker density and sample size grow until reaching a plateau. Several studies have shown that the prediction

accuracy is affected by the training population size (Voss-Fels et al. 2019, Guo et al. 2019). The number of lines to be genotyped and phenotyped is of high importance for breeders since determines the training dataset and at the same time is a large investment. A reduced accuracy of GP was observed by reducing the training dataset while accuracy values varied more (Nielsen et al. 2016). The desirable size of the training population is associated with the heritability of the target trait and population relatedness. For example, when the heritability is low (e.g., h^2=0.2), the training population size must be more than 1000 individuals (Voss-Fels et al. 2019). Moreover, the necessary size of training population is much smaller for a closely related population than that for a distantly related population. As studies have shown (Jia 2017) the predictive ability is closely influenced by the heritability of traits. High heritability traits such as plant heigh often have higher predictive abilities than low heritability traits such as grain yield. Another important factor affecting the prediction ability is the genetic relationship between the training and the test population. It seems that the prediction accuracy grows when the populations are genetically similar (Daetwyler et al. 2014, Wang et al. 2018). Isidro et al. (2015) observed the highest prediction accuracies when training data represented the whole population and had a strong relationship to the testing data. On the contrary, a decrease in the prediction accuracies was observed by Lorenz et al. (2015) and Nielsen et al. (2016) when using less related individuals. A higher predictive ability can be obtained by adding more related materials in the training population rather than increasing the size of the training population with unrelated materials. However, we have to increase relatedness with caution since this might damage the genetic gain

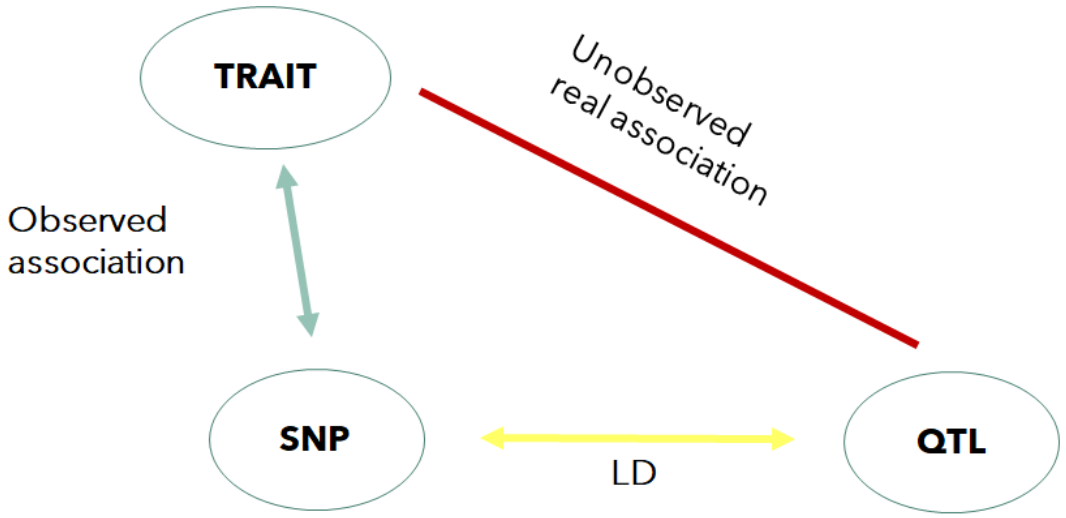

Figure 1.5: GP relies on the assumption that a QTL will be in LD with at least one SNP. Therefore, instead of detecting all the QTL associated with a trait we can find the indirect association between SNP and trait. Figure adapted by Lourenco et al. 2017.

in the long run as the genetic variation will be limited if the related populations are overused. As a result, a balanced relationship between training and testing population is the optimal in practical breeding (Xu et al. 2020). The predictive ability is also closely related to population structure as GP in stratified populations can lead to biased effect estimates and predictive ability (Yu et al. 2006). Finally, the degree of LD between markers and QTL influences GP. The training population needs to be

updated regularly because the LD between markers and QTL will gradually decrease as the number of generations grows (Desta et al. 2014).

1.4.2.5 Genetic variance explained by markers

It is important to know what part of the genetic variance is explained by each marker. If a marker has an effect of α_i for each copy of the A allele and the frequency of AA is p^2, then these individuals have value of $u = +2\alpha_i$; individuals aa with frequency q^2 have value $u = 0$; individuals Aa with frequency $2pq$ have a value of $u = \alpha_\iota$. Hence, the variance explained by this marker is $V(u) = E(u^2) - E(u)^2$. Finally, the variance explained by one marker is $4p\alpha_i^2 + 2pq\alpha_i^2 - (2p\alpha_i)^2 = 2pq\alpha_i^2$. It is expected that markers with intermediate frequencies will explain most genetic variance (Legarra et al. 2014). However, how do we calculate the total genetic variance explained by markers? Assuming that in most cases the marker effects are not known, a prior information might be used as their variance. Then, the total variance is

$$\boldsymbol{\sigma_u^2 = Var(u) = 2\sum_i^{nsnp} p_i q_i \sigma_{\alpha i}^2} \qquad \textbf{(1.13)}$$

If the markers have effect coming from distribution with the same variance α priori $\sigma_{\alpha 0}^2$ (say $\sigma_{\alpha 1}^2$=$\sigma_{\alpha 2}^2$=$\sigma_{\alpha 3}^2$ =…=$\sigma_{\alpha 0}^2$), Then σ_u^2= $2\sum_i^{nsnp} p_i q_i \sigma_0^2 = 2\sigma_0^2 \sum_i^{nsnp} p_i q_i$. Factoring out σ_0^2 from the formula, we end up with the famous identity (Gianola et al. 2009; VanRden 2008).

$$\boldsymbol{\sigma_{\alpha 0}^2 = \frac{\sigma_u^2}{2\sum_i^{nsnp} p_i q_i}} \qquad \textbf{(1.14)}$$

Based on equation (1.14) the priori variance of the markers is defined as a function of the genetic variance of the population, a formula used constantly in most applications in GP.

In case we want to calculate the genetic variance explained by markers after fitting the model, then estimates $\hat{\alpha}$ for each marker are available. Hence, each marker i explains a variance $2p_i q_i \hat{\alpha}_i^2$. Therefore, the genetic variance contributed by each marker is not the same across all markers (Legarra et al. 2014). Also note that $2\sum p_i q_i \hat{\alpha}_i^2$ underestimated the total genetic variance since estimates $\hat{\alpha}_i$ are shrunken towards 0. In the following paragraphs, different estimators are presented.

1.4.2.6 Models of genomic prediction

As we mentioned above, the prediction ability of GP methods is influenced by the proportion of variance on the traits the SNPs can explain. Conceptually GP is a large p and small n (genotyped animals) scenario since the number of variables *p* is far larger than the number of *n* of observations. As a result, methods can have a large sampling variance and mean-squared error. To overcome this limitation, variables must be selected or restrictions on the solutions must be applied or sometimes both can be applied simultaneously. Here GP methods are divided into two main classes based on the fact if they estimate the marker effects or not:

1) SNP effect-based method
2) Genomic relationship-based method

1.4.2.6.1 SNP effect-based methods

SNP effect-based methods estimate marker-effects for all genome-wide markers simultaneously. For practical reasons, markers are assumed to be uncorrelated even if they are close. For instance, if two markers are in strong LD, they will likely show a similar effect after fitting the model but before fitting the model, we cannot say that their effects will be similar or not. The most common methods used in this class are:

1. Random Regression BLUP (RR-BLUP), SNP-BLUP

These methods assume that the marker effects are random ("random regressions") coming from a normal distribution with constant variance for all loci. As a result, all markers explain the same proportion of variance on the traits. The methods are described by the following mixed model:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z\alpha} + \boldsymbol{e} \qquad \textbf{(1.15)}$$

Where, y is the vector of pre-corrected phenotypes, X is the incidence matrix for the fixed effects in b, Z is a matrix of the marker genotypes, α is a vector of marker effects, e is the residual term. Although method does not estimate breeding values, u, these can be derived as linear combination of the SNP effects, $Z\alpha$.

2. Bayesian Methods

One possible problem of the methods of Regression is the assumption of homogeneity across the marker effects, that is all markers have a constant variance. Thus, assuming homogeneity may not be optimal if some markers are in LD with QTL while some others are not. To deal with these, methods such as the Bayesian ones perform variable selection and shrinkage on the effects simultaneously. In Bayesian methods variable selection and differential shrinkage of estimates of effects can be applied using priors other than the Gaussian. For example, heavy-tailed prior distributions or mixture distributions are used as the distribution of marker-effects allowing for some markers to contribute more to genomic variance than others. Here a family models called Bayesian Alphabet is briefly introduced which share the same likelihood function but differ on the prior (or else shrinking) used for marker effects (Lourenco 2017; de los Campos 2018):

2.1 BayesA and BayesB (Meuwissen et al. 2001)

BayesA uses prior on the marker effects corresponding to a student-t distribution (Gianola et al. 2009) which has the property of having "fat tails". It assumes that all SNPs have effect on the traits with the majority of markers have small effect and very few have large effect. Thus, different variances are assumed for each marker. Note that Bayesian methods are non-linear and likely to be affected by shrinkage, that is the small effects became even smaller and the big effects even bigger. A very common thought at the beginning of Genomic Evaluation was that there were not many QTLs. Hence, it was commonly assumed that many markers do not have effect because they cannot trach QTLs. This originated the method known as BayesB (Legarra et al. 2014). BayesB uses priors that are mixtures of a spike of mass at zero of a continuous density (e.g., t, or normal). In other words, make

the same assumption with BayesA but for a fraction of markers. The method states that a proportion (π) of the SNPs have no effect and $1-\pi$ have a non-zero effect.

2.2 BayesC (Habier et al. 2011)

BayesC combines properties of BayesB and SNP-BLUP but uses as a prior a normal distribution with unknown variance. Particularly, it assumes that markers are coming from a distribution with constant variance (as in SNP-BLUP) and assumes that some fraction π of markers have no effect (as in BayesB).

$$p(\beta_i|\sigma_\beta^2) = \begin{cases} N(0,\sigma_\beta^2) \text{ with probability } (1-\pi) \\ 0 \text{ with probability } \pi \end{cases} \quad (1.16)$$

Where β_i is the effect of each marker with variance, σ_β^2 has a scaled inverse chi-square distribution; $\sigma_\beta^2 \sim \chi^{-2}(\nu_\beta, S_\beta^2)$ with S_β^2 scale and ν_β degrees of freedom. Note that the advantage of BayesC over BayesB is that is much faster. If assume $\pi = 1$ BayesC becomes SNP-BLUP.

1.4.2.6.2 Genomic relationship-based methods

These methods use markers to infer relationships among individuals, quantifying the number of alleles shared between two individuals. Genomic relationships are identical by state (IBS) because they account for the probability that two alleles randomly picked from each individual are identical, independently of origin. On the other hand, Pedigree relationships are identical by descent (IBD) because they consider the shared alleles come from the same ancestor.

1. Genomic Best Linear Unbiased Predictor (G-BLUP, VanRaden 2008)

G-BLUP is one of the most widely used models in genomic prediction. The method is equivalent to SNP-BLUP but genomic breeding values ($Z\alpha$) are estimated instead of SNP effects. The basic assumption is that all markers explain the same amount of variance therefore the majority of SNP have a small effect and very few moderate to large effect. It assumes a genomic relationship matrix (GRM) instead of a conventional pedigree-derived numerator relationship (A). Breeding values are more accurately estimated based on GRM.

$$y = X\beta + Zg + e \quad (1.17)$$

where X and Z are design matrices, β is a vector of fixed effects, g is a vector of additive genetic effects for an individual and e is a vector of random residuals with variance σ_e^2 . It was assumed that $g \sim N(0, G\sigma_g^2)$ where G is the GRM and σ_g^2 is the additive genetic variance. VanRaden (2008) suggested that the matrix G can be established as follows:

$$G = \frac{(M-P)(M-P)'}{2\sum p_k(1-p_k)} \quad (1.18)$$

Where M is a genotypic matrix $n \times m$ with n for the number of individuals and m for the number of markers, p_k is the minor allele frequency (MAF) of ith marker and P is the matrix in which the kth column elements are $2p_i$. GBLUP is robust, fast and more suitable for polygenic traits.

2. Reproducing Kernel Hilbert Spaces (RKHS) regressions

RKHS methods are used for semi-parametric modelling in different areas of application. Gianola et al. (2006) suggested using this method for semiparametric genomic enabled prediction. Since then, it has been widely applied in plant and animal breeding. Kernel-methods are very popular in particular to predict non-additive effects and to handle complex multi-environment multi-trait models (Sousa et al. 2017; Cuevas et al. 2018). RKHS considers as the best method in plants by Reinoso-Pelaez et al. (2022). Kernel methods can apply relationship or similarity (distance) matrices. The method uses the Gauss kernel function to fit the following model:

$$\boldsymbol{y = Xb + K_h\alpha + \varepsilon} \qquad \textbf{(1.19)}$$

where α has a multivariate normal distribution with mean zero and covariance matrix $K_h\sigma_\alpha^2$; $\varepsilon \sim N(0, I_n\sigma^2)$; K_h is a kernel function that represents the correlation between individuals and is defined as.

$$\boldsymbol{K_h(x_i, x_j) = exp(-hd_{ij})} \qquad \textbf{(1.20)}$$

Where d_{ij} is the squared Euclidean distance between individuals i and j calculated based on their genotypes, h is defined as $h = {}^{2}/_{d*}$ and $d*$ is the mean of d_{ij}. Under a Bayesian framework the model can be solved using a Gibbs sampler or a mixed linear model (Wang et al. 2018).

1.4.2.7 Deep Learning

Deep learning (DL) or artificial neural networks (ANN), is a subset of machine learning (ML) algorithms that are used to solve complex problems without being explicitly programmed. The ANN are trained in such a way to find complex relationships between traits. Usually, ML algorithms leverage structured labeled (or not) data to make predictions using a prior information as input data from the user. On the other hand, DL can be applied to unstructured data such as text, images and it automates feature extraction removing some of the dependency on feature engineers. The function of DL mimics and is inspired by the structure and function of human brain through a combination of data inputs, weights, and bias (Figure 1.6). In the early 1940´s Warren McCulloch a neurophysiologist worked together with logician Walter Pitts to create a model of how brain works. It was a simple linear model that produced a positive or negative output, given a set of inputs and weights:

$$\boldsymbol{f(x, w) = x_1w_1 + \ldots + x_nw_n} \qquad \textbf{(1.21)}$$

where $f(x, w)$ is the output, x_n the inputs and w_n, the weights. This model of computation was called neuron because it tried to mimic how the core building block of the brain worked. The brain neuron contains dendrites which are extensions of the nerve and propagate the electrochemical stimulation received from other neural cells to the cell body (soma) of the from which the dendrites project. If the received electrical signal is enough to trigger an impulse, then it is sent along the axon and it passes to other neurons. Just like brain neurons, McCulloch and Pitt´s neuron received inputs and if these signals were strong enough, passed them on the other neurons. However, this model had a problem. It could not learn like the brain does. Frank Rosenblatt gave the solution to this problem

almost a decade later, developing an algorithm that could learn the weights to generate an output. That model is called perceptron which is the simplest form of an artificial neuron network (ANN).

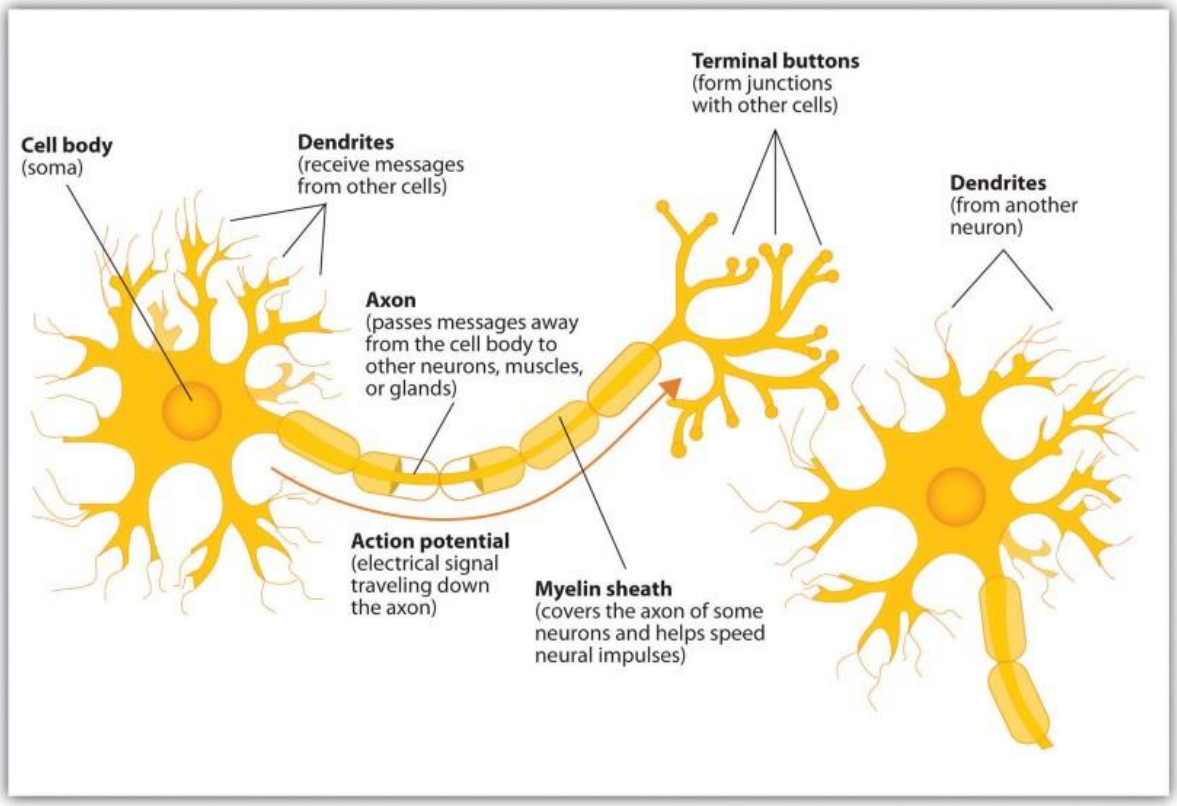

Figure 1.6: Brain neuron function Image used from Components_of_neuron.jpg (2046×1412) (wikimedia.org)

DL are classified in different categories based on their architecture with the most famous to be the Multilayer Perceptron network (MLP) and the Convolutional neural network (CNN). An MLP (Figure 1.7, A) has at least three layers: the input, a hidden layer, and the output layer. All the neurons are connected to every neuron in the previous layer and then connected to every neuron in the next layer by a weight that is assigned to each of them. After they receive the input that might be either the initial inputs or the output from other neurons, they make a decision of what to pass to the next layer of neurons. The layers between the input and output are referred to as “hidden layers”. DL networks consist of multiple layers of interconnected nodes. Each layer uses as input the output of the previous ones to optimize the prediction or classification. Neurons mathematically transform the data they receive before passing them forward. All these transformations allow the network to learn more complex relationships between the features and make predictions which other algorithms cannot easily discover.

A CNN (Figure 1.7, B) captures the spatial relationships of the data, that is the proximity and the position between the pixels can be taken into consideration. These networks have convolution layers

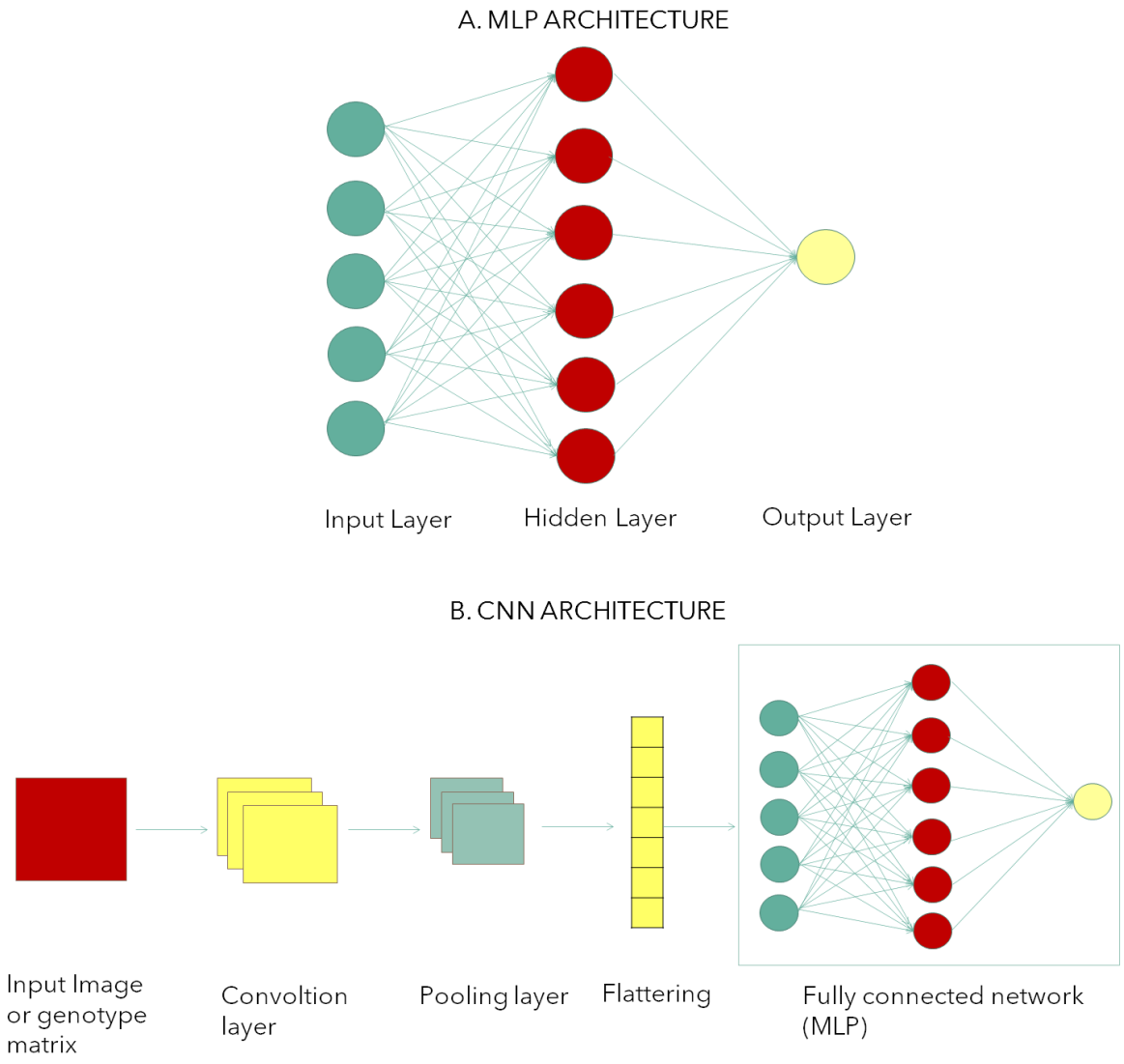

Figure 1.7: Figure depicts two DL architectures, Multi-Layer Perceptron (A) and Convolutional Neural Networks (B).

which are trying to capture the spatial patterns in the data. Particularly, convolution is the first layer to extract features from the input matrix. Convolution preserves the relationship between input variables by learning features using small squares of input data. It is a mathematical operation that takes two inputs, the input matrix and a filter or kernel. The filter or kernel is the learnable parameter of the network, that is the weight and is the same along the input matrix. The filter moves along windows of the input matrix of the same sizes consist of variables and performs a multiplication operation that is a dot product, until the entire matrix is transferred. The output matrix is called "feature map". After convolution is over, pooling layers are used to reduce the abundant information and keep the important one. The last layer of the networks is flattened out to be used as input in a fully connected network (MLP). Flattening is converting the data into a 1-dimensional array for inputting it to the next layer. We flatten the output of the convolutional layers to create a single long feature vector. The performance of the networks strongly depends on the hyperparameter choice, the dataset, and the size of the training sizes. Many tasks can be performed today by DL networks

like classification, prediction, image, and face recognition, forecast prediction robotics, Natural language processing, healthcare, and finally genomic prediction. A more detailed description about DL networks can be found in Chapter 5.

1.4.2.7.1 Deep learning in genomic prediction

In GP applications a genotype matrix is usually used as input in the network. Although DL has been applied and compared in various works seems to perform poorly under a GP framework (González-Recio et al. 2014; Ma et al. 2017; Bellot et al. 2018; Montesinos-López et al. 2018). However, Gianola et al. (2011) showed that MLP performed better than a Bayesian linear model in wheat. Pérez-Rodríguez et al. (2012) compared the prediction performance of Radia Basis Function Neural Networks and Bayesian Regularized Neural Networks against several linear models and semiparametric models such as (RKHS). The authors used wheat datasets and concluded that the non-linear models (DL, RKHS) had higher prediction ability than the linear models. Ehret et al. (2015) found non-relevant differences between a GBLUP and a MLP model. Similarly, Zingaretti et al. (2020) did not observe significant advantage of DL over linear models, except when epistasis component was important. DL outperformed GBLUP method when used to predict phenotypes from genotypes in wheat (Ma et al. 2017). An extensive review of studies in GP using neural networks can be founnd in (Montesinos-López et al. 2021).

The models for GP have been extensively compared. Meuwissen et al. (2001) compared four different statistical methods, observing that BLUP outperformed Least-square estimation (LS). Also, BayesA and BayesB increase prediction accuracy compared to GBLUP. Reviews of comparison of prediction models can be found in Heslot et al. (2012), Maltecca et al. (2012) and de los Campos (2013). In these comparisons was observed that when prediction is applied in close relatives and assuming a trait affected by many genes of small effect, the differences between the methods are small and methods like GBLUP, BLUP, and ridge regression are very robust. However, when traits have some larger QTL or when considering prediction of distant relatives, Bayesian and machine learning methods can be extremely effective (Robertsen et al. 2019).

1.5 Genomic prediction in rice breeding

Rice (Oryza sativa) was domesticated around 10,000 years ago and had developed into one of the most important food crops (Groen et al. 2022). Only in 2021/2022 about 509.87 million metric tons of rice was consumed worldwide up from 437.18 million metric tons in the 2008/2009 crop year (https://www.statista.com/statistics/255977/total-global-rice-consumption/). However, considering that the world population is increasing and at the same time we face a climate change, the conventional breeding techniques cannot meet the demand (Hickey et al. 2019). Conventional breeding needs almost ten years for a new cultivar to grow, meaning that the current pace of rice breeding is very slow. Consequently, we need methods that will secure nutritional requirements and at the same time increase the quality and quantity of yield. In addition, there is an imperative need for the new cultivars to have two significant traits, being disease resistant and climate smart. GP is an extremely important and efficient tool for achieving all the pre-mentioned

Table 1.3: Summary of prediction methods.

General method	Model	Distribution of marker effects	Software	Main functions
Parametric methods	RR-BLUP (BLUP)	Normal distribution	R/ rrBLUP, R/BGLR, R/sommer	Marker effect estimation, Genomic prediction, mixed model solver
	SNP-BLUP	Normal distribution		
	GBLUP	Normal distribution		
	BayesA	Student (t) distribution	R/PopVar R/BGLR R/BWGS	Genetic variance prediction, Genomic Prediction
	BayesB	Mixture of Student (t) distribution and spike at 0		
	BayesCPi	Mixture of Normal distribution and spike at 0	R/BGLR R/BWGS	Genomic Prediction
	BayesC	Normal distribution with unknown variances	R/qgg	Estimation of genomic parameters and genomic prediction
	GREML			
	Bayesian Lasso	Double exponential	R/glmnet R/STGS	
	Lasso	Laplace distribution		
	SSVS	Mixture of a large and small normal distribution	R/bvartools	
	Elastic-net	No priori	R/BWGS	Genomic Prediction
Non-Parametric methods	SVM		R/BWGS	
	RKHS		R/BGLR	
	RF		R/randomForest R/STGS	Genomic Prediction, Marker effect estimation
	RBFNN	--	R/rbf	Prediction
	DL		DeepGS	Genomic Prediction

*GREML: Restricted maximum likelihood estimation, DL: Deep Learning, SSVS: Stochastic Search Variable Selection (SSVS), SVM: Support Vector Machine, RF: Random Forest

Table 1.4: Summary of genomic selection studies in rice adapted by Xu et al. (2021).

Population	Genotype	Model	Trait (predictive ability)	Reference
110 Japanese cultivars	2071 SNPs	BL, EN, RF, GBLUP, wBSR, LASSO, RKHS	Flowering date (0.7–0.85), panicle length (0.5–0.7), panicle number (0.35–0.45), grain length (0.35–0.45), grain width (0.5–0.7)	Onogi et al. 2015
413 diversity inbred lines	36,901 SNPs	GBLUP	Florets per panicle (0.6), flowering time (0.6), plant height (0.7), protein content (0.45)	Isidro et al. 2015
386 inbred lines	1311 SNPs	PLS, Kernel PLS, RR Kernel RR	Grain shape (0.55–0.62)	Iwata et al. 2015
363 elite breeding lines	73,147 SNPs	BL, RKHS, RRBLUP, RF	Grain yield (0.15–0.31), flowering date (0.35–0.63), plant height (0.15–0.34)	Spindel et al. 2015
343 S2:4 lines	8336 SNPs	BL, BRR, GBLUP, LASSO, RRBLUP	Grain yield (0.31), flowering date (0.30), plant height (0.54), panicle weight (0.33)	Grenier et al. 2015
284 inbred lines and 97 F5-F7 lines	43,686 SNPs	GBLUP, RKHS, BayesB	Flowering date (0.35), nitrogen balance index (0.33), 100 panicle weight (0.38),	Hassen et al. 2018
128 Japanese cultivars	42,508 SNPs	GBLUP, PLS	Grain weight distribution (0.28–0.53)	Yabe et al. 2018
161 African accessions and 162 USDA accessions	36,901 SNPs	GBLUP, BayesA, BayesC	Rice blast (0.15–0.72)	Huang et al. 2019
210 recombinant inbred lines and 278 hybrids	1619 bins	GBLUP, LASSO, SSVS	Grain yield (0.31–0.36), grain number (0.59–0.61), tiller number (0.45–0.48), 1000 grain weight (0.82–0.83)	Xu et al. 2014
120 inbred lines and 575 hybrids	2,395,866 SNPs	GBLUP	Grain yield (0.39), grain number (0.64), plant height (0.86), 1000 grain weight (0.88)	Wang et al. 2017
120 inbred lines and 575 hybrids	116,482 SNPs	BayesB, GBLUP, PLS, LASSO, SVM, RKHS	Grain yield (0.38–0.41), grain number (0.64–0.65), plant height (0.86), 1000 grain weight (0.87–0.88)	Xu et al. 2018
1495 hybrids derived from incomplete NC II design and 100 hybrids derived from half diallel crosses	102,795 SNPs	GBLUP	Grain yield (0.54), grain number (0.62), plant height (0.58), 1000 grain weight (0.54)	Cui et al. 2020
738 accessions from five groups in rice, AUS/Boro, Indica, Aromatic, Admixed, Japonica	228,871 SNPs 52,120 MITE/DTX 21,571 RLX/RIX	RKHS, BayesC	Culm Diameter (0.26 0.40), Culm strength (0.28,0.16), Flag leaf angle (0.45,0.28), Grain length (0,69,0.66), Grain width (0.83, 0.64), Leaf length (0.41 0.52), Leaf senescence (0.47,0.54). Grain weigh (0,30,0.14), Salt injury (0,28,0.49), Time to flowering (0.65,0.73), Panicle threshability (0.29, 0.24)	Vourlaki et al. 2022

requirements, accelerating the breeding project. In rice GP can be used for inbred selection as well as hybrid breeding.

The main GP framework in rice studies includes a training population based on which the prediction takes place and the evaluation of the predictive ability within and between populations using a testing population. GP has been performed in rice for predicting various quantitative traits, while moderate to high predictive ability has been reported (Xu et al. 2021). Onogi et al. (2015) performed GP over six different rice traits using nine different methods. Using a diverse population of 413 rice inbred lines from 82 countries, Isidro et al. (2015) compared five different sampling strategies with stratified sampling resulting in highest predictive abilities over four different traits.

Huang et al. (2019) showed that GP was very helpful for predicting rice blast. GP is an effective tool, not only for the selection of pure lines in rice breeding but also in hybrid breeding, contributing to overcome limitations caused by the numerous potential crosses. GP can predict the performance of all combinations of a given set of genotyped parents. From these crosses only a small proportion is required to be evaluated in the field. Using GBLUP method, Xu et al. (2014) predicted hybrid performance of rice for first time. Particularly, they randomly paired 278 crosses from 210 recombinant inbred lines and predicted the remaining 21,667 untested hybrids. However, compared to other major crops such as maize and wheat, the studies focusing on applying GP to rice breeding practice are still limited. Table 1.3 shows a summary of GP models and reported values of predictive ability in rice breeding, and Table 1.4 displays a summary of the GP studies in rice adapted by Xu et al. (2021).

1.5.1 Incorporating new genetic markers into the plant breeding

Most studies focusing on the genetic variability at the whole-genome level in plants have been concentrated in SNPs as the main type of genetic variability (3KRGP 2014). However, other relevant sources of genetic variation can be found in the genomes. Transposable elements (TEs) and Structural variations (SVs) have been shown to form an important fraction of genetic variation in plant species playing a significant mutational role in crop domestication and breeding. Genetic difference caused by SVs and TEs can lead to phenotypic variations in a species. Hence, the use of these kinds of variation in GP could lead to improve the predictive ability of traits of interest. Here, we analyze the importance of Transposable elements in plant breeding as well as of Structural variations.

1.5.1.1 Transposable elements

TEs, also known as “jumping genes” or transposons, are sequences of DNA that move from one location in the genome to another. TEs represent the largest fraction of “junk DNA”, that is, DNA fragments without an obvious protein-coding or regulatory functional relevance for the organism (Dubin et al. 2018). They were first discovered by Barbara McClintock in the 1940s who was awarded a Nobel prize for this discovery. McClintock suggested that these mysterious mobile elements of the genome might play some kind of regulatory role, determining which genes are turned on when this activation takes place (McClintock 1965). Following McClintock research, scientists Roy Britten and Eric Davidson further proposed that TEs not only play a role in regulating gene expression but also in generating different cell types and different biological structures based on where in the genome they insert themselves (Britten and Davidson, 1969). Despite the groundbreaking research of scientists such as McClintock, Britten and Davidson, only recently the scientific community started to understand the importance of TEs as source of genetic variation (Pray 2008).

1.5.1.1.1 TEs come in many different forms and shapes

TEs come in a variety of forms and shapes because of their deep evolutionary origins and continuous diversification. TEs can be divided into two major classes based on their mechanism of transposition and each class can be further subdivided into subclasses based on the mechanism of chromosomal integration. The TEs belong to the Class I are known as retrotransposons, "jumping" through the mechanism of "copy-and-paste", whereby an RNA intermediate is reverse-transcribed into a DNA copy that is integrated elsewhere in the genome (Figure 1.8, Bourque et al. 2018). For long terminal repeat (LTR) retrotransposons, integration occurs by means of a cleavage and strand-transfer reaction catalyzed by an integrase much like retroviruses (Brown et al. 1987). For the case of non-LTR retrotransposons, including both long and short interspersed nuclear elements (LINEs and SINEs), chromosomal integration is coupled to the reverse transcription though a process referred to as target-primed reverse transcription (Luan et al. 1993). The most abundant superfamilies of Class I found in rice are RLX (LTR retrotransposons) and RIX (LINEs, NON—LRT retrotransposons). Class II elements are known as DNA transposons, and they are mobilized via the mechanism of "cut-and-paste" or "peel-and-paste". Particularly, CLASS II elements are jumped via a DNA intermediate, either directly through the mechanism of "cut-and-paste" (Figure 1.8, Greenblatt et al. 1963) or in the case of Helitrons, a "peel-and-paste" (Grabundzija et al. 2016) replicative mechanism involving a circular DNA intermediate. In rice, the most representative superfamilies of CLASS II are MITEs (Miniature inverted-repeat transposable elements) and the DTX (DNA TEs with terminal inverted repeats).

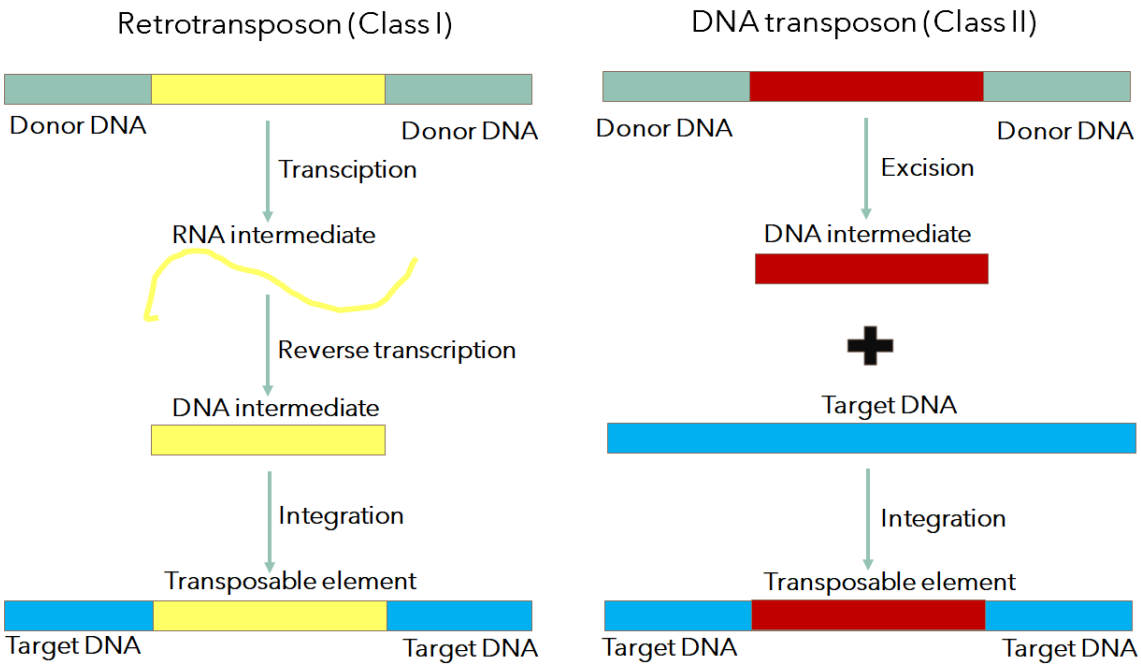

Figure 1.8: Different classes of TEs based on mechanism of transposition.

1.5.1.1.2 TEs are significant source of mutations and genetic polymorphisms

A substantial portion of the genome of a species is occupied by TEs including a large fraction of the DNA unique to that species. Particularly, in maize 60 to 70% of the genome is consist of LTR

retrotransposons, while the most of these are unique to this species or its close wild relatives being the most active and mutagenic in the genome (Schnable et al. 2009). In rice the percentage of TEs in the genome varies from 18.4% to 37.5% (Li et al. 2017) while in *A. thaliana* TEs make up only 10% of the genome (Arabidopsis Genome Initiative, 2000). The content of TIPs in the genome but also the proportion of TE classes depends on species and variety (Dubin et al. 2018). In rice DNA transposons are four times more than retrotransposons (Song et al. 2017). The vast majority of TE insertions in *D. Melanogaster* are absent at the orthologous site in its closest relative *D. Simulans,* while the most are not fixed in the population (Kofler et al. 2015). In rice, MITE insertions are present at high frequencies while being absent from its wild ancestor, indicating that they have been fixed at the same time with domestication. Also, MITE insertions present at low frequencies among rice varieties have transposed after domestication. Furthermore, it seems that MITE target gene-rich regions for integration (Castanera et al. 2021). In the same study, Castanera et al. (2021) discovered association between Transposable insertion polymorphisms (TIPs) that were not detected with SNPs. In melon, TEs may be the origin of an important fraction of the variability, in addition to the variation due to SNPs and SVs. Approximately 60% of the polymorphic TEs are present in only one variety, indicating that they have had an important activity during the recent melon evolution (Sanseverino et al. 2015). In humans, two haploid genomes differ by approximately a thousand TE insertions primarily from L1 or Alu families. It is well established that TEs account for at least 50% of the human genome (Bourque et al. 2018). TEs insertions rarely provide an immediate fitness advantage to their host and those reaching fixation are driven mainly by genetic drift while are subsequently eroded by point mutations that accumulate neutrally (Lynch 2007). It is interesting though that even TEs don't bring immediate benefit to their host and are largely decaying neutrally once inserted, they persist in evolution. One explanation of this ability of TEs is the fact that not only propagate vertically but also horizontally between individuals and species. Many studies provide evidence of the idea that horizontal transposon transfer is a common phenomenon that affects virtually every major type of TE and all branches of the tree of life (Gilbert et al. 2018).

1.5.1.1.3 TEs importance in plant evolution

The genomes of most eukaryotes as plant species are dominated by TEs which are now known to have a major role in driving genome evolution. The most abundant TE classes in plant genomes are LTRs and MITEs. Over the last years it has become known that certain TE families can experience bursts of activity under stress conditions and other environmental stimuli resulting in new TE insertions. The high transposition activity can result in altered gene expression patterns and phenotypes. The connections between TE- mediated increases in diversity and an accelerated rate of genome evolution provide powerful mechanisms for plants to adapt more rapidly to new environmental conditions. Considering that plants are sessile organisms with low migration capacity, the survival of native populations undergoing climate changes relies on evolutionary responses. Many TEs families are responsive to environmental cues, apparently integrate preferentially within genes involved in the environmental response and generate large-effect mutations, some of which are potentially adaptive (Baduel and Quadrana 2021). Although most TE insertions are highly deleterious, some can provide key adaptive variation (Baduel and Quadrana 2021). The important role of TEs as a source of genetic variation in plants is highlighted by the massive changes in TE abundance and

diversity that occurred during domestication and as a result of breeding efforts (Dubin et al. 2018). New TEs insertions are triggered by stress conferring new transcriptional responses to the target genes. This provides an additional source of variation on which selection can act quickly evolve phenotypes adapted to the stress. Transposition activity as constant source of new and major-effect mutations can be proved significantly useful for rapid adaptation. This ability to artificially boost TE activity supplied an extra source of variation for breeding. The fact that selection acts on TEs during local adaptation, speciation, domestication, and breeding (Dubin et al. 2018) highlights the essential adaptive role of these elements. It seems that understanding the important role of TEs as source of genetic variation and incorporating them in plant breeding programs could provide us a highly advantage on predicting new phenotypes adapted to drastic environmental changes such as ongoing global warming. Studies in rice and tomato have shown that TEs can reveal significant association with traits that are not detected with SNPs since TEs can be recent insertions and may not be in high LD with the surrounding SNPs. Also, the fact that TEs have been shown to mediate large phenotypic changes in a number of studies (Daborn et al. 2002; Butelli et al. 2012) indicates the need to apply then in plant breeding programs. Chapter 4 of this thesis aims to study whether the incorporation of Transposable insertion polymorphism (TIPs) in rice GP can increase the prediction ability compared to SNPs.

1.5.1.2 Structural variation

Genomic structural variation is the variation in structure of an organism's chromosome that can lead to gene loss, gene duplication and the generation of novel genes. Particularly, structural variations (SVs) are genomic polymorphisms that could originate from insertion, duplication, deletion, translocation, or inversion (Figure 1.9, Alkan et al. 2011). SVs can lead to polymorphisms affecting the gene content called copy-number variations (CNVs) and presence-absence variations (PAVs). SVs are considered to be longer (> 50bp) and can have a greater influence on gene expression and protein function than SNPs (Chiang et al. 2017). These types of SVs have been shown to be frequent in plant species (Saxena et al. 2014). Also, it has been established that such SVs have been associated with a diversity of phenotypes for major traits in plants (Sutton et al. 2007; Cook et al. 2012).

1.5.1.2.1 SVs in plant evolution

In the last years, studies of structural variation in plants have been increasing, extending our knowledge of genomic changes during evolution, domestication and breeding. Montenegro et al. (2017) identified PAVs associated with important agronomic traits such as environmental stress and defense response from 18 wheat cultivars. Golicz et al. (2016) discovered that SVs affected the presence of flowering time genes such as FLOWERING LOCUS C (FLC) in *Brassica oleracea*. In another study in tomato, Gao et al. (2019) found 4873 genes demonstrating PAV and identified a rare allele deletion associated with the flavor of tomato. Fuentes et al. (2019) showed that rice genome regions with frequent SVs were enriched in stress response genes. They also demonstrated how SVs could help in finding causative variants in genome-wide association analysis. The important role of SVs in fungal evolution and adaptation was highlighted by (Gorkovskiy et al. 2021). The majority of

examples of adaptive structural variation correspond to Single nucleotide variants (SNVs) but others can be attributed to transcriptional changes of the genes located within or near the SV event. SV plays an important role in genetic diversity in plants and in phenotypic variation. The limitations associated with the technology and methods used to analyze SVs did not improve our understanding of these variations. However, the current advances in DNA sequencing and optical mapping have increased the resolution of SVs identification. Nowadays, an increasing number of plant studies use SVs demonstrating the importance of SVs compared to SNPs and small indels (Wellenreuther et al. 2019). Furthermore, SV-specific genome-wide association study approaches are needed to associate SVs with phenotypes (Yuan et al. 2021). To further benefit plant breeding extensive databases of crop genome sequences are needed that will not be mainly restricted to SNPs. Mining SV-related genes may provide a useful tool to breeders and crop researchers to produce improved varieties. In Chapter 5 of this thesis, SVs along with TIPs are applied in a GP framework using Deep learning networks aiming to investigate whether this strategy can improve prediction of complex traits in rice.

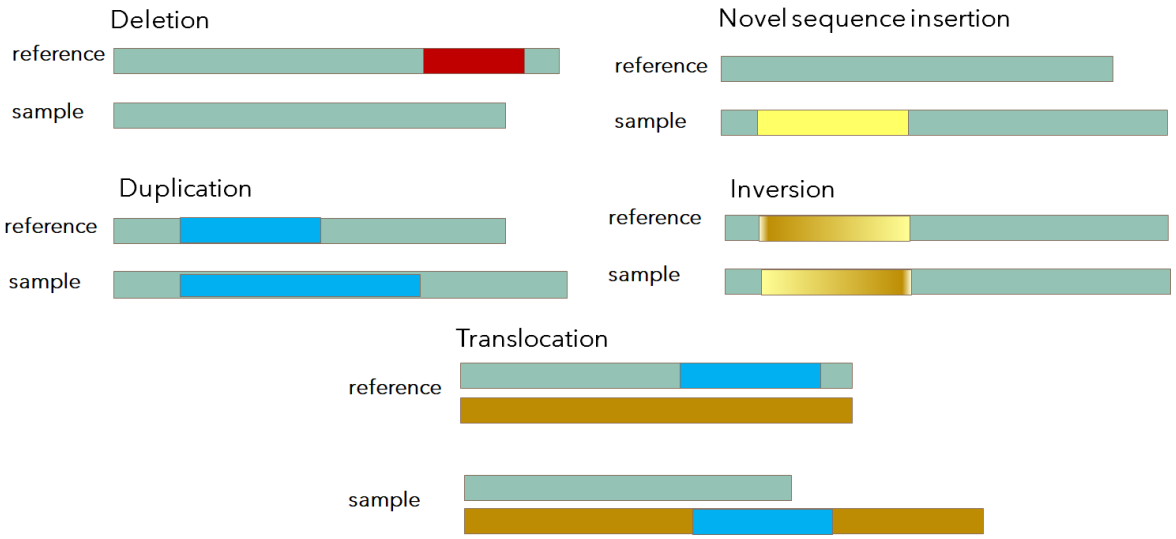

Figure 1.9: SVs are classified as a deletion, an insertion, a duplication, an inversion, or a translocation.

Chapter 2

Objectives

The general goal of this thesis is to detect signals of polygenic variability under two different frameworks, (i) from a population genetics perspective, trying to infer the effects of polygenic adaptation under domestication and (ii) from a quantitative genetics perspective, trying to predict the effects of polygenic variation on the phenotype using a plant breeding strategy.

The Specific objectives are:

1. Perform a simulation study on the effects of polygenic variability in a context of wild versus domestic populations and study the patterns of variability.
2. Infer the Distribution of Fitness Effects (DFE) of the simulated wild and domestic populations and compare their patterns to evaluate the capacity of detection of the polygenic adaptation.
3. Incorporate Transposable Insertion Polymorphisms (TIPs) and Structural Variants (SVs) in Genome Prediction (GP) for determining genetic potential in rice.
4. Evaluate the performance of deep learning in genomic prediction using Single Nucleotide Polymorphisms (SNPS), TIPS and SVs.

Chapter 3

Detection of Domestication Signals through the Analysis of the Full Distribution of Fitness Effects using Forward Simulations and Polygenic Adaptation

DOI: 10.1101/2022.08.24.505198 (under revision)

Abstract

As a consequence of the process of domestication, wild and domestic individuals are adapted to very different environmental conditions. Although the phenotypic consequences of domestication are observable, the genetic causes are not evident in many cases. Artificial selection could be modifying the selection coefficients of new and standing variation in the population under domestication. Here, we aim to detect a genome-wide signal of domestication under a model of polygenic adaptation. We use forward simulations to investigate the 1D and 2D site frequency spectra (SFS) of mutations in two populations (Wild and Domestic) with divergent histories (demographic and selective) following a domestication split. We simulate ten different scenarios, varying the strength of selection upon beneficial mutations and the proportion of mutations whose selection coefficients change after domestication. First, we describe that in domesticated populations selection at linked sites needs to be invoked to explain the SFS of neutral mutations and that the mode of linked selection affecting the neutral SFS depends on the duration of the domestication bottleneck. Second, we find that some aspects of the full distribution of fitness effects (DFE), such as the shape and strength of the deleterious DFE, are accurately estimated in both populations when using only the 1D-SFS. However, the detection of significant differences in the beneficial DFE between populations remains challenging in most, but not all simulated scenarios when only the 1D-SFS is used. Third, when considering the 2D-SFS and a new joint DFE model, we are able to detect more subtle differences in the full DFE that are hidden in the 1D-SFS analysis. In conclusion, our work highlights the strengths and limitations of detecting a polygenic signal of domestication under a variety of domestication scenarios and genetic architectures.

Introduction

The increase in human population size and the emergence of modern society are linked to the domestication of plants and animals (Purugganan and Fuller 2009; Driscoll et al. 2009; Larson and Burger 2013; Amills et al. 2017; Stetter et al. 2018). Human civilization as we know it was possible thanks to the domestication of surrounding life forms, where people started to domesticate plants and animals such as wheat, dogs, pigs, or chickens (Avni et al. 2017; Dayan 1994; Redding 2015; Zeder 2012). Domestication is a process that allows humans and other species to obtain a mutualistic long-term relationship that implies an advantage for both species (Zeder et al. 2006). The domestication of fauna and flora by humans started approximately 10-15 thousand years ago and is still in progress (Larson et al. 2014; Zeder 2015). Although human civilization relies on domestication, we still lack a full genomic and evolutionary understanding of domestication. Domestication is a rapid process in terms of its evolutionary time scale, but it is not a discrete event, and it implies the gradual improvement of domesticated traits. The artificial selection generated by humans during domestication can be considered relatively stronger and hence faster than natural selection because the selective pressure imposed by humans tends to be very extreme. Moreover, domestication tends to be associated with bottlenecks; only a small number of individuals from the wild population become domesticated, which is expected to decrease the efficiency of natural selection (Wright et al. 2005). Another important difference between natural and artificial selection is that modern breeders usually apply truncation selection, that is, the selection of the top percentage individuals for the

desired trait (Granleese et al. 2019). The prevalence of truncation selection in nature, or before the industrialization era, is unknown. Truncation selection is known to be an easy and efficient form of directional selection (Crow and Kimura 1979), and no severe accumulation of genetic load is expected in outcrossing species (Kondrashov 1988; Ohta 1989) if the population size remains large enough (Marsden et al. 2016). A recent comprehensive meta-analysis about the genetic cost of domestication (Moyers et al. 2018) found that harmful variants are more numerous (or segregate at higher frequencies) in domesticated populations compared to their wild relatives. This pattern is likely driven by a number of processes, which jointly decrease the efficacy of selection in domesticated populations as first suggested in rice genomes (Lu et al. 2006).

Selection, both natural and artificial, can occur either through a few loci with strong effects or via many loci with small effects (Jain and Stephan 2017a; b), depending on the genetic architecture of the trait and the strength of selection on it. Different patterns of genetic diversity around the selected loci are expected in response to these two models of selection (Stephan and John 2020). Classic hard selective sweeps have been reported in few candidate loci under important domesticated traits (Andersson 2012), such as the IGF2 gene region associated with lean domestic pigs (van Laere et al. 2003), the thyroid-stimulating hormone receptor (*TSHR*) in domestic chickens (Rubin et al. 2010) or the loci *sh4* and *qSW5* (Li et al. 2018b; Shomura et al. 2008), which are involved in the traits of seed shattering and grain width in domestic rice, respectively (Huang et al. 2012). These examples are consistent with a simple Mendelian genetic architecture where few loci determine most of the variance in the domesticated trait.

Polygenic adaptation, in contrast, describes a process in which a constellation of small changes in allele frequencies modifies the differences in the trait under selection. A wide range of population genetics models and simulations describing polygenic adaptation has been investigated (e.g., Stephan 2016; de Vladar and Barton 2014). Some models analyze the polygenic response of a trait in the presence of mutation and stabilizing selection (Jain and Stephan 2015; de Vladar and Barton 2014), while others capture the response of a trait under mutation and stabilizing or directional selection after an environmental shift in a finite size population (Stephan and John 2020). The interplay between hard selective sweeps and polygenic adaptation has been theoretically studied by Höllinger et al. (2019) by measuring how selection and genetic drift favour sweeps over small allele frequency shifts and vice versa using a finite population size. Thornton (2019) simulated the dynamic effect of selected and neutral mutations at a single quantitative trait, showing that the new optimum trait in a population of finite size is reached before selective sweeps are completed in polygenic models with many involved loci. Stetter et al. (2018) showed the impact of different demographic conditions, such as population bottlenecks or exponential growth, on domestication in maize via polygenic adaptation and for traits with major effect loci. They performed an exhaustive simulation analysis using a limited number of QTLs under an additive model and observed the presence of selective sweeps, even for small effect size mutations, after sudden environmental changes in traits under stabilizing selection. They concluded that the effect size of new mutations, as well as demography, are the main parameters influencing the observed genetic architecture. In practice, polygenic adaptation is harder to detect than classic selective sweeps (Pritchard et al. 2010). However, polygenic adaptation has been detected in some particular studies in wild and domesticated populations. Genome-wide association studies (GWAS) combined with population genetic modelling could potentially lead to the

detection of signals of polygenic adaptation (Berg and Coop 2014). Combining gene association and multivariate (redundancy analysis) methods have contributed to the identification of polygenic adaptation in the threatened fish species *Maccullochella peelii* (Harrisson et al. 2017). In *Drosophila*, Evolve and Resequencing (E&R) experiments that modified the regime of the temperature of populations have detected a polygenic adaptive response that could be explained by functional redundancy and quantitative traits (Barghi et al. 2019). Under climate change, the coral ecological divergence has been explained by the polygenic adaptation of many amino acid-changing variants (Rose et al. 2018). Small effects of polygenic variants seem to explain most of the existing variation in maize domesticated traits (Xue et al. 2016). It has been suggested that in animals, the initial domestication process may have acted on complex behavioural traits such as taming (Trut et al. 2009), which others have found to be affected by a large number of genes (Jasinska and Freimer 2009).

In this study, we wonder to what extent we are able to detect a genomic signal of domestication using a different approach: the comparison of the full distribution of fitness effects (DFE) on new and standing variation. To do that, we first compared the 1-dimensional (1D) and 2-dimensional (2D) unfolded site frequency spectra (SFS) of synonymous and nonsynonymous variants between domesticated and wild *in silico* populations. Second, we compared the inferred full DFE of new nonsynonymous mutations in those populations. The DFE of new deleterious mutations has been previously estimated contrasting the 1D-SFS of synonymous and nonsynonymous mutations of a multitude of species assuming that beneficial mutations only contribute to divergence but not to polymorphism because of their rapid fixation in the population (Barton and Zeng 2018; Boyko *et al.* 2009; Keightley and Eyre-Walker 2007; Kim et al. 2017; Tataru et al. (2017) proposed a model, polyDFE, for the inference of the full DFE and the proportion of adaptive substitutions (α) by using polymorphism data exclusively. Castellano et al. (2019) used polyDFE to compare the full DFE of new amino acid mutations across great apes, finding that the shape of the deleterious DFE is constant across this set of closely related species. Recently, using the 2D-SFS, a new method to jointly estimate the full DFE between two populations that have recently diverged and share many polymorphisms has been proposed (Huang et al. 2021). However, there are few studies comparing the DFE between domesticated and wild populations (Leno-Colorado et al. 2020). Here, we use forward simulations, considering the domestication process under different demographic and selective models. Several combinations of genetic architectures and selective effects have been simulated, from one considering a relatively small number of loci changing their selective effects to another of polygenic adaptation where many loci have divergent selective effects. In all cases, the selective effects of a proportion of existing variants can change (from deleterious to beneficial, and vice versa) in the domesticated population.

Materials and Methods

Simulation of the Domestication Process with Selection and Demographic Changes

A simulation analysis of the domestication process is developed using the forward-in-time simulator SLiM2 (Haller and Messer 2016). This tool is very versatile and allows the introduction of a number of variable scenarios and parameters. The general model for the domestication process is

developed in the SLiM script in the Zenodo database (10.5281/zenodo.7017885). Ten different scenarios of domestication are analyzed, and the parameters for each of the scenarios are shown in Table 3.1. All the options (flags) used for running the SLiM script are shown in a file in Zenodo (10.5281/zenodo.7017885). Briefly, the constructed model starts from a single panmictic population of Ne = 500 diploid individuals, with a genome containing 10 independent chromosomes, each with 1000 loci of 1500 base pairs of length, and each locus having one-third (4-fold) neutral synonymous positions and two-thirds of (0-fold) selected nonsynonymous positions interspersed along the locus. Each locus is separated from each other at different recombinational distances, following a convex curve in which the loci located at the telomeres are at a longer recombinational distance between loci (2.5e-2), while the loci near centromeres are closer in the recombinational distance (2.5e-4). The recombination within loci is fixed to a rate of 2.5e-7 between positions. Given the high computational burden of simulating entire chromosomes using variable recombination values and a considerable number of coding sites (1,500,000 coding sites per chromosome and 10 independent chromosomes) and the large number of mutations obtained, we perform a single run for each of the 10 scenarios. Note that the higher recombination among loci aims to mimic their real genetic distance separation but severely speeds down the simulation.

The demographic parameters for each scenario (Figure 3.2) are as follows: the initial blank population run for 10**Ne* generations to reach mutation-selection-drift equilibrium, then splits into two equal populations (outgroup population and the wild population), and after 10**Ne* generations, the wild population splits again into domestic and wild populations. Hereafter we refer to the Wild and Domestic populations. We aim to mimic a realistic animal domestication process, such as pig, where ancestral *Ne* estimates were around 10,000 (Groenen et al. 2012) and the domestication process occurred around 10,000 years ago (Zeder et al. 2006). The generation time was here assumed at 3 years per generation (Zhang et al. 2022). Two very different conditions for the bottleneck process are studied. The Domestic population suffers a bottleneck, reducing its population size temporarily to 50 diploid individuals, to recover again to *Ne* diploid individuals after the bottleneck. The bottleneck elapsed either 2*Ne**0.016 generations for scenarios with a short bottleneck or 2*Ne**0.161 generations for long bottleneck scenarios. Moreover, the simulation finished either in 2Ne*0.15 generations after the initiation of the short bottleneck process or 2Ne*0.005 generations of the long bottleneck. The selective effects produced by domestication are modelled by changing the fitness values of a proportion of the existing and new mutations in the domestic population (at the time of the split).

Most new nonsynonymous mutations occurring in ancestral and Wild populations are under negative selection (97.5%), and only 2.5% of the mutations have positive fitness effects. Domestic populations show different proportions of (new and standing) beneficial and deleterious variation depending on the scenario (Table 3.1). The negative effects in all scenarios and populations follow a gamma distribution with a shape value of 0.2 and a mean of S_d = -10 ($2N_e s_d$ when homozygote), while variants with positive effects follow an exponential distribution with a mean S_b = 1 or S_b = 10 ($2N_e s_b$ when homozygote) depending on the scenario (Table 3.1).

Types of Sites

The sites are initially divided into seven different types (named m_1 to m_7), being m_1 neutral (synonymous) and m_2 to m_7 functional (nonsynonymous) sites having a different selective effect when mutated (see Table 3.2). Mutations at m_5, m_6 and m_7 sites generate deleterious variants in the Wild population, and mutations at m_2, m_3 and m_4 sites generate beneficial mutations in the Wild population. The selection coefficient of mutations generated at m_2 (beneficial) or at m_5 (deleterious) sites are invariant for the Wild and Domestic populations. However, the mutations at m_3, m_4, m_6 and m_7 sites will change their selective effect in the Domestic populations relative to the Wild populations. That is, the new selective effect is drawn from the corresponding DFE section (positive or negative), independently of their value in the wild population. Those can be understood as sites with divergent selective effects. The selection coefficient of a given beneficial mutation at m_3 sites will remain beneficial in the Domestic population, but it will be different from the original beneficial effect at Wild. A mutation at m_4 sites will change its selection coefficient from beneficial in the Wild to deleterious in the Domestic population. Equivalently, the selection coefficient of a deleterious mutation at m_6 sites will remain negative in the Domestic population but it will be different from that found at Wild. A mutation at m_7 sites will change its selection coefficient from deleterious in the Wild to beneficial in the Domestic population. This hard-coding of selective effects on different sites allows us to gain insight into the relative importance of each mutation type for the domestication process.

Type of variants

The variants are classified into total (all observed variants), exclusive (variants that are present in a single population) and shared (variants that are present in both the Wild and Domestic populations). Note that exclusive and shared variants are not exactly coincident with new and standing variants, that is, new variants are those mutations that appear after the split between Wild and Domestic populations while standing variants are variants that appeared before the Domestication split. However, a new variant happening in the Wild population can be shared if a migration event transferred this variant to the Domestic population, and a standing variant can become exclusive if this variant disappears in one of the populations (e.g., Lee and Coop 2017). The proportion of the different types of sites, polymorphisms and substitutions across scenarios can be found at Supplementary Figure A.0 and Table A.0.

Simulating the Domestication Process only with Demographic Changes

To disentangle selective from demographic effects in our comparative analysis of the SFS we used the *ms* (Hudson 2001) coalescent simulator to simulate samples according to demographic parameters under short and long bottleneck periods, with or without migration (Figure 3.2). The demographic parameters used here are the same as the ones used for simulating the different scenarios described above with SliM2. However, the parameters are re-scaled according to software requirements (that is, considering *Ne,* see the scripts at Zenodo). Here, the size of the simulated coding regions is 2000 bp. *ms* outputs are processed using an R script (10.5281/zenodo.7017885) to calculate the site frequency spectrum for total, shared, and exclusive variants, divergence, and the estimates of genetic diversity.

Genetic Diversity Summary Statistics and the Fraction of Adaptive Substitutions (α)

The estimates of genetic diversity, population differentiation and divergence, and the parametric inference of the DFE and demographic patterns are computed in a sample of 20 diploid individuals (n=40 haploid chromosomes) per population. The 1D and 2D SFS are polarized using the simulated outgroup population. The number of mutations occurring in each population and the variants that become fixed and remained polymorphic in one or both populations is recorded. The fraction of adaptive substitutions (α) is estimated per population using the estimation methods proposed by Smith and Eyre-Walker (2002) and by the asymptotic McDonald & Kreitman test method (MKT, Messer and Petrov 2013), which corrects for the underestimation of α due to slightly deleterious mutations. The asymptotic MKT is computed using the web service available at http://benhaller.com/messerlab/asymptoticMK.html. Here, the default cutoff interval of x [0.1, 0.9] is used. PolyDFE (Tataru and Bataillon 2019) α estimates are computed using only polymorphisms counts to estimate the negative DFE (using model f1 from Table 3.5). Then, the expected number of non-adaptive nonsynonymous substitutions is contrasted with the observed number of nonsynonymous substitutions to estimate α. We also estimated α not using substitutions, but these estimates turn out to be very noisy (data not shown). Since we know the selection coefficients of mutations before and after the domestication split, all these summary statistics are investigated for different types of sites. The estimates of variability per site using shared (and exclusive) variants are calculated considering the total of positions under study (e.g., synonymous or m_1, and nonsynonymous or m_2…). That is, the sum of shared plus exclusive variability is equal to the total variability for such type of positions.

Distribution of fitness effects (DFE): Two complementary approaches

polyDFE: 1D-SFS

We use the polyDFEv2.0 framework (Tataru and Bataillon 2019) to estimate and compare the DFE across Wild-Domestic population pairs by means of likelihood ratio tests (LRTs). We use the R function *compareModels* (from https://github.com/paula-tataru/polyDFE/blob/master/postprocessing.R) to compare pairs of models. The inference is performed only on the unfolded SFS data (divergence counts to the outgroup are not fitted), and unfolded SFS data are fitted using a DFE model comprising both deleterious (γ-distributed) and beneficial (exponentially distributed) mutations. Note that in polyDFE *s* is defined to be the selection coefficient on the heterozygote (like in *dadi*), but the scaled selection coefficient is defined as $4N_e s$. The DFE of each Wild-Domestic population pair is inferred using the 1D-SFS of each population. polyDFE assumes that new mutations in a genomic region arise as a Poisson process with an intensity that is proportional to the length of the region and the mutation rate per nucleotide (μ). We assume that μ remains constant across simulations (as it is the case). Both an ancestral SNP misidentification error (ε) and distortion parameters (r_i) are estimated. However, we notice that the exclusion of ε does not affect the rest of estimated parameters because under the simulation conditions few sites are expected to be misidentified. The r_i parameters are fitted independently for each frequency bin (from n = 1 to n = 39), and they are able to correct any distortion that affects equally the SFS of synonymous and nonsynonymous variants (such as, in principle, demography or linked selection). To obtain the sampling variance of parameter estimates and approximate confidence intervals, we use a bootstrap

approach. Bootstraps are generated by resampling the SNPs in chromosomal "chunks" or segments 100 times using an ad hoc R script (zenodo link). These segments are non-overlapping and are 100,000 bps long. Hence, we assume that SNPs are only independent across segments but not within segments. Model averaging provides a way to obtain honest estimates that account for model uncertainty. To produce the model average estimates of the full DFE we weight each competing model according to their AIC following the equation 6.1 shown in the polyDFEv2 tutorial. We use the R function getAICweights (from https://github.com/paula-tataru/polyDFE/blob/master/postprocessing.R to do the model averaging R) to obtain the AIC values.

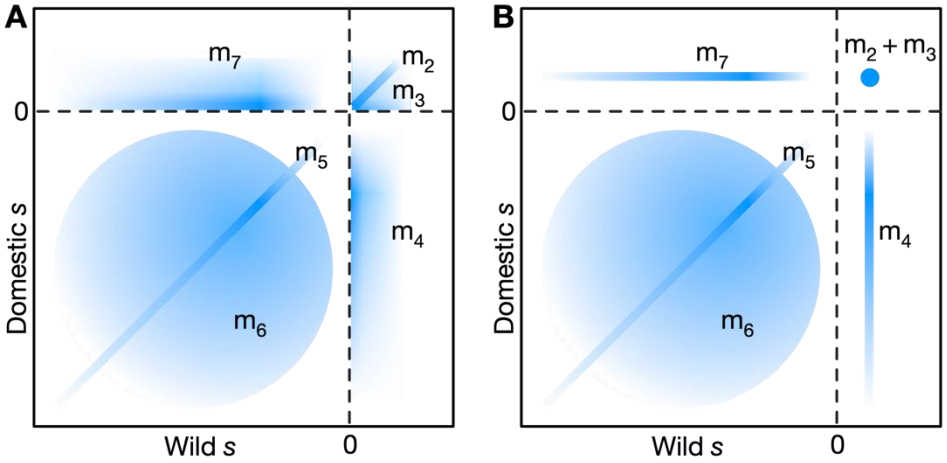

Figure 3.1: Joint DFE models simulated and fit. **A:** Illustration of the joint DFE model used in the SLiM simulations, with mutation types illustrated. **B**: Illustration of the joint DFE model used in the dadi inferences, in which a fixed positive selection coefficient is assumed.

dadi: 2D-SFS

dadi (Gutenkunst et al. 2009) is employed to infer the joint distribution of fitness effects (Jerison et al. 2014; Ragsdale et al. 2016; Huang et al. 2021) and the demographic history of all simulated population pairs. Our new model for the joint DFE between the two populations is a mixture of multiple components designed to mimic the selected mutation types in the simulations (Table 3.2; Figure 3.1). The major exception is that beneficial mutations are modelled to have a single fixed selection coefficient, rather than arising from a distribution. Let p_{+w} be the fraction of mutations that are positively selected in the Wild population, p_c be the fraction of mutations that change selection coefficient in the Domestic population, and p_{c+} be the fraction of those mutations that become beneficial in the Domestic population. To model mutation types m_2 and m_3, a proportion p_{+w} $(1\text{-}p_c)$ + p_{+w} p_c p_{c+} of mutations are assumed to have the same fixed positive selection coefficient in both populations. To model m_4, a proportion p_{+w} p_c $(1\text{-}p_{c+})$ are assumed to have a fixed positive selection coefficient in the Wild population and a gamma-distributed negative selection coefficient in the Domestic population. To model m_5, a proportion $(1\text{-}p_{+w})(1\text{-}p_c)$ of mutations are assumed to have equal negative gamma-distributed selection coefficients in the two populations. To model m_6, a proportion $(1\text{-}p_{+w})$ p_c $(1\text{-}p_{c+})$ are assumed to have independent gamma-distributed selection coefficients in the two populations. To model m_7, a proportion $(1\text{-}p_{+w})$ p_c p_{c+} mutations are assumed to have a gamma-distributed negative selection coefficient in the Wild population and a fixed positive

selection coefficient in the Domestic population. All gamma distributions are assumed to have the same shape and scale. This model is implemented as in *dadi* as the function dadi.DFE.Vourlaki_mixture. For inference, demographic models are first fit to neutral mutations from each simulation, then the new proposed joint DFE model is fit to the selected mutations. Demographic models (Figure 3.2) are estimated by running 100 optimizations per simulated dataset. Then, the 2D-SFS for selected sites are precomputed conditional on the demography for 104 values of γ ($2N_A * s$, a population scaled selection coefficient for the heterozygote where N_A is the ancestral population size), 102 negative and 2 positives. For the negative part of the DFE, γ values were logarithmically equally spaced between -2000 and -10^{-4}. The expected DFE for selected sites can then be computed as a weighted sum over these cached spectra (Kim et al 2017). The DFE parameters shape (α), scale (β), p_{+w}, p_c, and p_{c+} are then estimated by maximizing the Poisson likelihood of the simulated data, with the nonsynonymous rate of mutation influx θ fixed to twice that inferred for neutral sites in the demographic history fit. For the DFE inference, optimization is repeated until the best three results are within 0.5 log-likelihood units. Ancestral state misidentification is not modelled, because under the simulation conditions few sites are expected to be misidentified. Uncertainties of DFE parameter inferences are calculated by conventional bootstrapping, holding the demographic model fixed and dividing the simulated data into non-overlapping regions of 100,000 basepairs. Note that this procedure does not propagate uncertainty in demographic parameters through to the DFE parameters.

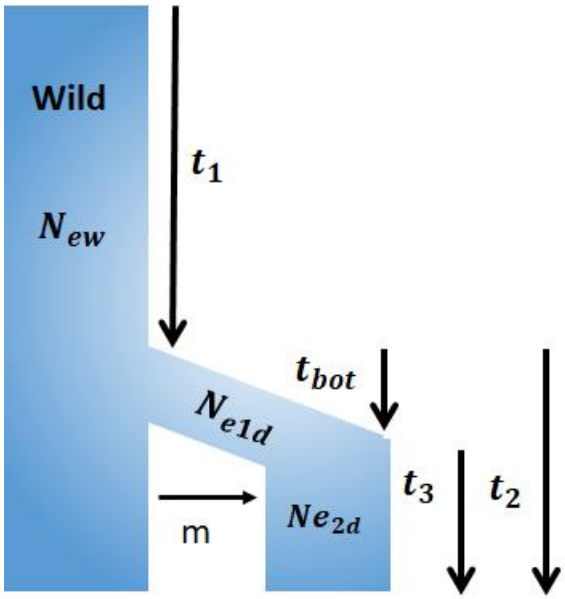

Figure 3.2: Diagram of the demographic model. New: Effective population size of the Wild population. Ne_{1d}: Effective population size of the Domestic population at bottleneck. Ne_{2d}: Effective population size of the Domestic population after a bottleneck. t_1: Number of generations from the Wild-Outgroup split to the Domestic-Wild split. t_{bot}: Number of generations in the bottleneck period. t_2: Number of generations from the bottleneck to the present. t_3: Number of generations from the end of the bottleneck to the present. m: Wild to Domestic migration rate from the bottleneck to the present (migration occurs along t_2).

Table 3.1: Parameters for each analyzed scenario.

Scenario	Bottleneck	Positive DFE1 (S_b)	Domestication		Positive DFE2 (S_b)	m W->D
	t_bot (2Ne)	**exp mean (2Nes)**	**% Change**	**% Positive**	**exp mean (2Nes)**	**(2Nem)**
1	0.016	10	0	2.5	10	200
2	0.016	1	5	25.0	1	200
3	0.016	1	25	10.0	1	200
4	0.016	1	0	2.5	1	0
5	0.016	1	25	25.0	1	0
6	0.016	10	25	2.5	10	0
7	0.161	10	0	2.5	10	0
8	0.161	1	5	2.5	1	0
9	0.161	1	25	25.0	1	0
10	0.161	10	5	10.0	10	0

Ne: Ancestral effective population size, S_b: Population selection coefficient of new beneficial mutations $2N_es$ in homozygotes, %change: percentage of sites that change their selection coefficient in the Domestic population, %Positive: from the percentage of sites that change their selection coefficient (% change) this is the percentage that change toward positive values in the Domestic population. Exp mean: Mean value of the exponential distribution for new positive derived variants in the Domestic population. m W->D: migration rate of Wild to Domestic population.

For the purpose of this work, *dadi* software is downloaded and installed according to the instructions provided at the following link: https://bitbucket.org/gutenkunstlab/dadi/src/master/. Since *dadi* operates as a module of Python, the Anaconda3 and Spyder (Python 3.7, Rossum and Drake 2006; Anaconda 2016, Raybaut 2009) versions are used in this study.

Results

Studying the effect of domestication on the DFE of natural populations can be very challenging, especially if the available methods for inferring and comparing the DFE have not been benchmarked using exactly the same dataset. Within the scope of the present work, we conduct a simulation study using different combinations of parameters relevant to the domestication process. We aim to investigate the ability to detect the effect of domestication in the SFS and the full DFE of a domesticated population that is experiencing a large or small change in the number and selective effects of loci under domestication/selection after a short or a long bottleneck period, with or without migration. Hereafter, we refer to the Wild and Domestic populations. The Wild populations have a constant DFE and constant population size. Beneficial mutations arise at Wild populations at a relatively low percentage (2.5%) following an exponential distribution, the rest of mutations are drawn from a gamma distribution with shape 0.2 and mean S_d=-10 (where $S_d=2N_es$, s is the selection coefficient in the homozygote, and N_e = 500 is the ancestral effective population size, see Material and Methods: Simulating the Domestication Process). All mutations, beneficial and deleterious, are co-dominant. The Domestic population originates from the Wild population through a bottleneck and

a concomitant change of the selective effects in a proportion of nonsynonymous sites (Figure 3.2; Table 3.1).

The change in selective effects impacts new mutations in the Domestic population and pre-existing variants that originated before the domestication event. In other words, not only mutations that were deleterious (or beneficial) before the split can become beneficial (or deleterious) in the Domestic population, but even when the sign of the selection coefficient remains the same, the strength of selection can be modified. See Table 3.2 for all the combinations of changes in selective effects between Wild and Domestic populations. The simulated scenarios aim to cover a variety of possible changes in the genetic architecture (number of loci) and the strength of selection (selection coefficients) of the trait/s under domestication. Two selection coefficients for beneficial mutations are assumed: (i) strong, with a mean population-scaled selection coefficient of S_b=10 and (ii) weak, with S_b=1 when homozygote. Depending on the scenario, a selective change occurs only at a few (5%) or at a large proportion (25%) of positions in the Domestic population (Table 3.1, "%change" column). We leave three scenarios (scenario 1, 4 and 7) as negative controls; the DFE in the Domestic and Wild populations is the same. Among those positions showing an alteration of fitness effects, we also vary the proportion that is beneficial (Table 3.1, "%Positive" column). In all scenarios, the fraction of positions (and polymorphisms) with beneficial mutations in the Domestic population is equal or larger than in the Wild population and varies across scenarios (Table 3.1, see Supplementary Figure A.0 for the exact site composition of the Domestic populations in each scenario). Moreover, demographic changes affect only the Domestic population. Three demographic models are simulated: (i) short bottleneck with migration, (ii) short bottleneck without migration and (iii) long bottleneck without migration. There is only migration from the Wild to the Domestic populations. Table 3.2 shows the different types of mutations according to their fitness in Wild and Domestic populations.

Table 3.2: Types of mutations in simulated scenarios.

	Wild	Domestic
m_1	Neutral	No change, remain Neutral
m_2	Beneficial	No change, remain Beneficial
m_3	Beneficial	Change to a different Beneficial effect
m_4	Beneficial	Change to Deleterious
m_5	Deleterious	No change, remain Deleterious
m_6	Deleterious	Change to a different Deleterious effect
m_7	Deleterious	Change to Beneficial

Descriptive Summary Statistics of the Simulated Populations

Approximately 120K mutations are observed under each of the scenarios in the Wild population, with variable numbers, between 53 to 60%, of polymorphic variants (Supplementary Table A.1A), while the rest are fixed (relative to the outgroup). The comparison among scenarios shows that those with a higher positive DFE mean (S_b=10, scenarios 1, 6, 7 and 10) exhibit a slightly higher proportion of fixations (~54K vs 47K) than those scenarios having an S_b=1 (first column in Table 3.3). This excess of fixed variants affects only nonsynonymous positions, while synonymous positions have the same number of fixations in all scenarios as expected (~22K, Supplementary Table A.1B)

(Birky and Walsh 1988). Nucleotide diversity within the Wild and Domestic populations and the divergence (from the simulated outgroup) at nonsynonymous versus synonymous positions are shown in Supplementary Table A.2. The lower divergence ratio (D_n/D_s) versus polymorphism ratio (P_n/P_s) suggests an excess of deleterious polymorphisms at nonsynonymous sites, as expected given the simulated DFE parameters. The number of exclusive variants in the Domestic populations and shared variants between Wild and Domestic populations across all scenarios is shown in Table 3.3, as well as the mutations that were initially shared but that eventually got fixed in one of the populations. In Domestic populations, we observe that the absolute number of exclusive polymorphic (SxD) and fixed (SfD) variants are more affected by the demographic bottleneck than by the strength of positive selection. As expected, the number of shared mutations (polymorphic in both populations, Ssh) is very high in scenarios with migration (scenarios 1-3) and very low in models with a long bottleneck and no migration (scenarios 7-10). The number of fixed variants in Wild which are still polymorphic in Domestic (SfWxD) is very low in scenarios with long bottlenecks without migration, while the number of fixed variants in Domestic which are still segregating in Wild (SfDxW) is the largest in scenarios with long bottlenecks with no migration.

The Fraction of Beneficial Substitutions: Comparing α across Scenarios and Types of Sites

Table 3.3: Number of fixed, exclusive and shared variants observed in the Domestic populations for each scenario.

Scenarios	SfWD	SfW	SfD	SxW	SxD	Ssh	SfWxD	SfDxW
1	55446	0	0	14440	16539	51081	281	211
2	47645	0	0	25526	35707	46442	813	344
3	47193	0	0	28436	36185	44599	765	499
4	47377	5	24	50933	42248	20691	900	2385
5	47785	47	132	51511	40213	16641	783	3274
6	54503	62	139	48257	38018	13427	884	2966
7	54909	312	4267	56671	4351	4	9	9260
8	47221	318	4552	61805	4228	3	9	10928
9	47582	270	4362	61077	4362	4	17	11058
10	55340	449	4301	56019	4293	1	16	9351

SfWD: Fixed variant in the species in relation to the outgroup. SfW: Exclusive fixed variant in Wild. SfD: Exclusive fixed variant in Domestic. SxW: Exclusive polymorphism in Wild, SxD: Exclusive polymorphisms in Domestic. Ssh: Shared polymorphic variants. SfWxD: Fixed variants in Wild and polymorphic in Domestic. SfDxW: Fixed in Domestic and polymorphic in Wild.

Next, we ought to know how the strength of positive selection and the number of loci under positive selection affect the fraction of beneficial nonsynonymous substitutions (α) across scenarios (Table 3.4, see also Supplementary Table A.3 for the absolute number of fixations). At Wild, true α values correlate with the simulated conditions in Table 1; higher α values can be found in scenarios where strong selection is assumed (scenarios 1, 6, 7 and 10). In contrast, in Domestic (to observe the effects of domestication, α is calculated counting variants that are fixed in the Domestic and are polymorphic or absent in Wild), the highest α values are found in scenarios 1, 3,

5 and 9. These scenarios (except scenario 1) have a large fraction of deleterious sites in the Wild population that become beneficial in the Domestic population (m_7 sites, Table 3.2). We also find that a substantial amount of the beneficial substitutions (25-60%) in those scenarios come specifically from m_7 sites.

In relation to understanding the contribution of exclusive variants (mostly new mutations that occurred after the domestication split) to current α estimates, we observe that exclusive beneficial substitutions are contributing very modestly to the total number of beneficial substitutions for short bottleneck scenarios (Supplementary Table A.3). This suggests that adaptive amino acid substitutions after short bottlenecks come mostly from standing variation that were present before the domestication split. In contrast, for long bottleneck scenarios, the number of exclusive fixed beneficial mutations is higher at the Domestic population than the Wild population and around 5 times larger than short bottleneck scenarios without migration (Supplementary Table A.3). This is because in small populations the time to fixation decreases for all mutations, neutral, beneficial and deleterious. For long bottleneck scenarios, exclusive beneficial substitutions, or new mutations, are playing an important role and they contribute substantially to the realized α at the end of the simulation.

Table 3.4: True α for Total, Shared and Exclusive mutations in relation to all nonsynonymous fixations.

	All variants				Shared Variants			Exclusive variants		
	Wild	Domestic			Wild	Domestic		Wild	Domestic	
	α_{total} *	α_{total} †	α_{m2+m3} §	α_{m7} ¶	α_{total} *	α_{total} †	α_{m7} ¶	α_{total} *	α_{total} †	α_{m7} ¶
1	0.211	0.156	0.156	0.000	0.211	0.156	0.000	.	.	.
2	0.056	0.063	0.045	0.017	0.056	0.063	0.017	.	.	.
3	0.061	0.106	0.081	0.025	0.061	0.106	0.025	.	.	.
4	0.059	0.060	0.060	0.000	0.059	0.061	0.000	.	.	.
5	0.061	0.105	0.041	0.064	0.061	0.105	0.063	0.097	0.101	0.089
6	0.213	0.089	0.078	0.011	0.213	0.089	0.011	0.175	0.085	0.000
7	0.208	0.055	0.055	0.000	0.208	0.070	0.000	0.219	0.026	0.000
8	0.060	0.040	0.039	0.001	0.060	0.047	0.001	0.112	0.026	0.001
9	0.058	0.094	0.034	0.060	0.058	0.101	0.062	0.072	0.080	0.058
10	0.212	0.055	0.050	0.005	0.212	0.072	0.005	0.254	0.023	0.004

* fix (m_2, m_3, m_4)/fix(Nsyn); † fix (m_2, m_3, m_7)/fix(Nsyn); § fix(m_2,m_3)/fix(Nsyn); ¶ fix(m_7)/fix(Nsyn), where numerator and denominator belong to Total, shared or exclusive.. s: scenario.. § The variants that are fixed at Domestic and at Wild are not counted in the Domestic α calculation, as they are considered occurred before the split of the populations.

True α vs. estimated α

In order to investigate if current α estimators can be used to extract meaningful conclusions from natural wild and domesticated populations, we estimate α for Wild and Domestic populations using different methods (Figure 3.3 and Supplementary Tables A.4-A.6 for confidence intervals).

As expected, the standard MK test (Smith and Eyre-Walker 2002) underestimates α due to the segregation of deleterious mutations at low frequency. The estimates are improved when the asymptotic MK test is used (Haller and Messer 2017), which corrects for the excess of deleterious mutations at low frequencies. Nonetheless, the asymptotic MK test is not very accurate in scenarios with long bottlenecks. The asymptotic MK test detects positive selection in the Wild population only for scenarios with strongly beneficial mutations (scenarios 1, 6, 7 and 10). In the Domestic populations, significantly positive α values are estimated only for scenarios with short bottlenecks and strong positive selection (scenarios 1 and 6). Finally, we also estimate α using the polyDFE framework (Tataru et al. 2017). polyDFE α estimates tends to be less noisy and slightly more accurate than the asymptotic MK test. However, qualitatively speaking it suffers the same weaknesses than the asymptotic MK test under long bottlenecks. Our list of tested α estimators have very limited power to detect the subtle differences in α between these simulated population pairs.

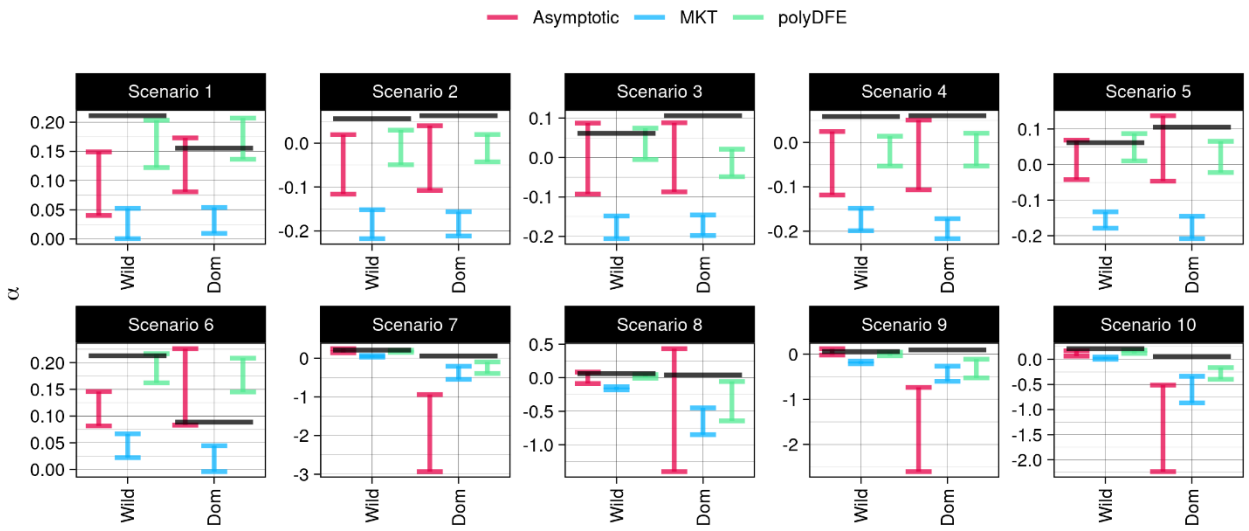

Figure 3.3: Comparison of true and inferred α values across different methods. Asymptotic: MKTa (Messer and Petrov 2013), MKT: α calculation using Smith and Eyre-Walker (2002). polyDFE: α inference using the algorithm from Tataru et al. (2017). True α is depicted with a solid black line. Confidence intervals at 95% are obtained by bootstrap.

Comparison of the Allele Frequency Distributions across Populations and Scenarios

Figure 3.4 shows the ratio (in $\log_2$ scale) between the observed derived allele frequency and the expected derived allele frequency, or unfolded site frequency spectrum (uSFS), under the same demographic model but without selection (obtained by coalescence simulation with *ms*, Hudson 2001). In the absence of (direct or indirect) selection, the observed derived allele frequency and the expected derived allele frequency should be the same and the $\log_2$ ratio equal to zero. The small and uniform $\log_2$ ratio for synonymous mutations in Wild populations suggests that there is good agreement between the simulated data and our naive neutral expectation in the absence of selection. Thus, we do not find evidence of linked selection affecting the synonymous SFS in Wild populations. In contrast, the nonsynonymous derived allele frequency distribution in Wild populations shows an

excess of low-frequency variants due to the direct effect of negative selection on those mutations and exhibits a considerable lack of intermediate- and high-frequency variants. However, in scenarios with strong positive selection (scenarios 1, 6, 7 and 10), beneficial nonsynonymous variants leave a signal in the SFS. Beneficial variants generate an increase of high-frequency variants that resembles (or matches) the neutral expectation at nonsynonymous positions (see Supplementary Figures A.4-A.13 under these models).

The Domestic populations have suffered a shift in the selection coefficients at a sizeable number of mutations (both shared and exclusive) together with a demographic bottleneck, with or without migration from the Wild population. For synonymous variants, we show that the observed allele frequency distribution seems to be disturbed by indirect selection on nearby nonsynonymous mutations in most scenarios (Figure 3.4). In scenarios with long bottlenecks (scenarios 7-10), only the lowest and highest frequency bin are in excess, and all other frequency bins show a remarkable lack of synonymous variants. We find that the allele frequency distribution for nonsynonymous mutations in Domestic populations under long bottlenecks is very similar to those found for synonymous variants but very different to the expected in the absence of selection under a purely demographic model (Figure 3.4 and Supplementary Figure A.1). Thus, under long bottlenecks allele frequency distributions seem deeply affected by (direct and indirect) selection. Similarly, synonymous variants also seem to be under indirect selection at scenarios with short bottlenecks having a large percentage (25%) of nonsynonymous mutations shifting their selection coefficients (scenarios 3, 5 and 6). Synonymous variants show an excess of singletons, a lack of low- to intermediate-frequency variants, and, again, a remarkable excess of mutations at high frequency. This result suggests that synonymous (and other non-beneficial nonsynonymous) mutations might be hitch-hiking with beneficial nonsynonymous mutations after the domestication split (Supplementary Figures A.4-A.13) (Hartfield and Otto 2011).

Interestingly, scenarios 3, 5, and 6 also show an increase of nonsynonymous variants at high frequency but not as strong as the one shown for synonymous variants. We show that for scenarios 3, 5 and 6, the allele frequency distribution for nonsynonymous mutations occurring at sites m_2, m_3 and m_7 (which are beneficial at the Domestic population) resemble that described for synonymous variants. However, this pattern is much weaker for deleterious nonsynonymous mutations occurring at deleterious sites (m_5 and m_6 sites, Supplementary Figures A.4-A.13). Finally, we analyze the uSFS for shared and exclusive variants. Here our aim is to determine the contribution of shared and exclusive variants to the observed uSFS and the process of domestication, particularly for selectively neutral synonymous variants that later will be important to correct (or infer) the demographic changes suffered by nonsynonymous variants. Again, we compare the uSFS of synonymous shared and exclusive variants in relation to the expected neutral shared and exclusive uSFS under the same demographic model in the absence of selection (Supplementary Figures A.14-A.33 for all scenarios and Supplementary Figures A.2-A.3 for the ratio of the demographic neutral pattern versus the Wright-Fisher Standard Neutral Model). In general, for short bottleneck scenarios (scenarios 1-6) we observe a modest impact of domestication on shared synonymous variants (Supplementary Figures A.14-A.23), except in scenario 3.

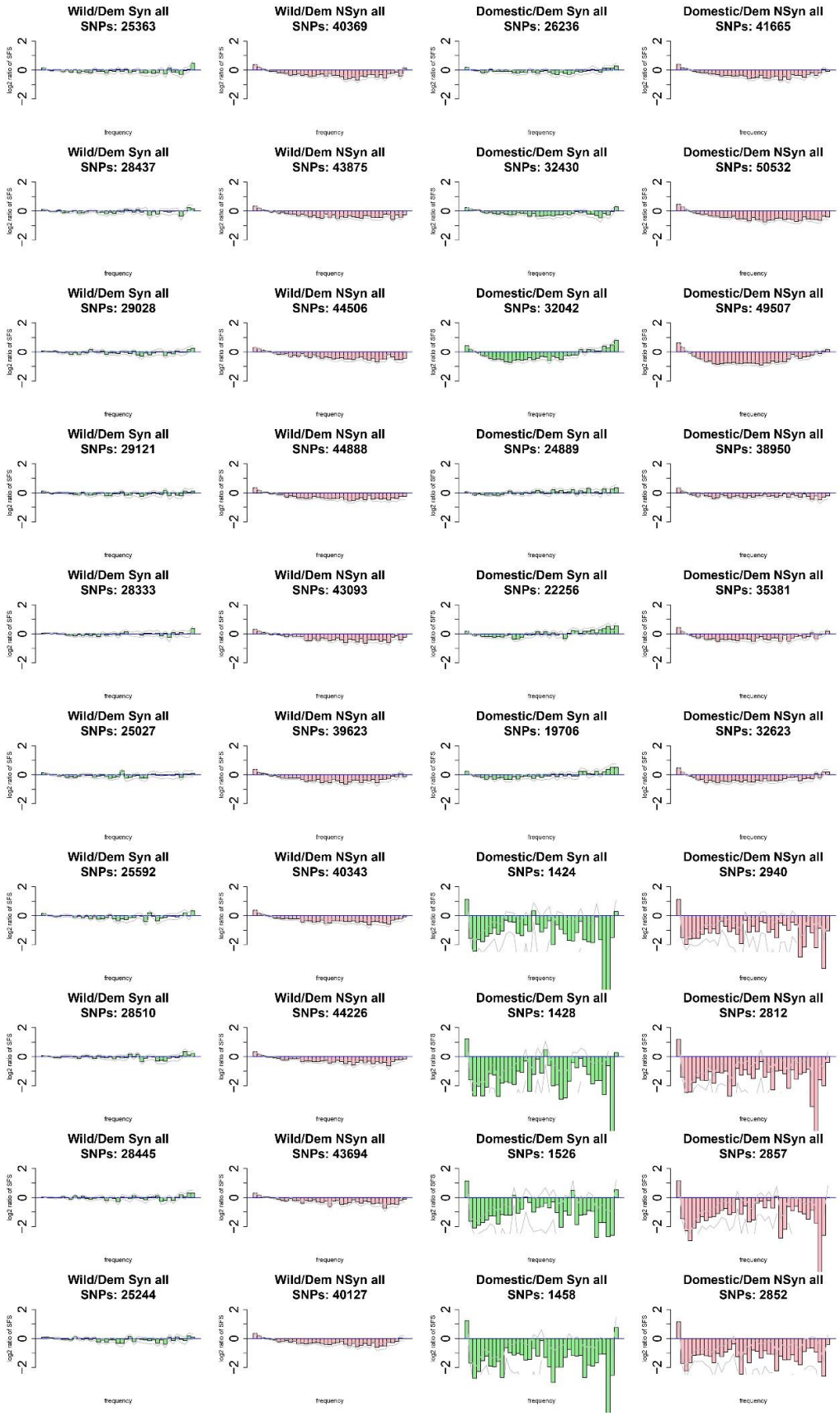

Figure 3.4: Log_2 ratio of the simulated SFS (with selection) versus the expected SFS (in the absence of selection) but under the same demographic model.

There is an absence of intermediate frequency shared synonymous variants and an excess of low and high frequency shared synonymous variants in scenarios 3 relative to the expected in the absence of (linked) selection. Indeed, in scenario 3, the uSFS of shared synonymous variants is markedly disturbed, and surprisingly, all types of shared mutations, including deleterious mutations (m_5, m_6, m_4 in Domestic), exhibit this excess of high frequency variants that is not explained by demography alone (Supplementary Figure A.16). The high proportion of sites that change their selective effect together with migration from Wild to Domestic populations may explain this excess of shared polymorphisms at high frequency in scenario 3. Other scenarios which do not share this parameter combination do not show this pattern. In scenarios with long bottlenecks (scenarios 7-10) the SFS of shared synonymous variants in Domestic populations is massively disturbed. There is a large excess of shared singletons and a large absence of shared variants at the rest of frequency classes. In relation to exclusive mutations (Supplementary Figures A.24-A.33), in scenarios with migration, short bottlenecks and a change in the DFE (scenario 2 and 3), there is an excess of exclusive synonymous variants at all frequencies (except singletons) suggesting that the contribution of exclusive variants to neutral diversity levels can be boosted in the presence of migration from the Wild population. For scenarios with short bottlenecks without migration (scenarios 4-6) there is a lack of exclusive synonymous variants in most frequency bins except for very low and high frequency bins where there is an excess. In scenarios with long bottlenecks (scenarios 7-10) the SFS of exclusive synonymous variants in Domestic populations is again massively disturbed. There is a large excess of almost fix mutations and a large absence of exclusive variants at the rest of frequency classes.

In conclusion, our hypothetical domestication processes (in the absence of free recombination) tend to affect more exclusive synonymous variants than shared synonymous variants. In the absence of migration, domestication tends to deplete the SFS of exclusive neutral mutations (and under long bottlenecks this depletion also affects shared synonymous variants). In the presence of migration, exclusive neutral variants become more abundant than expected in the absence of selection (and if the change in the DFE is big, also shared synonymous variants at high frequency become more abundant than expected).

Estimation of the DFE: detecting differences in DFE between Wild and Domestic populations

In the previous section we showed the pervasive impact of linked selection on synonymous polymorphisms. In the Domestic populations, the neutral SFS cannot be explained by the underlying true demography alone. The two methods we employ to estimate the DFE of nonsynonymous mutations either are agnostic to the underlying demography and just aim to correct for any distorter that affects equally the synonymous and nonsynonymous SFSs (polyDFE, Tataru and Bataillon 2019), or aim to first estimate the underlying demography using the synonymous SFS and then use that inferred demography to estimate the DFE parameters (*dadi*, Gutenkunst et al. 2009). Hence, the question is to what extent the nuisance *r* parameters from polyDFE or the inferred demographic model from *dadi* will be enough to recover the true DFE parameters in our simulations. In this work we are not particularly interested in recovering the true demography (given the pervasiveness of linked selection), but in recovering the true changes in the DFE.

polyDFE: 1D-SFS

We investigate whether polyDFE is able to capture differences in the DFE of Domestic and Wild populations across 10 possible domestication scenarios (Table 3.1). We run five nested models (Table 3.5) and compare them by means of likelihood ratio tests (LRTs) (Table 3.6).

Table 3.5: List of nested polyDFE models and (co)estimated parameters. Independently estimated parameters for the domesticated and wild population (Var). Jointly estimated parameters for the Domestic and Wild populations (Fix).

Model name	**Deleterious DFE**		**Beneficial DFE**		**Population mutation rate**	**Nuisance parameters**	**Polarization parameter**
	β	S_d	p_b	S_b	ϑ	r_i	ε
f1	Var	Var	None	None	Var	Var	Var
f10	Fix	Var	None	None	Var	Var	Var
f2	Var	Var	Var	Var	Var	Var	Var
f20	Fix	Var	Var	Var	Var	Var	Var
f30	Fix	Var	Fix	Fix	Var	Var	Var

Under an agnostic approach (not assuming any *a priori* information about our datasets), LRTs between a number of nested models allow us to address several relevant questions related to the DFE. First, we compare whether the estimated shape of the deleterious DFE is the same in both populations and whether the inclusion of beneficial mutations affects the estimation of the shape parameter. When comparing models that do not consider beneficial mutations (models f1 versus f10, first row at Table 3.6), the model with a different shape for Domestic and Wild populations is accepted for scenarios 1, 3, 6 and 10. This means that an artificial change in the shape of the deleterious DFE between Domestic and Wild populations is invoked. Scenarios 1, 6 and 10 include strong positive selection (S_b = 10), and a change in fitness effects occurs at 0%, 25%, and 5% of sites, in each scenario, respectively. Scenario 3 involves weak positive selection (S_b = 1), but many sites change their fitness effects (25%). In contrast, when comparing models that consider beneficial mutations (models f2 vs f20, third row at Table 3.6), a shared shape of the deleterious DFE is preferred in all scenarios (this is expected given the simulation parameters). This result suggests that not accounting for beneficial mutations can generate an artefactual change in the shape of the deleterious DFE between populations (as first noticed by Tataru et al. 2017). Thus, we find that when assuming no beneficial mutations the wrong shape is estimated (f1 vs f10) in the domesticated population when positive selection is strong (scenarios 1, 6, and 10) or when a large fraction of mutations become weakly beneficial (scenario 3).

Second, we find that including the positive DFE is only statistically significant for scenarios with strong positive selection (scenarios 1, 6 and 10) (Table 3.6 second and fourth row). Note, however, that all of our simulations include beneficial mutations. There is one scenario with strong positive selection that remains undetected, scenario 7. The distinct characteristic of this scenario is that although the domesticated population has undergone a long bottleneck, as in scenario 10, the fitness effects of mutations do not change between populations. Hence, our ability to detect beneficial mutations with polyDFE relies on the strength of positive selection but also on the

proportion of sites that change their fitness effects and the duration of the domestication bottleneck. It is also worth mentioning that the power to detect beneficial mutations in the unfolded SFS increases when the shape of the deleterious DFE is jointly estimated (see the decrease in the p-value for scenarios 1, 6 and 10 for the comparisons of models f1 vs. f2 and models f10 vs. f20).

Table 3.6: Likelihood ratio test p-value for each scenario. Significant and marginally significant comparisons (< 0.1) are highlighted in bold.

Models	**Domestication Scenarios**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
f1 vs. f10 (constant vs. variable shape in deleterious DFE models)	**0.007**	0.533	**0.018**	0.167	0.239	**0.000**	0.592	0.308	0.252	**0.008**
f1 vs. f2 (variable shape while comparing deleterious DFE vs. full DFE)	**0.095**	0.990	0.996	0.649	0.969	**0.007**	0.816	0.902	0.846	0.147
f2 vs. f20 (constant vs. variable shape under full DFE model)	0.442	0.539	0.380	0.423	0.351	0.317	0.321	0.550	0.398	0.137
f10 vs. f20 (constant shape while comparing deleterious DFE vs. full DFE)	**0.002**	0.987	0.246	0.334	0.900	**0.000**	0.844	0.730	0.775	**0.030**
f30 vs. f20 (constant shape while comparing constant positive DFE vs. variable positive DFE)	0.597	0.919	0.064	0.342	0.826	**0.011**	0.823	0.846	0.466	0.658

Finally, we wonder whether the differences in the beneficial DFE between Domestic and Wild populations are detectable or not. Interestingly, when comparing models f20 and f30 (last row at Table 3.6), our ability to detect differences in the beneficial DFE between Domestic and Wild populations is very limited in most scenarios. With the exception of scenario 6, where there is both a large proportion of mutations that show a change in fitness (25%) and strong positive selection (S_b = 10). The short domestication bottleneck in scenario 6 might also increase the power to detect this change in the positive DFE.

DFE parameter estimates with polyDFE

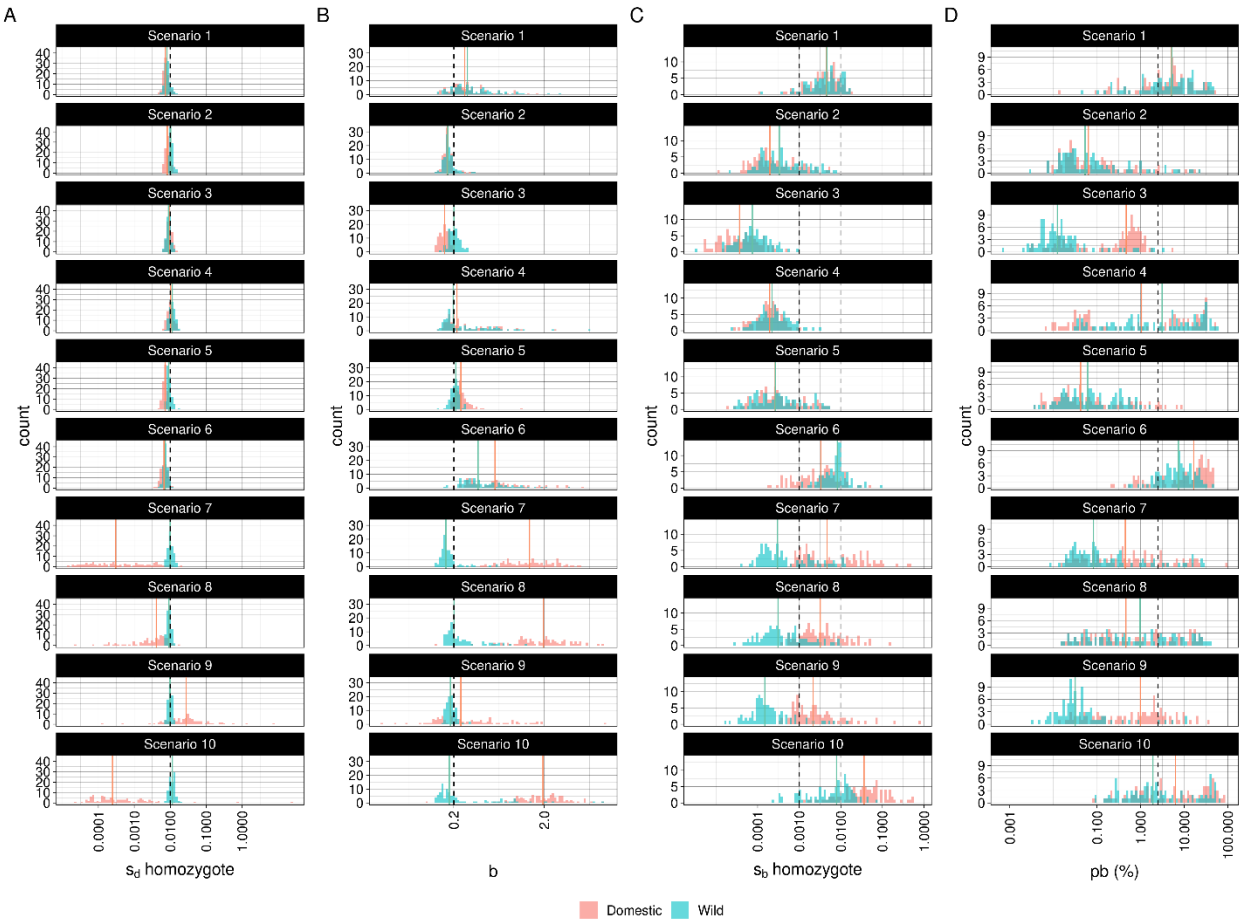

Figure 3.5: Sampling distributions of estimated DFE parameters from polyDFE obtained by 100 bootstrap replicates. Dashed vertical lines show the true simulated values of the parameters and the coloured vertical lines show the median inferred value across replicates. For s_b we show two dashed lines; one in grey when $S_b = 10$, and the other in black when $S_b = 1$. We show the true p_b found in Wild populations, which is the lower bound of what can be found in Domestic populations. To obtain s_d and s_b from S_d and S_b values we use π at synonymous sites and the true simulated mutation rate (1e-6) to get the realized N_e after the bottleneck and/or the action of linked selection. Note that in polyDFE *s* is defined to be the selection coefficient on the heterozygote (like in *dadi*), but the scaled selection coefficient is defined as $4N_es$.

Since the true model generating the data in natural wild and domesticated populations is unknown, we first extract the AIC of each model and then compute the AIC-weighted parameters for all models (Supplementary Table A.7) (Tataru and Bataillon 2019; Castellano et al. 2019). Figure 3.5 shows the distribution of parameters estimated by polyDFE using bootstrap analysis. The mean deleterious coefficient (s_d) of the DFE is accurately estimated for both populations and all scenarios, except when there are long bottlenecks in Domestic populations, where the mean s_d is generally underestimated (except for scenario 9 which is slightly overestimated). Similarly, the shape of the negative DFE is generally well estimated except in scenario 6 and long-bottleneck scenarios, where the shape is overestimated. The estimation of the parameters of the positive DFE is more challenging, and for all scenarios, the estimates include high variance and, thus, a wide range of possible values.

Moreover, in our particular case the true probability that a new mutation is beneficial (p_b) in Domestic populations is not easily defined because there are beneficial mutations in Domestic populations that were already segregating as effectively neutral or deleterious mutations just before the start of the domestication process. This violates the implicit model assumption of constant selection coefficients along generations. Plus, in some scenarios there is migration from the Wild to the Domestic population that will be re-introducing beneficial alleles. Supplementary Figure A.0 shows the true composition of beneficial sites and segregating sites (m_2, m_3 and m_7) after the domestication split which we believe is the best proxy of the current p_b in Domestic populations. However, for Wild populations p_b is well defined (due to the constant population size and DFE, and the lack of migration) but the estimates, as expected, tend to be noisy with very wide confidence intervals.

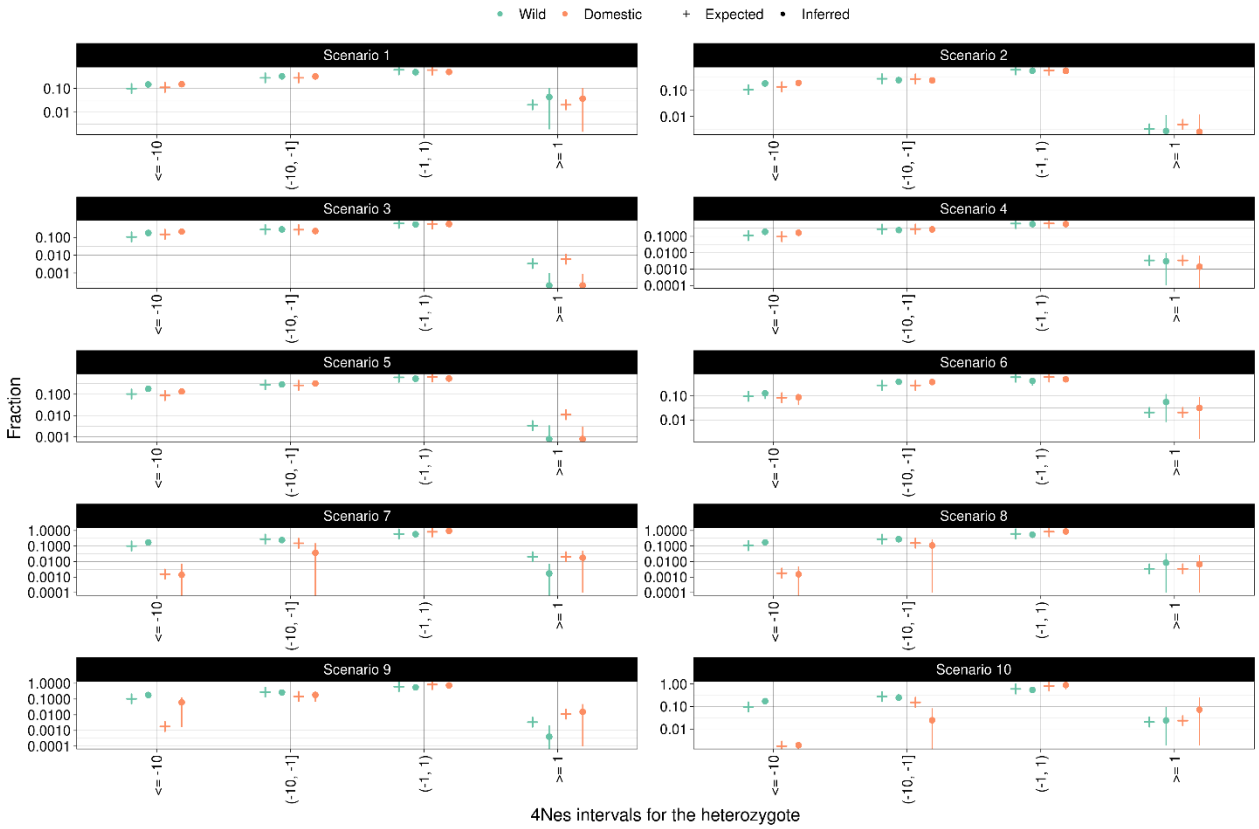

Figure 3.6: Discretized full DFE inferred by polyDFE, showing the Wild and Domestic populations under each scenario.

Figure 3.6 shows the inferred distribution of the AIC weighted DFE. Discrete DFE ranges tend to be less noisy than individual DFE parameters (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). The discrete ranges for the deleterious DFE show an excellent agreement between the inferred and the simulated values especially for the fraction of effectively neutral mutations and for scenarios with short bottlenecks (scenarios 1-6), while for scenarios with long bottlenecks (scenarios 7-10) the inferred fraction of weakly and strongly deleterious mutations becomes noisy and sometimes inaccurate. The confidence intervals for the number of effectively beneficial mutations seem to capture the real values indicating a lack of power and not a bias. However, there

are slight underestimations for scenarios 3 for both populations, scenario 5 for the Domestic population and scenario 7 and 9 for Wild populations.

dadi: 2D-SFS

Demography inference

A joint DFE between two diverged populations can be inferred with the *dadi* software using the 2D-SFS (Huang et al. 2021). This provides us with a potential advantage relative to the 1D-SFS approach because the joint DFE can quantify the fraction of mutations with divergent selective effects even when the underlying parameters describing the full DFE remain the same between populations. In contrast to polyDFE, *dadi* estimates first the demographic parameters to then estimate the DFE. The parameters of a demographic scenario describing two recent divergent populations are estimated by fitting a demographic model to the 2D-SFS of synonymous sites (Huang et al. 2021). Those parameters are estimated by maximizing a Poisson composite likelihood and contrasted with the corresponding expected SFS. Supplementary Tables A.8-A.9 show the estimated demographic parameters and their confidence intervals respectively. Here we describe the inferred parameters one by one:

- The population size of the wild (N_{ew}) before the domestication split is slightly but significantly underestimated in all scenarios, except for scenarios 1, 4 and 8 where it is well-recovered.
- The time of the domestication split (t_1) is in general well-recovered but in scenario 2 is slightly underestimated and in scenario 3 is slightly overestimated.
- The population size of the Domestic populations under the bottleneck (N_{e1d}) is overestimated only for scenario 2 and significantly underestimated for scenarios with long bottlenecks (scenarios 7-10). For the rest of scenarios N_{e1d} is well-recovered.
- The duration of the bottleneck (t_{bot}) is well-recovered in all scenarios except in scenarios with migration. In scenario 1 (no change in the DFE) the duration of the bottleneck is significantly underestimated and very close to zero, this is expected given the high migration rate. In the scenarios with migration and a change in the DFE (scenarios 2 and 3) the duration of the bottleneck is substantially overestimated and in fact they are similar to the long bottleneck scenarios.
- The estimates of the population size after the bottleneck in Domestic populations (N_{e2d}) are noisier and they are significantly overestimated in scenarios 2, 3 and 10, and significantly underestimated in scenarios 5 and 6.
- The time of the population size recovery (t_3), or the time since the end of the bottleneck, is well-recovered except in scenarios with migration and a change in the DFE. For scenarios 2 and 3 t_3 is significantly underestimated. The recovery time seems again affected by the migration parameter and the change in the DFE.
- Finally, for scenarios without migration from the Wild to the Domestic population the migration parameter (m) is consistently low and in all cases not significantly different from 0. In scenarios with migration (scenarios 1-3), there seems to be an interaction between the change in the DFE and the migration parameter. Scenario 1 which is a negative control (there is no change in the DFE between Domestic and Wild populations) the migration rate is significantly underestimated and half of its true value. However, in scenarios 2 and 3 which

show a change of fitness effects in 5% and 25% of the sites, respectively, there is a large underestimation of the migration rate (13X lower than expected).

The previous comparison between the expected number of exclusive mutations in the absence of selection and the observed number of exclusive mutations in the presence of (indirect) selection showed a substantial excess of exclusive variants in scenarios 2 and 3. This emergent excess of exclusive variants might be "interpreted" by *dadi* as a low migration rate and a long but very mild bottleneck. Note that N_e estimates from levels of neutral diversity in these two scenarios are larger than N_e estimates from their paired Wild populations (Supplementary Table A.7). These results suggest that the linkage effects produced in the domestication process can bias some aspects of the inferred demographic model, especially in the presence of migration and when many sites change their fitness effect in the domesticated population.

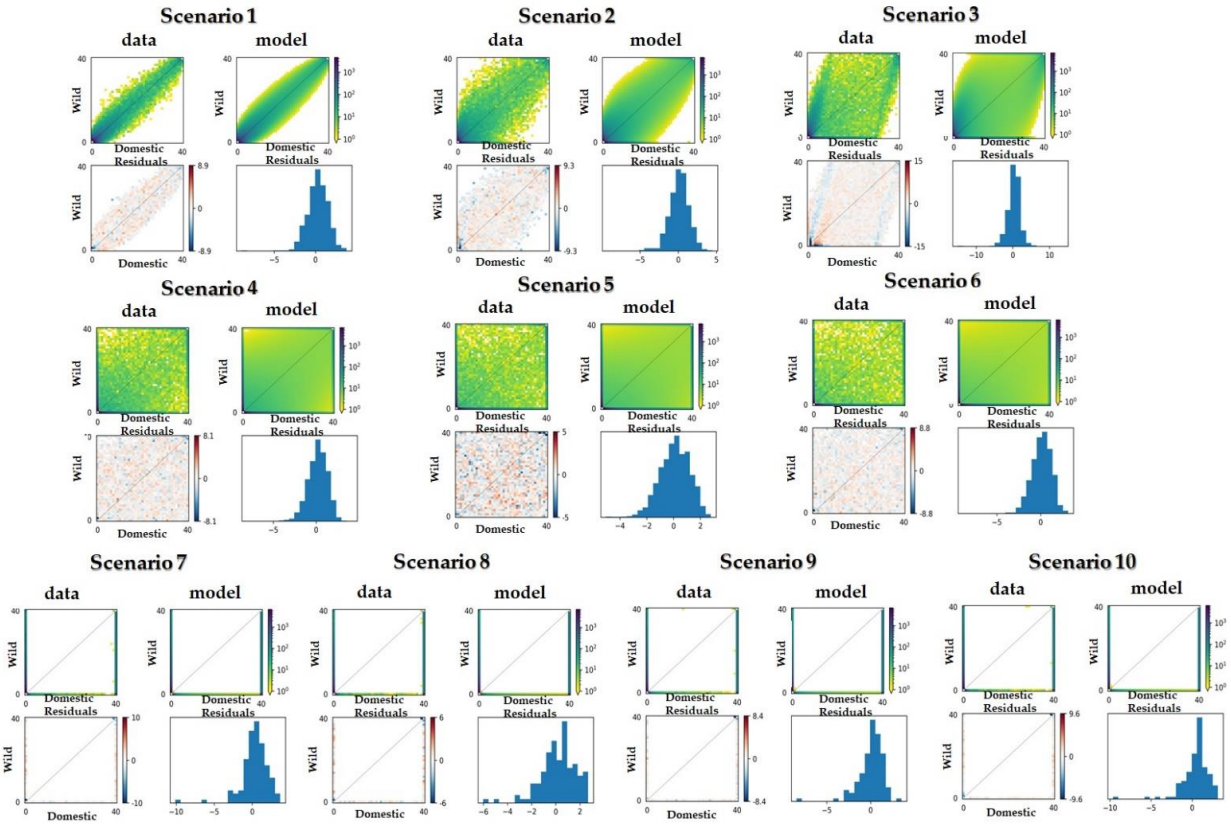

Figure 3.7: Estimation of demographic parameters from synonymous mutations under each scenario. Each plot includes four subfigures or panels. The one entitled "data" (top left panels) represents the two-dimensional simulated SFS for both populations, while the right one labeled "model" (top right panels) depicts the expected 2D-SFS for the inferred dadi model. The two plots at the bottom illustrate the residual performance (bottom left panels) and the residual histogram (bottom right panels) of the model.

Figure 3.7 reports the validation plots of the fitted model using the *dadi* framework. The narrow ranges of the distribution in the residual histogram plot (Figure 3.7, bottom right panels) reveals a good fit of the model under each distinct scenario (even in the absence of shared polymorphisms as in long bottleneck scenarios 7-10). Note that although scenarios 2 and 3 have the same demographic model than scenario 1, the 2D-SFS fills almost all the cell combinations (Figure

3.7, top left panels), while in scenario 1 (no change in the DFE) the 2D-SFS is concentrated in the diagonal. This spread of the synonymous 2D-SFS in scenarios 2 and 3 must be driven by the change in the DFE and selection at linked sites. This spread of the 2D-SFS is also observed in scenarios with short bottlenecks but without migration (scenarios 4-6). Scenarios with long bottlenecks show no shared polymorphisms (scenarios 7-10) and all observed 2D-SFS cell combinations are in the margins. Particularly revealing for scenarios 2 and 3 are the model residuals of each 2D-SFS cell combination (Figure 3.7, bottom left panels). Here model residuals are computed as the expected counts derived from the model minus the observed counts in the (simulated) data. Cell combinations with more observed variants than expected are in blue, while cell combinations with less observed variants than expected are in red. In scenarios 2 and 3 there are less observed counts in the diagonal of the 2D-SFS than those expected by the model. In contrast, particularly evident for scenario 3, there are two blue coloured vertical "stripes".

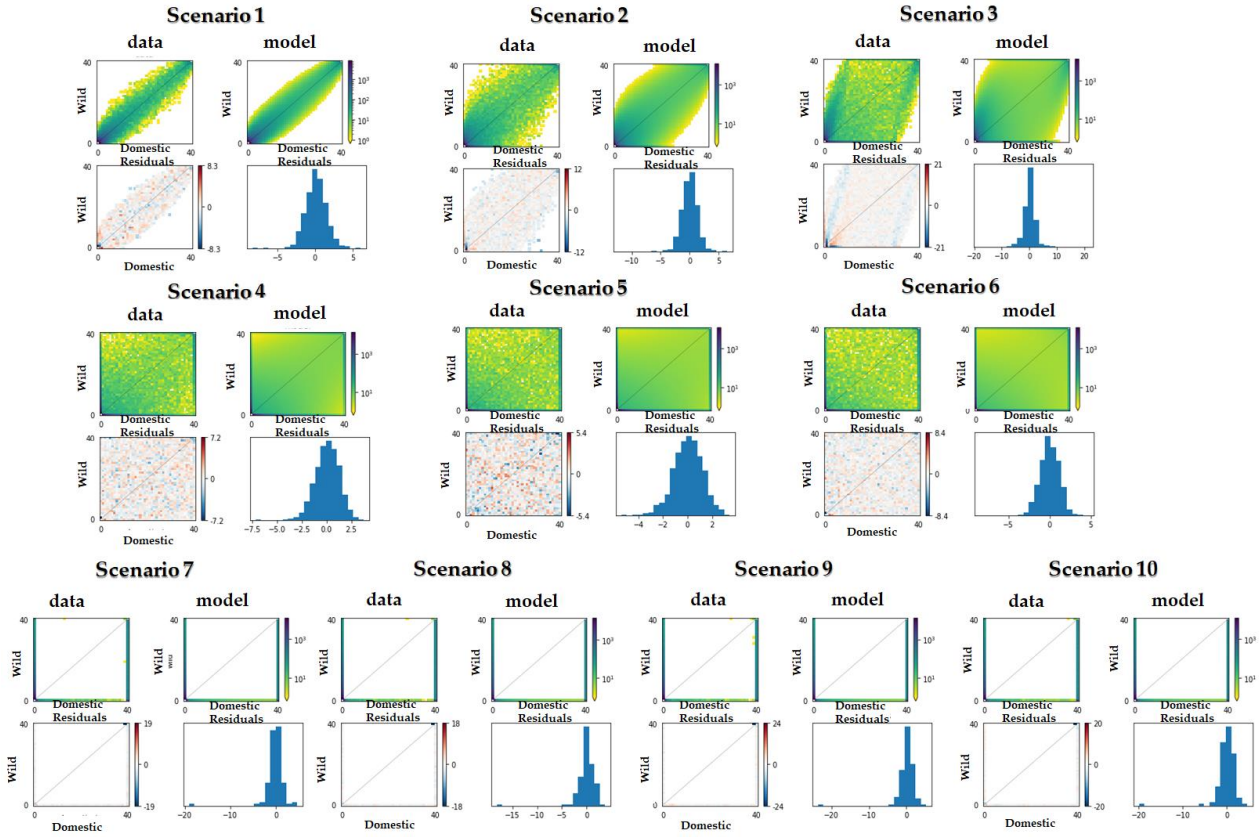

Figure 3.8: Estimation of the DFE from nonsynonymous mutations under each scenario. Each plot includes four subfigures or panels. The one entitled "data" (top left panels) represents the two-dimensional simulated SFS for both populations, while the right one labeled "model" (top right panels) depicts the expected 2D-SFS. The two plots at the bottom illustrate the residual performance (bottom left panels) and the residual histogram (bottom right panels) of the model.

The left stripe indicates an excess of observed mutations at low frequency in the Domestic population that can be at any frequency in the Wild population (this could correspond to synonymous variants in linkage with nonsynonymous mutations that were effectively neutral in Wild but that became more deleterious in Domestic). The second stripe indicate an excess of observed mutations at high frequency in the Domestic population that can be at any frequency in the Wild population

(this could correspond to synonymous variants in linkage with nonsynonymous mutations that were effectively neutral in Wild but that became more beneficial in Domestic). In the rest of scenarios, the 2D-SFS model residuals show a random distribution of colours across cell combinations which is indicative of the absence of systematic biases.

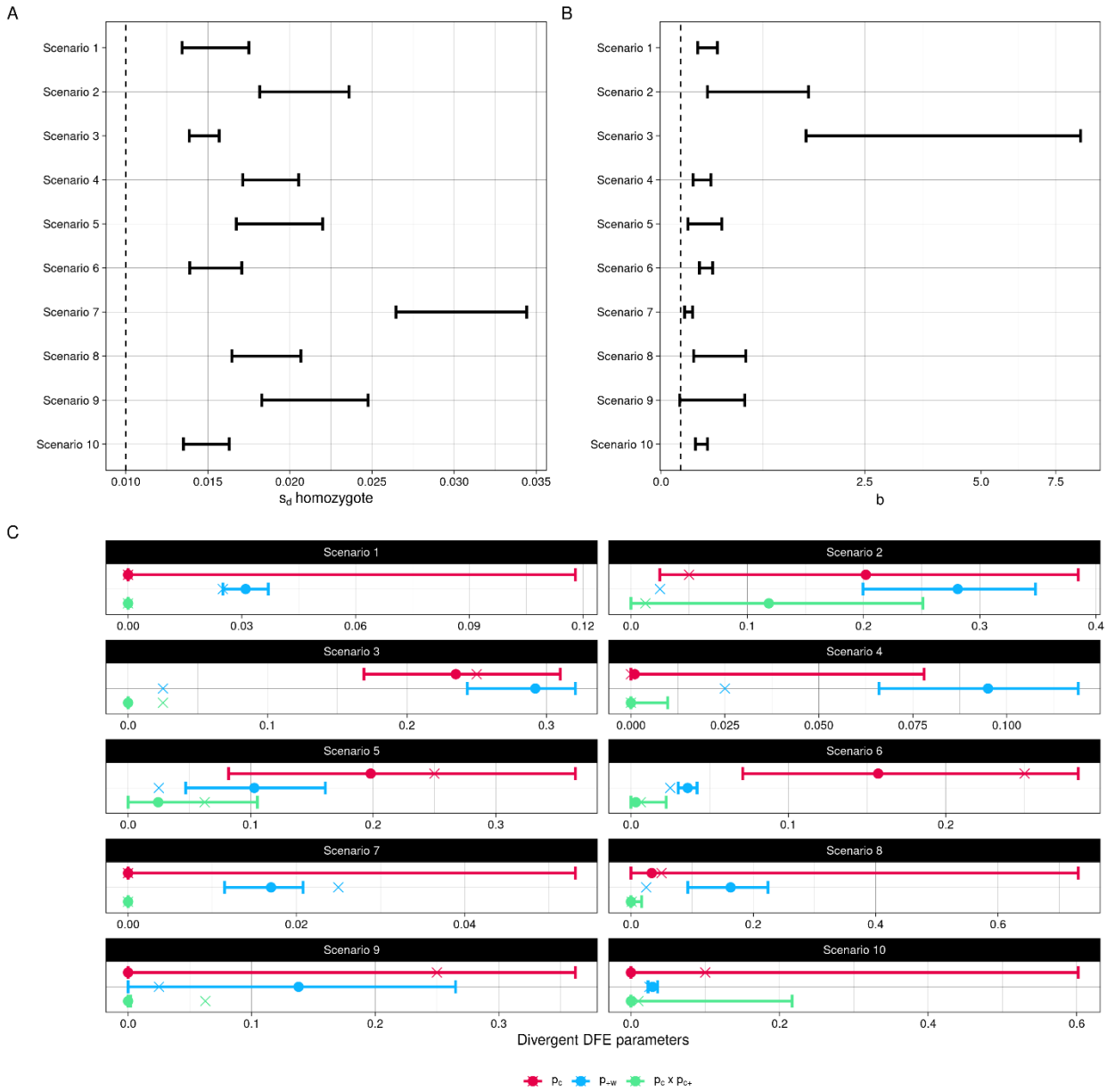

Figure 3.9: Confidence intervals and values for the inferred parameters versus the real ones (vertical dashed lines and crosses) for each population and scenario for s_d (A), the shape parameter b (B) and the DFE parameters describing divergent selective effects (C). To obtain s_d from S_d values we use π at synonymous sites in the inferred ancestral population size and the true simulated mutation rate (1e-6) to get the realized N_e after the bottleneck and/or the action of linked selection. Note that in *dadi s* is defined to be the selection coefficient on the heterozygote (like in polyDFE), but the scaled selection coefficient is defined as $2N_es$.

DFE parameter estimates with ∂a∂i

The estimated demographic parameters above are used for the inference of the DFE of nonsynonymous mutations using the 2D-SFS. Figure 3.8 shows a comparison between the 2D-SFS inferred by the most likely model vs. the observed 2D-SFS and the residuals of the model for each 2D-SFS cell combination. The residual histograms for nonsynonymous mutations (Figure 3.8, bottom right panels) show a nicely distributed residuals with a mean around zero for all scenarios with the exception of few very negative outliers in scenarios with long bottlenecks. We observe again the blue coloured vertical "stripes" pattern found before for synonymous mutations at scenario 3. The left stripe could be generated by nonsynonymous mutations that have become more deleterious in the Domestic population (mutations at m_4 sites plus some mutations at m_6 sites). While the right stripe can be interpreted as nonsynonymous mutations that have become more beneficial in the Domestic populations (mutations at m_7 sites plus some mutations at m_3 sites). Both of these stripes are likely driven by direct selection on selectively divergent mutations falling at m_3, m_4, m_6 and m_7 sites, and also by other nonsynonymous mutations that might have a constant selection coefficient but turn out to be in linkage with those selectively divergent mutations.

Table 3.7: Comparing models with and without positive selection from *dadi* likelihoods.

Scenario	True S_b	AIC_0	AIC_1	AIC_{10}	Lower AIC Model
1	10	5514.9	5360.1	5325.4	$S_b = 10$
2	1	8277.9	8221.5	8208.1	$S_b = 10$
3	1	13578	13456	13517	$S_b = 1$
4	1	8866.7	8810.3	8833.6	$S_b = 1$
5	1	8540.8	8520.1	8475.3	$S_b = 10$
6	10	8513.0	8368.4	8264.4	$S_b = 10$
7	10	1125.8	1076.4	1073.7	$S_b = 10$
8	1	1081.9	1053.1	1065.5	$S_b = 1$
9	1	1192.5	1183.1	1182.9	$S_b = 10$
10	10	1236.4	1125.2	1121.5	$S_b = 10$

AIC_0: Akaike Information Criteria ($2k-2\ln L_{S0}$) for a model 0, considering $S_b = 0$. k=5 after fixing demographic parameters.

AIC_1: Akaike Information Criteria ($2k-2\ln L_{S1}$) for a model 1, considering $S_b = 1$. k=5 after fixing demographic parameters.

AIC_{10}: Akaike Information Criteria ($2k-2\ln L_{S10}$) for a model 10, considering $S_b = 10$. k=5 after fixing demographic parameters.

Figure 3.9 and Supplementary Tables A.10-A.11 show the inferred DFE parameters under the new joint DFE model. The inferred strength of negative selection (s_d) is overestimated between 1.5X to 3X times. For the shape parameter (b) there is also a general overestimation of this parameter, especially in scenario 3. Regarding the inference of the strength of positive selection note that with *dadi* the positive DFE is precomputed with a unique point mass distribution of positive selection, or no beneficial mutations at all. Under the polyDFE framework beneficial mutations are drawn from an exponential distribution mimicking our forward simulations. However, with *dadi* we assume either

no beneficial mutations or a given proportion of beneficial mutations with S_b = 1 or S_b = 10 when homozygote. Model comparison shows that *dadi* can distinguish between weak and strong positive selection in most cases (Table 3.7). However, for scenarios 2, 5 and 9 which show a high proportion of positive change on Domestic populations (25%), the likelihood is somewhat better considering a strongest selection coefficient (S_b = 10).

As we shown before with polyDFE, the estimation of the positive DFE and its changes between populations is very challenging mainly because beneficial mutations tend to be rare, and the true p_b in Domestic populations is not well defined due to the action of positive selection on standing variation and the re-introduction of beneficial mutations due to migration. Two violations that we believe might be commonplace in nature. This is the reason why we propose a new joint DFE model which might be able to tackle some of this complexity (Figure 3.1 and Materials and Methods for a full description). Under this new joint DFE model, we can measure the fraction of mutations that change their selection coefficient (p_c). This fraction p_c is accurately estimated in scenarios with shared polymorphisms (scenarios 2, 3, 5 & 6), and it is not statistically different from zero for scenarios 1, 4, and all the long bottleneck scenarios (scenarios from 7 to 10). Scenarios 1, 4 and 7 are negative controls without a change in the DFE between Wild and Domestic populations. However, scenarios 8, 9 and 10 do suffer a change in the DFE but remain undetected very likely due to the lack of shared polymorphisms. Even in long bottleneck scenarios where p_c seems to be poorly estimated, confidence intervals seem to capture the real values indicating a lack of power and not a bias. The estimated proportion of positive mutations in Wild populations (p_{+w}) is very close to the true value in scenarios where strong positive selection is driven these mutations (scenarios 1, 6, 7 and 10). In scenarios with weak positive selection, p_{+w} tends to be overestimated, particularly in scenarios with migration and a change in the DFE (scenarios 2 and 3). Note that in Figure 3.9 is illustrated the proportion of nonsynonymous mutations that change their selection coefficient toward positive values in the Domestic populations as the product of $p_c * p_{c+}$. Unfortunately, this product, which aims to estimate the fraction of m_7 sites in our simulations, is always not significantly different from zero. Overall, the new joint DFE model seems able to capture the fraction of mutations that change their selection coefficient in the presence of shared polymorphisms, but it is not able to discern how many deleterious mutations have become beneficial. We believe p_c could be used in future studies to quantify the number of mutations with divergent fitness effects between natural wild and domesticated populations.

Discussion

The literature about the genomic basis of domestication (e.g., Ross-Ibarra et al. 2007, Flood and Hancock 2017; Moyers et al. 2018; Flori et al. 2019; Frantz et al. 2020; Leno-Colorado et al. 2020) revolves around three main questions: (1) Is domestication driven by few loci of large effect or by many loci of weak effect (polygenic adaptation)? (2) Is domestication dominated by standing variation (variation that was present before the start of domestication) or by new exclusive mutations in the domesticated populations? (3) Does domestication have a genetic cost due to the bottleneck and inbreeding (reviewed by Moyers et al. 2018)? In this work, we ask an intimately related question: Can we investigate domestication as a change in the full distribution of fitness effects comparing the SFS of wild and domesticated populations? This question is motivated by empirical results. For example,

a small number of fixed functional variants have been detected in domestic pig populations in relation to wild populations (Leno-Colorado et al. 2020 but see Groenen et al. 2012; Rubin et al. 2010). Although frequent introgression has been described between pig populations (Frantz et al. 2015; Ramírez et al. 2015), there are clear phenotypic features that distinguish wild from domestic populations. The absence of major genetic signals to explain most traits under domestication makes us think that polygenic adaptation might be a plausible hypothesis. We have simulated potential domestication scenarios involving both a population bottleneck and a change in the selection coefficients in new and standing variation in the domesticated population. Thus, in our simulations, we consider that a hypothetical environmental change (*i.e.,* domestication), makes a number of deleterious mutations (and effectively neutral mutations) in the wild populations to become beneficial in the domesticated populations. Also, beneficial mutations in the wild populations can become deleterious or effectively neutral in the domesticated populations. However, since the rate of new beneficial mutations in the wild populations is relatively small (2.5% of all new mutations), then beneficial variants that become deleterious are a minority of the minority and we expect a second-order effect of those mutations in our simulations. We also investigate the impact of migration from the wild to the domesticated population on our ability to measure a change in the full distribution of fitness effects.

First, we start describing the fraction of adaptive substitutions (α) across scenarios and populations. We wonder which is the contribution of new beneficial mutations and mutations that were already segregating at the beginning of domestication to α. We are also interested in quantifying the contribution of deleterious polymorphisms in the wild populations that become beneficial in the domesticated populations. Most contemporary adaptive fixations came from variation that predates the initiation of domestication. In other words, we find that in all scenarios, the largest proportion of adaptive substitutions came from polymorphisms that were already beneficial before the domestication split. However, in some scenarios, deleterious polymorphisms that became beneficial in domesticated populations can also explain a large fraction of all adaptive substitutions (Table 3.4, Supplementary Table A.3). Interestingly, we find that the contribution of new beneficial variants (exclusive of the domesticated populations) on α depends on the duration of the bottleneck. Most beneficial mutations that have occurred after the domestication split, and only in the domesticated population, have not reached fixation in scenarios with short bottlenecks, and hence their exclusion from the computation of α affects very little the realized α values. In contrast, new exclusive beneficial mutations in scenarios with long bottlenecks contribute substantially to the current fraction of adaptive substitutions in domesticated populations. We hypothesize this is due to the reduction in the time to fixation expected in small populations. It is also important to note that α in domesticated populations is not much larger than in wild populations even when around 5%-10% of the mutations are beneficial in the domesticated population (as in scenarios 3, 5 and 9, see Supplementary Figure A.0) compared to the 2.5% in the wild populations. In fact, in many scenarios, α in domesticated populations can be smaller than in wild populations. These modest α values in domesticated populations may be due to the bottleneck associated with the domestication process, but also to the short time since the initiation of domestication. The small effective population size, on one hand, reduces the efficacy of positive and negative selection (and α) in domesticated populations but on the other hand, accelerates the fixation of mutations (neutral, deleterious, and

beneficial). Unfortunately, we find that our list of α estimators (MKT, MKTa, polyDFE), when applied to our simulations, have little power to detect statistically different α values between populations. We do not recommend extracting strong conclusions from applying α estimators alone to natural wild and domesticated populations.

Second, we study the impact of direct and indirect selection and demography on the 1D-SFS. We compare the 1D-SFS of synonymous (neutral) and nonsynonymous (selected) variants in our simulated scenarios against the expected 1D-SFS for (only) neutral mutations under the same demographic model. We show that in domesticated populations selection at linked sites needs to be invoked to explain the 1D-SFS of synonymous mutations. We hypothesize that the mode of linked selection affecting the 1D-SFS depends on the duration of the bottleneck. For those scenarios with long bottlenecks, we think that background selection is the main distorter of the synonymous 1D-SFS. Under long bottlenecks, haplotypes carrying deleterious variants (and neutral variants in linkage) can raise in frequency (Hartfield and Otto 2011). After the bottleneck these haplotypes will decrease in frequency dragging with them neutral variation. These anti-sweeps (or sweeps toward extinction) might produce the observed synonymous 1D-SFS in our simulations which show an excess at low frequency variants and a lack of variation in the rest of frequency classes (Figure 3.4). The 1D-SFS of synonymous and nonsynonymous mutations is very similar in our simulations with long bottlenecks but very different from that expected in the absence of selection (Supplementary Figures A.1-A.3). This makes us think that a particularly strong form of background selection might be taking place. Surprisingly, the synonymous 1D-SFS also seems affected by indirect selection, or selection at linked sites, in short bottleneck scenarios when a substantial fraction of standing deleterious (and effectively neutral) variation become beneficial (Figure 3.4, scenarios 2, 3, 5, 6). Short bottlenecks are expected to have a modest impact on the population frequencies of standing variation, particularly in the presence of migration from the wild population to the domesticated population. In these scenarios, we think that the shift in the selection coefficients in the domesticated populations is setting the ideal conditions for the emergence of soft sweeps (Hermisson and Pennings 2005; Stetter et al. 2018). This multitude of soft sweeps could explain the excess of synonymous variants at a high frequency that we observe in our simulations (Figure 3.4). Note that in scenarios 3, 5 or 6 the excess of (neutral) synonymous variants at high frequency is even larger than the excess found for (selected) nonsynonymous variants. We hypothesize that soft sweeps might be behind this distortion of the synonymous 1D-SFS in our short bottleneck simulations. This detailed analysis of the 1D-SFS is a prerequisite to interpret the comparison of the full DFE using polyDFE and *dadi*, because both methods contrast the (1D or 2D) SFS of synonymous and nonsynonymous mutations to infer the DFE parameters.

Third, the comparison of the full DFE for domesticated and wild *in silico* populations finds that polyDFE seems to provide more accurate estimates of the deleterious DFE than the new joint DFE model build on *dadi*. The discretized DFE analysis with polyDFE shows that the confidence intervals for the number of effectively beneficial mutations seem to capture the real values in most scenarios indicating a lack of power and not a bias. However, due to the lack of power with polyDFE is not possible to detect statistically significant differences between populations for the beneficial DFE, except when selection is strong, and many sites change their selection coefficients (as in scenario 6). The new joint DFE model seems to accurately estimate the strength of positive selection across most

scenarios, but when a very large fraction of mutations become beneficial it tends to invoke stronger positive selection. We find that, even in the demographically and selectively "stable" wild populations, the inference of the probability that a new mutation is beneficial (called p_b in polyDFE and p_{+w} in *dadi*) can be extremely challenging with both methods. More importantly, for domesticated populations, the meaning of this probability is unclear because the selection coefficients of new and standing variation are not static, migration from the wild to the domesticated population can re-introduce beneficial mutations and our domesticated populations are far from demographic equilibrium. We find that joint DFE models offer one key advantage in the study of domestication. This is the computation of the fraction of mutations that *changed* their selection coefficient in the domesticated population relative to the wild population (p_c) and from those how many became beneficial during domestication (p_{c+}). We find that in the presence of shared polymorphisms p_c is well estimated, but p_{c+} turned out to be very noisy. Thus, our new joint DFE model contrasting the 2D-SFSs is able to quantify a signal of domestication, understood as the fraction of mutations with divergent selective effects, when two populations have diverged recently enough to share many polymorphisms. Given the strong distortion of the synonymous SFS that we described before it is not surprising that polyDFE only detects a significant difference in the positive DFE in one scenario or that *dadi*'s deleterious DFE and p_{c+} estimates are inaccurate and noisy. polyDFE does not estimate the underlying demography, but it corrects for the potential impact of demographic changes on the estimation of the DFE. The way polyDFE corrects for demographic changes, and/or linked selection, is through the nuisance parameters r_i which contrast the relative observed synonymous SFS to the relative expected neutral SFS under the Wright-Fisher model and then uses these inferred r_i parameters to correct the nonsynonymous SFS and estimate the DFE (Eyre-Walker et al. 2006; Tataru et al. 2017). It is likely that the strong impact of selection at linked sites in our simulations is making the r_i parameters overcorrect the nonsynonymous SFS. This might be hampering our ability to detect differences in the positive DFE between populations. In contrast, *dadi* infers first the demographic parameters with synonymous mutations. The inference of the demographic model with *dadi* shows that under some domestication scenarios, even in the presence of realistic recombination rates, important aspects of the inferred demographic model, such as the migration rate or the duration of the bottleneck, can be substantially biased.

Finally, we want to discuss the limitations and future directions of this work. This work is not an exhaustive benchmarking across all one hundred and twenty-eight possible domestication scenarios that we could have been simulated given the value of the different parameters that we play with at Table 3.1. We do not investigate the prevalence of hard and soft sweeps in our simulated scenarios, but this could certainly enrich the study of domestication under polygenic adaptation. Our work also posits the question of how we can infer (or at least correct) the demographic changes in the presence of pervasive selection at linked sites and changes in the selection coefficients of standing variation. Machine learning approaches might be a natural choice for such complex scenarios involving rich non-equilibrium dynamics. But perhaps methods that first estimate demography using patterns of linkage disequilibrium, which seem more robust to linked selection than SFS statistics (Ragsdale and Gutenkunst 2017; Novo et al. 2022), in combination with SFS methods can provide more general, accurate enough, and faster estimates than machine learning tools. We have not even incorporated the complexity introduced by the distribution of dominance (but see Arunkumar et al.

2015) or the impact of population size changes in the wild populations. Nonetheless, the limits to comprehend the domestication process by comparing the DFE between wild and domesticated populations has become more evident.

Chapter 4

Transposable element polymorphisms improve prediction of complex agronomic traits in rice

DOI: https://doi.org/10.1007/s00122-022-04180-2
Theoretical and Applied Genetics volume 135, pages 3211–3222 (2022)

Abstract

Transposon Insertion Polymorphisms (TIPs) are significant sources of genetic variation. Previous work has shown that TIPs can improve detection of causative loci on agronomic traits in rice. Here, we quantify the fraction of variance explained by Single Nucleotide Polymorphisms (SNPs) compared to TIPs, and we explore whether TIPs can improve prediction of traits when compared to using only SNPs. We used eleven traits of agronomic relevance from five different rice population groups (Aus, Indica, Aromatic, Japonica and Admixed), 738 accessions in total. We assess prediction by applying data split validation in two scenarios. In the within population scenario, we predicted performance of improved Indica varieties using the rest of Indica accessions. In the across population scenario, we predicted all Aromatic and Admixed accessions using the rest of populations. In each scenario, BayesC and a Bayesian reproducible kernel Hilbert space regression were compared. We find that TIPs can explain an important fraction of total genetic variance and that they also improve genomic prediction. In the across population prediction scenario, TIPs outperformed SNPs in nine out of the eleven traits analyzed. In some traits like leaf senescence or grain width, using TIPs increased predictive correlation by 30 – 50%. Our results evidence, for the first time, that TIPs genotyping can improve prediction on complex agronomic traits in rice, especially when accessions to be predicted are less related to training accessions.

Introduction

More than half of the world population consumes rice (*Oryza sativa*) in their daily diet. To secure nutritional requirements of a growing human population, the improvement of grain yield, both in quantity and in nutritional quality, is imperative. This is a significant challenge in the face of climate change and limited cultivable land. Current pace of rice genetic improvement may be too slow to meet these demands (Rosegrant and Cline 2003; Zhao et al. 2018). Genomic selection can be a useful tool to accelerate genetic progress (Meuwissen et al. 2001). Numerous studies in rice and in other plant species (Jighly et al. 2019; Tessema et al. 2020; Xu et al. 2020; Krishnappa et al. 2021) have already shown that genomic prediction (GP) can increase breeding speed. GP is particularly effective when traits are controlled by numerous loci which are difficult to map individually, such as yield and other traits of agronomic interest. For a recent review in rice, see Xu et al. (2021).

Conceptually, genomic prediction (GP) is a 'large p, small n' scenario where the number of variables p (molecular markers) is typically far larger than the number of observations n. In this setting, either variables must be selected or restrictions on the solutions must be imposed, or a combination of both. Methods such as LASSO (Tibshirani 2011) or BayesC (Meuwissen et al. 2001) are examples of the first choice, whereas ridge regression or GBLUP (VanRaden 2008) involve restrictions on the square of solutions (L2 norm). Numerous metrics exist for measuring predictive ability. Among others, it can be measured as the correlation between predicted and observed phenotypes by splitting the data in training and test sets. Prediction accuracy is affected by different factors such as the size of the training data, heritability, similarity between training and testing populations, or choice of marker sets (Robertsen et al. 2019; Xu et al. 2021; Goddard and Hayes 2007).

In general, there is no consensus on which GP method is best. A recent review by Reinoso-Peláez et al. (2022) point at Reproducible Kernel Hilbert Space (RKHS) as the best overall method in

plants. But there is variability. For instance, Tehseen et al. (2021) compared GBLUP, Ridge Regression (RR), LASSO, Elastic Net (EN), Bayesian Ridge Regression (BRR), Bayesian alphabet (A, B, C, …), RKHS for different traits, observing that no single method outperformed the rest for all traits. Kaler et al. (2022) conducted a comparative study among 11 different methods for two traits in soybean, rice, and maize, reporting better predictive abilities using Bayes B. Xu et al. (2018) found that GBLUP and LASSO performed best in hybrid breeding. Other authors have suggested integrating genomic prediction with crop growth models to evaluate the efficiency of phenotypic strategies and the impact of the different yield components on the prediction accuracy (Bustos-Korts et al. 2019; Cooper et al. 2016). Selecting SNPs based on genome wide association studies (GWAS) has also been proposed, e.g., Spindel et al. (2016).

Irrespective of the algorithm chosen, single nucleotide polymorphisms (SNPs) are the main class of markers used so far in GP due to their genome wide abundance and genotyping automatization. SNPs are not, however, the only source of phenotypic variability in the genome. In the last few years, data has accumulated on the importance of presence-absence variation and structural variation as a source of phenotypic variability in plants, including in rice (e.g., Fuentes et al. 2019). Transposon Insertion Polymorphisms (TIPs) can account for a major fraction of intraspecific structural variation, as it has been recently found in maize (Haberer et al. 2020). In fact, transposable elements are considered as one of the main drivers for plant genome variability, impacting on genome coding capacity and regulation in numerous ways (Lisch 2013). However, until the recent development and evaluation of reliable methods for calling TIPs from short-read resequencing data (Vendrell-Mir et al. 2019), it was not possible to use TIPs for GWAS approaches.

Importantly, recent studies in rice and in tomato have shown that the use of TIPs as genetic information can result in an increase of association signals as compared to SNPs in GWAS (Akakpo et al. 2020; Carpentier et al. 2019; Domínguez et al. 2020; Castanera et al. 2021). These results prompt us to investigate whether transposons can also improve prediction accuracy. For this purpose, we used the TIP genotypes from Castanera et al. (2021) and the phenotype database hosted in IRRI (Jackson 1997; Mansueto et al. 2017). Note that a better model fit, as observed in GWAS, does not necessarily imply a more accurate prediction and thus the question posed here is pertinent. Further, any improvement in prediction albeit small can translate into large genetic gains when accumulated through generations.

Materials and Methods

Rice accessions and traits

The 738 accessions used in this study (Supplementary Table B.1) are from the collection conserved at IRRI used for the 3,000-rice genome project (Jackson 1997; Li et al. 2014) and were chosen because they were sequenced at least at 15x depth. The 738 accessions retained pertain to all main rice population groups: Aus/Boro (AUS, N=75), Indica (IND, N=451), Japonica (JAP, N=166), Aromatic (ARO, N=17). The accessions that cannot be assigned to a specific rice group are categorized as Admixed (ADM, N=29). We used the SNP-based group assignment from Sun et al. (2017) to identify the different subsets of this study.

Out of the 56 traits originally available at IRRI SNP-Seek database (https://snp-seek.irri.org/), we chose the 11 traits for which data was available in the 738 accessions selected. Some discrete traits were binned to balance the number of observations per class and time to flowering was log-transformed. Supplementary Table B.2 shows basic statistics and transformations applied. Principal Component Analysis (PCA) for the 11 phenotypes was obtained with the "prcomp" function available in R.4.1.0 (Team 2021) environment. For plotting loading variables of PCA, package "factorextra" (Kassambara and Mundt 2020) and packages "ggrepel" (Slowikowski 2020) and "ggbiplot" (Vu 2011) for the biplot were used.

Markers

A binary ped file format with the Core SNP dataset for all chromosomes was downloaded from the SNP-Seek database. The original dataset consisted of 404,388 bi-allelic SNPs from 3,034 rice accessions, including the 738 accessions selected. Markers with minor allele frequency ≤ 0.01 and missing rate > 1% were filtered out using plink2 (Purcell et al. 2007; Chang et al. 2015). Missing genotypes were imputed using Beagle 5.2 with default parameters (Browning et al. 2018). The final dataset consisted of 228,871 SNPs, which were used for the analyses reported here. Of those, 50,485 SNPs were in gene regions.

Transposable Elements (TEs) are divided into two main classes "copy and paste" (Class I TEs) or "cut and paste" Class II TEs. In rice, the most abundant Class I elements are RLX (LTR retrotransposons) and RIX (Non-LTR retrotransposons), whereas DTX (DNA TEs with terminal inverted repeats) and MITEs (Miniature Inverted-repeat Transposable Elements) are the most prevalent (Mao et al. 2000). Here we used markers from both classes, accounting for 94 % of the TIPs described in Castanera et al. (2021). Class I TIPs were represented by 21,571 RLX and RIX markers. Class II consisted of 52,120 MITE and DTX markers. In contrast to SNPs, TIPs can only be genotyped as presence / absence, recoded consequently as 0/1, and defined as genomic windows with an average size of 1.2 kb. TIP windows were taken from Castanera et al. (2021), and are based on the intersection of the individual TE insertion regions predicted for each accession with genome-wide windows of a fixed size (1kb, merging adjacent windows). These TIPs were further classified as genic or intergenic by intersecting the windows with MSU7 non-TE gene annotation (Kawahara et al. 2013). A TIP was considered genic if the window overlapped at least 1bp with the gene feature. There were 17,034 genic MITE/DTX and 5,024 genic RLX/RIX TIPS. The remaining TIPs were considered intergenic.

MITEs amplify by bursts from individual elements creating highly homogeneous families, as previously reported in Arabidopsis (Santiago et al. 2002) and rice (Lu et al. 2017). Different bursts of amplification at different evolutionary times may have different prediction potential for particular phenotypes. In an attempt to study the potential predictive capacity of individual families, we created individual TIP genotype matrices for each of the 18 largest MITE families described in Castanera et al. (2021) (Supplementary Table B.3). Each of these matrices included only TIPs originated from a single transposon, in this case MITE, family.

Genetic variance inference

We fitted the following linear model in order to estimate genetic variance components explained by each marker set:

$$\mathbf{y = \mu + Z\,u_1 + Z\,u_2 + Z\,u_3 + e} \qquad \mathbf{(4.1a)}$$

where μ is the general mean, y is the phenotype vector of size n, the number of accessions, Z is an identity incidence matrix, u_1, u_2, u_3 are random effects representing each of the marker groups and e is the residual. We assume $u_1 \sim N(0, K_S\,\sigma_S^2)$, $u_2 \sim N(0, K_M\,\sigma_M^2)$, and $u_3 \sim N(0, K_R\,\sigma_R^2)$, where K_S, K_M, K_R are genomic relationship matrices obtained from SNPs, MITE/DTX and RLX/RIX markers, respectively. These matrices were calculated using AGHMatrix (Amadeu et al. 2016). Model 1a was fitted with a Bayesian Reproducible Kernel Hilbert Space (RKHS, Herbrich et al. 1999) as implemented in BGLR package (Pérez and de los Campos 2014) using default priors to estimate σ_S^2, σ_M^2 and σ_R^2.

Genomic prediction

Plant breeding is primarily based on trials of new crosses, which can be lengthy and costly. The speed of development for new improved varieties depends largely on accuracy of prediction for new genotypes. We evaluated two distinct validation scenarios that cover two important issues: prediction of performance within population (rice group in this case) and prediction of individuals from different groups. In the first scenario, we measured accuracy when predicting performance of improved Indica varieties (N = 48) using the rest of accessions, including non-improved Indica accessions. Accessions from IRRI core collection are classified as “improved”, “breeding and inbred lines” and “traditional” varieties. We used this passport information to identify this subset of improved varieties. ‘Improved’ Indica varieties correspond to most modern and commercial lines available at IRRI collection. In this scenario, performance to be predicted is from highly related accessions to those in the training set. In the second scenario, we predicted performance of all admixed (ADM, N = 29) and aromatic (ARO, N = 17) accessions using the rest of groups. In this case, performance to be predicted is from accessions that may not be too related to accessions in the training set, and we expect prediction to be worse than in the former scenario. For instance, the ADM group is a small, highly heterogeneous collection of accessions.

The rationale for the first scenario is that new selected accessions can be crosses within the same population, and the breeder can be interested in designing new better performing crosses out of traditional varieties. The second scenario is more challenging, since we do not use any sample of the population to be predicted. These two scenarios, within and across populations, resemble main challenges faced in a breeding program. Note there are infinite designs for assessing predictive accuracy. For instance, we did not study prediction in Japonica because we preferred to focus on a larger number of traits, since genetic architecture is a main factor influencing predictive performance (Daetwyler et al. 2010).

Ample literature shows that no single method performs best for all traits and scenarios. Here, we compared two alternative modelling strategies: Bayesian RKHS as described above, and BayesC. RKHS is equivalent to ridge regression and GBLUP, whereas BayesC is a variable selection method. The two methods were applied to both predictive scenarios. For RKHS, we compared predictive performance using all markers (model 4.1a above) with sub models

$$\mathbf{y = \mu + Z\,u_1 + e,} \qquad \textbf{(4.1b)}$$

$$\mathbf{y = \mu + Z\,u_2 + e,} \qquad \textbf{(4.1c)}$$

and

$$\mathbf{y = \mu + Z\,u_3 + e,} \qquad \textbf{(4.1d)}$$

i.e., when using only SNPs (model 4.1b), only MITE/DTX (model 4.1c) or only RLX/RIX (model 4.1d) markers. For BayesC, the complete model was:

$$\mathbf{y = \mu + X_S\,\beta_1 + X_M\,\beta_2 + X_R\,\beta_3 + e,} \qquad \textbf{(4.2a)}$$

where X_S, X_M and X_R are the standardized genotypic values of each marker class, β_1, β_2 and β_3 are the corresponding vectors of effects for SNPs, MITE/DTX and RLX/RIX markers, respectively. As with RKHS, three partial models were also evaluated:

$$\mathbf{y = \mu + X_S\,\beta_1 + e,} \qquad \textbf{(4.2b)}$$

$$\mathbf{y = \mu + X_M\,\beta_2 + e,} \qquad \textbf{(4.2c)}$$

and

$$\mathbf{y = \mu + X_R\,\beta_3 + e} \qquad \textbf{(4.2d)}$$

In BayesC, a probability π of presence / absence of a SNP in the model is sampled from $\pi \sim \text{Beta}(p_0, \pi_0)$. Following Pérez and de Los Campos (2014, see their Tables 1 and S1), 'the beta prior is parameterized in a way that the expected value by $E(\pi) = \pi_0$; on the other hand, p_0 can be interpreted as the number of prior counts (prior "successes" plus prior "failures")'. Here we chose p_0 = 5 and π_0 = 0.01.

In a subset of cases, we evaluated whether using only genic SNPs improved prediction compared to using all available markers. Similarly, we conjectured that not all transposable elements are equally likely to cause phenotypic changes. We analyzed predictive performance of models containing TIPs from each of the largest 18 MITE families present in the rice genome (Supplementary Table B.3). To avoid repetitive, lengthy results we make the additional analysis using two agronomic traits of high importance on rice breeding, time to flowering and grain length. An earlier or later growing can determine seed production. Grain size related traits such as grain length/width are important breeding targets since they affect the quality of the crop yield. These two traits may also represent alternative genetic architecture (Begum et al. 2015; Xu et al. 2015; and Chen et al. 2021).

Using either RKHS or BayesC, phenotypes to be predicted were removed from the dataset, the model fitted using the remaining phenotypes and the correlation between predicted and observed phenotypes computed as a measure of predictive accuracy. From a practical point of view, it is important to assess whether predictions using TIPs, or all markers are better than the state-of-the-art method, i.e., with SNPs only. To assess variability of results, we generated 10,000 bootstrap sampling with replacement from the corresponding pairs of phenotypes observed and predicted with

each method and marker set. We then computed the correlation observed-predicted samples within each bootstrap sample, and we counted how many times correlation using SNPs only was lower than with each alternative strategy. Phenotypic measurements and variables were centered and scaled to mean 0 and variance 1. BGLR was run for 100,000 iterations using default priors for RKHS. This number of iterations seemed enough to attain convergence (Supplementary Figure B.1).

Results

Descriptive analysis

Figure 4.1A shows the loadings, i.e., the projections of variables into the lower dimensional space, of each trait to the principal components. In the figure, the length of the arrow is proportional to trait contribution and the angle between arrows, to their correlation. An analysis in two principal components shows that the first component depends on grain width and grain weight whereas culm diameter, time to flowering and leaf length are the main contributors to the second component. The rest of traits contribute more modestly to total phenotypic variation. A sample projection (Figure 4.1B) shows graphically how accessions differed in the traits studied. Supplementary Figure B.2 shows the differences in trait distributions across accessions. In general, populations differed for most traits although to varying extent. Figure 4.1 does indicate, e.g., that Japonica accessions tend to have higher grain weights and widths, as they are projected in the lower part of the figure, and as shown in Supplementary Figure B.2.

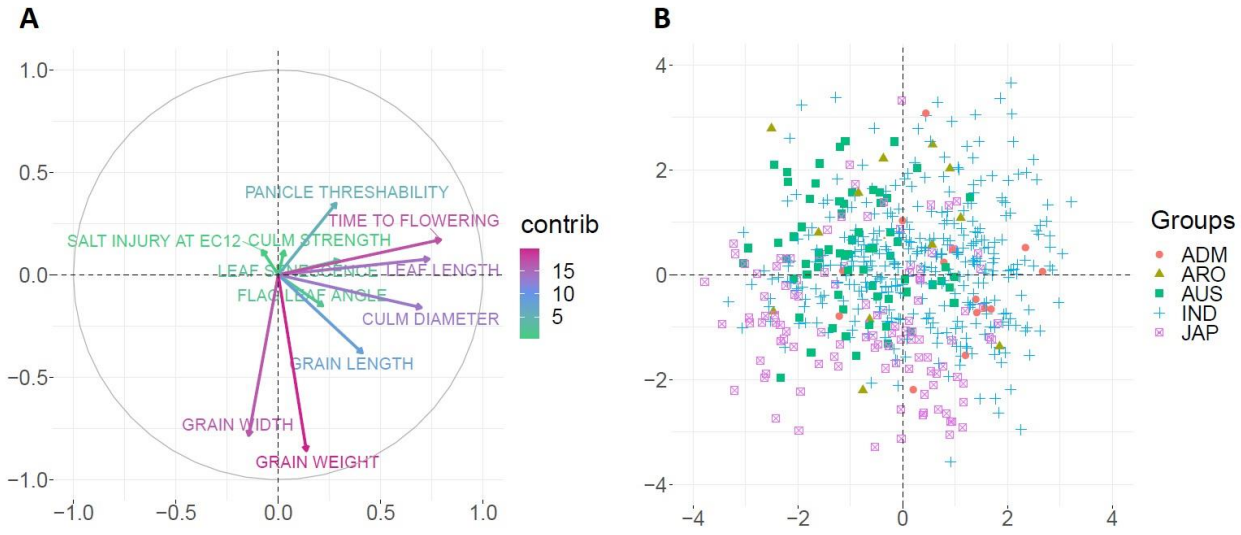

Figure 4.1: A: PC loadings of each trait for the two first standardized principal components. B: Plot showing the accessions projected. The first (x-axis) and second (y-axis) PCs explained 19% and 15.8% of variance, respectively.

Genetic variance estimates

The genetic variance explained by each marker set measures its relative importance in determining the observed phenotypes. Here we prefer not to use the classical term 'heritability'

because a proper interpretation assumes panmixia, a condition not fulfilled here. Having these cautionary remarks in mind, Table 4.1 does show that transposons can explain a sizeable fraction of genetic variance, which was larger than that explained by SNPs in five out of 11 traits. In seven traits, σ_S^2 was smaller than the sum of σ_M^2 and σ_R^2. Results are presented when all accessions were analyzed together and when using only data from Indica, the largest group (N = 451). Note model 1a assumes constant genetic variances across accessions, i.e., σ_S^2, σ_M^2, and σ_R^2 are the same in all rice groups. This is not necessarily the case. Nevertheless, variances were similar within Indica and across population groups.

Table 4.1: Means of posterior distributions of genetic variances explained by each marker set.

	All accessions(N=738)			**Indica accessions(N=451)**		
Traits	$\boldsymbol{\sigma_S^2}$	$\boldsymbol{\sigma_M^2}$	$\boldsymbol{\sigma_R^2}$	$\boldsymbol{\sigma_S^2}$	$\boldsymbol{\sigma_M^2}$	$\boldsymbol{\sigma_R^2}$
Culm Diameter	0.16	0.17*	0.16	0.13	0.17*	0.15
Culm strength	0.10	0.25*	0.16	0.11	0.19*	0.14
Flag leaf angle	0.22*	0.14	0.15	0.24*	0.14	0.14
Grain length	0.48*	0.11	0.11	0.41*	0.11	0.13
Grain width	0.49*	0.11	0.12	0.42*	0.11	0.14
Leaf length	0.26*	0.16	0.19	0.22*	0.16	0.19
Leaf senescence	0.12	0.25*	0.18	0.14	0.21*	0.16
Grain weight	0.40*	0.11	0.13	0.31*	0.12	0.13
Salt injury	0.10	0.11	0.12*	0.09	0.11*	0.11*
Time to flowering	0.45*	0.12	0.13	0.39*	0.13	0.13
Pan. threshability	0.11	0.13*	0.10	0.11	0.15*	0.11

σ_S^2: genetic variance explained by SNPs.
σ_M^2: genetic variance explained by DNA transposon markers (MITE/DTX).
σ_R^2: genetic variance explained by retrotransposons (RLX/RIX).
Traits are scaled such that phenotypic variances are 1.
** Best strategy*

Genomic prediction

We assess prediction in two validation scenarios that represent some of the main challenges in breeding, prediction within and across populations (see methods). In the first one, Indica improved varieties were predicted using the rest of accessions, including traditional Indica varieties. In this scenario, using TIPs increased prediction accuracy compared to using SNPs in six traits: culm diameter, grain length, leaf length, leaf senescence, grain weight and time to flowering (Figure 4.2). In the second scenario, phenotypes of all ADM and ARO accessions were predicted given the rest of the accessions. TIPs were especially beneficial in this case: TIPs improved prediction upon using only SNPs in nine out of the 11 traits analyzed (Figure 4.3). In some traits, such as grain width or leaf senescence, improvement in correlation using TIPs was remarkable, over 30%. In other traits, such as time to flowering, improvement was marginal. For some traits, notably grain weight and panicle threshability, prediction across populations was successful neither with SNPs nor with TIPs. We computed the bootstrap probability that using TIPs, or all markers resulted in better predictions than using only SNPs (see methods). Results are in Supplementary Tables B.4 and B.5 for the within and

across population scenarios, respectively. Even if gains in accuracy shown in Figures 4.2 and 4.3 may seem small in some cases, results are consistent. For instance, increase in correlation for leaf length is ~15% when using MITE/DTX compared to SNPs in the within population scenario, a somewhat modest figure. But this result is confirmed in 80% of the bootstrap samples. In contrast, SNPs are far better than TIPs for culm strength and this is also confirmed in bootstrap samples (Supplementary Table B.4, Figure 4.2).

On average, prediction across populations was less accurate than within Indica in seven out of 11 traits and irrespective of marker set used (Figures 4.2, 4.3). Importantly, gain using TIPs was larger in this scenario than in the within population scenario. Time to flowering and grain width were the traits for which prediction was most accurate. Nevertheless, prediction across populations for grain width was far less precise than within Indica. It is interesting to note that grain width and time to flowering are basically uncorrelated, but both contribute largely to total phenotypic variation (Figure 4.1). This suggests that genomic prediction combined with

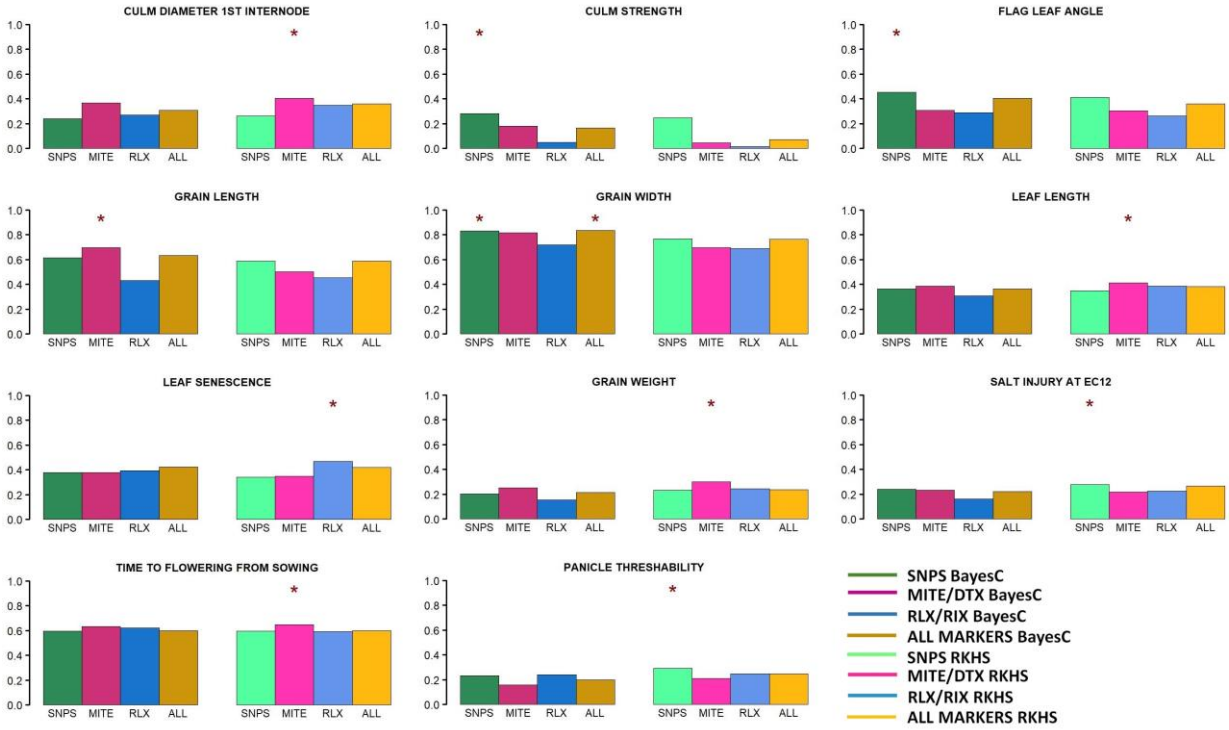

Figure 4.2: Correlation between observed and predicted phenotypes of Indica improved varieties. In each plot, the first four columns represent the correlation values using BayesC, while the last four values correspond to RKHS method. Colors represent marker information utilized: Green, SNPs; Magenta, MITE/DTX; Blue, RLX/RIX; Brown, all markers. The asterisc shows the best option for each trait.

transposable elements can be an effective tool for overall rice genetic improvement as it would enhance genetic progress in important agronomic traits. Note that using all markers is not necessarily the best option for predictive purposes: it only outperformed the rest of approaches in three out of the 44 (= 11 traits x 2 methods x 2 predictive scenarios) analyses. This indicates that adding additional markers may contribute to overfitting and reduce model performance in prediction. Overfitting is a well-known phenomenon in the machine learning literature when the model is not properly

regularized. This has been clearly observed with simulated data in a genomic prediction scenario (e.g., Pérez-Enciso et al. 2015).

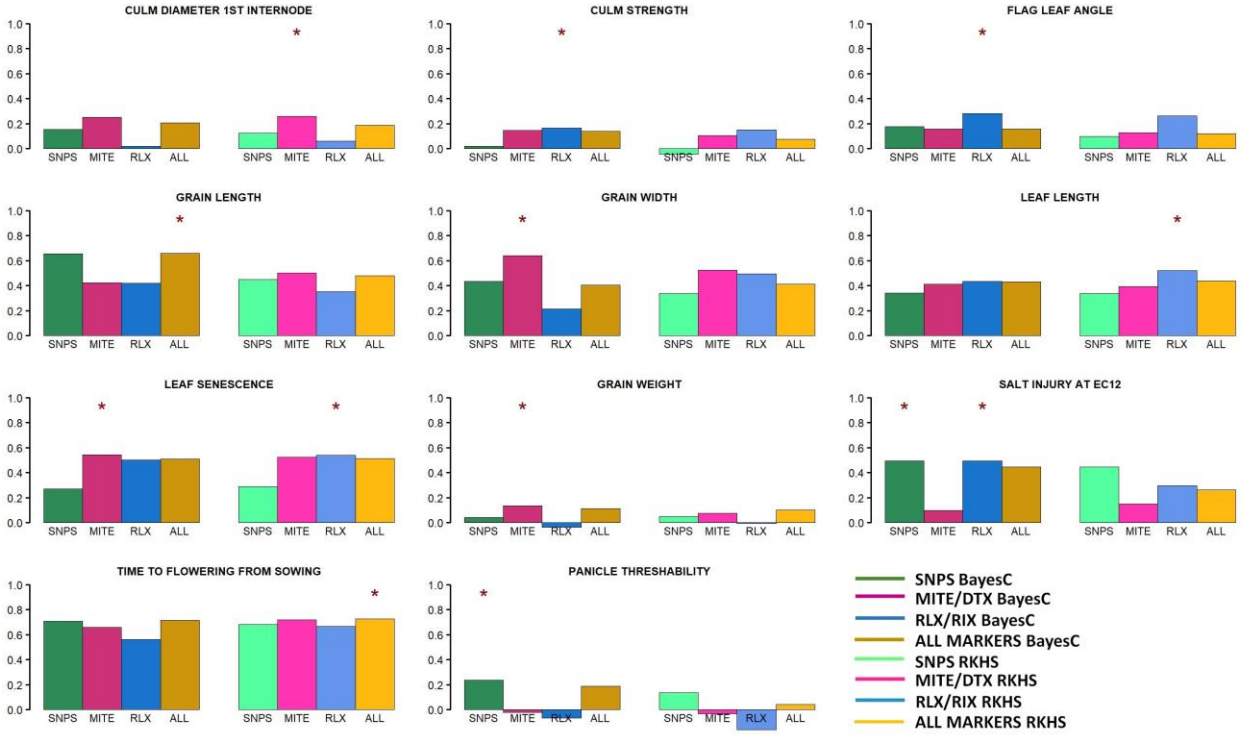

Figure 4.3: Correlation between observed and predicted phenotypes across accessions. All ADM and ADM accessions were predicted using the rest of groups. In each plot, the first four columns represent the correlation values using BayesC, while the last four values correspond to RKHS method. Colors represent marker information utilized: Green, SNPs; Magenta, MITE/DTX; Blue, RLX/RIX; Brown, all markers. The asterisc shows the best option for each trait.

Next, we wished to study how the different genetic architectures influence the statistical behavior of the three sets of markers. BayesC is a variable selection method and so we reasoned that the number of markers entering the model and their effects would differ between traits. Broadly, estimates of marker effects were quite similar across traits (for the same type of marker) as can be seen in Supplementary Figure B.3. The only exception was grain width and grain length, where we observed much larger estimated effects for MITE/DTX and SNPs respectively, in agreement with results in Figure 4.3. In turn, there were larger differences between the probabilities (d) of entering the model for each marker type (Supplementary Figure B.4). This occurred despite setting equal priors for all types of markers (p = 0.01). This was not due only to the priors or different number of TIPs compared to SNPs, because the pattern differed between traits. Using a subset of all markers available can improve prediction. For instance, the accuracy of a model which contains only the causative SNPs can approach one (Pérez-Enciso et al. 2015). The problem, of course, is that causative mutations cannot be identified in most cases. Several indirect approaches have been suggested instead.

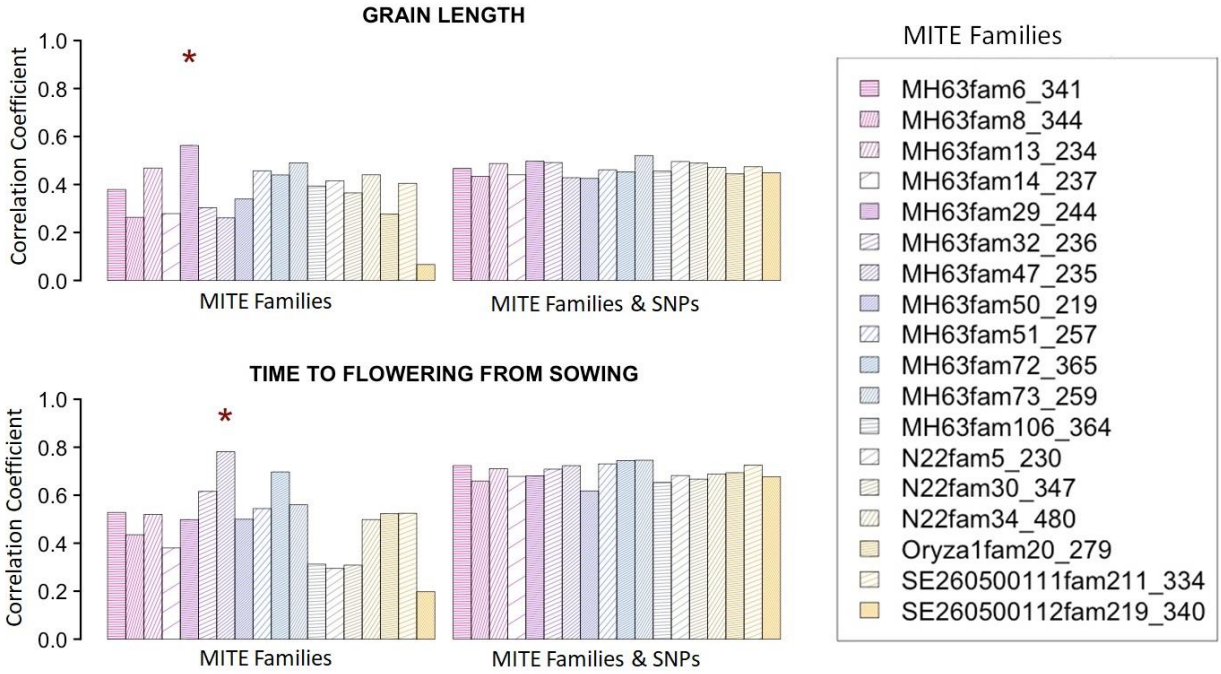

Figure 4.4: Predictive accuracy across populations using TIPS from each of 18 recognized MITE families. Each column corresponds to accuracy with one MITE family. Model included only MITEs or MITEs and all SNPs. The asterisc shows the best option.

For instance, Spindel et al. (2016) proposed to perform prediction using the most associated markers, e.g., selected via a GWAS P-values. We did not evaluate this strategy here, although we did consider two alternative approaches for preselecting markers. In a first attempt, we examined whether using only gene-based markers improved prediction performance. To avoid multiplying analyses, we selected grain length and time to flowering. As can be seen in Table 4.2, gene – based markers outperform all markers across population but minimally. The opposite was observed in the within population scenario.

We also studied performance of TIPs pertaining to each of the 18 largest MITE families (Supplementary Table B.3, see Material and Methods). Again, for brevity, we considered only prediction across accessions in grain length and time to flowering using RKHS (Figure 4.4). The most relevant conclusion is that predictive performance can vary largely according to MITE family and that using SNPs on top of MITEs may not improve prediction. Prediction of time to flowering improved using MITE family MH63fam47_235 (MITE-adh type B-like superfamily) TIPs compared to using the full MITE/DTX set (Figures 4.3, 4.4). Although it is tempting to conclude that a specific MITE family is enriched in genes affecting a given trait, one should be careful as disequilibrium can extend over long genome regions (Mather et al. 2017; Nachimuthu et al. 2015).

Table 4.2: Predictive accuracy when using all or only gene-based markers.

Prediction scenario	Trait[a]	Markers	BayesC				RKHS			
			SNPS	MITE/ DTX	RLX/R IX	ALL	SNPS	MITE/ DTX	RLX/R IX	ALL
Across	GL	Genic	0.57	0.51	0.41	0.67*	0.47	0.54	0.26	0.48
		All	0.65	0.42	0.42	0.66	0.45	0.50	0.35	0.48
	TF	Genic	0.71	0.58	0.50	0.68	0.71	0.65	0.69	0.75*
		All	0.71	0.66	0.56	0.71	0.68	0.72	0.67	0.73
Within	GL	Genic	0.56	0.55	0.43	0.57	0.56	0.38	0.36	0.55
		All	0.61	0.69*	0.43	0.63	0.59	0.50	0.45	0.59
	TF	Genic	0.59	0.55	0.51	0.62	0.57	0.61	0.56	0.57
		All	0.59	0.63*	0.62	0.60	0.59	0.65	0.59	0.60

* Best strategy

[a] GL: Grain Length; TF: Time to Flowering

Discussion

We have shown, for the first time to our knowledge, that transposable element polymorphisms can improve predictive accuracy for important agronomic traits in rice. The impact of using TIPs varied; here we found that they improved predictive performance in ~ 60% of the traits and scenarios considered. Table 4.3 presents a summary. The increase in accuracy also varied. Although the added benefit of using TIPs was sometimes modest, TIPs improved correlation by more than 30% in traits like grain width or leaf senescence.

All traits analyzed here have an economic impact in rice production. Unfortunately though, grain yield phenotypic data is not available for the 3k rice panel, and how grain yield is affected by TIPs remains to be studied. This trait is largely affected by genotype x environment interaction and so the relevance of TIPs may be harder to characterize. Among the traits studied, time to flowering is particularly important (Wang and Li 2019). Rice plants need approximately, 3-6 months to grow, meaning that earlier or later growing can strongly affect the yield. Productivity is also determined by morphological trait such as grain weight (Chen et al. 2021). Grain weight in turn correlates with grain width Figure 4.1 (Li et al. 2021). Most of these traits are polygenic. Some traits like time to flowering, grain weight, grain width and grain length seem controlled by large effect quantitative trait loci (Begum et al. 2015; Xu et al. 2015; Chen et al. 2021). For some traits, e.g., grain width, GP was quite accurate, and we confirm that GP can largely enhance rice genetic progress, in agreement with previous results (Xu et al. 2021). For other traits, e.g., leaf length, GP accuracy was lower, although it is interesting to note that bootstrap sampling suggests that results are repeatable (Supplementary Tables B.4, B.5). Since plant breeding builds on cumulative progress over generations, even a small advantage can be highly relevant in the medium to long term.

The reasons behind the high capacity of TIPs to predict phenotypes, which in some cases is far better than SNPs, could be manifold. Transposable element insertions can have stronger effects than SNPs as some transposon types tend to localize near genes.

Table 4.3: Maximum predictive accuracy and corresponding marker set.

	Scenario	
Trait	**Indica improved varieties**	**ARO/ADM accessions**
Culm Diameter	0.40 (MITE/DTX)	0.26 (MITE/DTX)
Culm strength	0.28 (SNPs)	0.16 (RLX/RIX)
Flag leaf angle	0.45 (SNPs)	0.28 (RLX/RIX)
Grain length	0.69 (MITE/DTX)	0.66 (ALL)
Grain width	0.83 (SNPs, ALL)	0.64 (MITE/DTX)
Leaf length	0.41 (MITE/DTX)	0.52 (RLX/RIX)
Leaf senescence	0.47 (RLX/RIX)	0.54 (MITE/DTX, RLX/RIX)
Grain weight	0.30 (MITE/DTX)	0.14 (MITE/DTX)
Salt injury	0.28 (SNPs)	0.49 (SNPs, MITE/DTX)
Time to flowering	0.65 (MITE/DTX)	0.73 (ALL)
Panicle threshability	0.29 (SNPs)	0.24 (SNPs)

Therefore, TIPs could be in some cases causative mutations linked to a specific trait. Indeed, transposable element insertions are known to have played a major role in plant genome evolution both in the wild and under breeding settings, and examples of TIP causative mutations for many agricultural important traits have been reported (Lisch 2013; Dubin et al. 2018). In some cases, the TIPs linked to the trait may be recent insertions and may not be in high LD with surrounding SNPs. This is what was shown in recent GWAS analyses performed with TIPs and SNPs in tomato and rice, where TIPs revealed associations with traits that are not detected with SNPs (Domínguez et al. 2020; Akakpo et al. 2020; Castanera et al. 2021). In contrast to SNP mutation rate, transposon activity is not constant over time, with bursts of transposition associated with stress situations or environmental stimuli (Dubin et al. 2018). Therefore, it can be hypothesized that the adaptation of a crop to a new environment, say as part of the breeding process, could be a period particularly prone to transposition activity (Baduel and Quadrana 2021). On the other hand, while SNPs accumulate relatively homogeneously throughout the genome, some TEs target gene-rich regions for integration, particularly RLXs and MITEs in rice (Castanera et al. 2021). Therefore, the potential for TEs to produce causal mutations and TIP associations with traits could be particularly high for some agronomic traits. Importantly, we found TIPs are especially helpful when prediction was across populations. These issues merit further research.

The main families of class I in rice are LTR-retrotransposons (RLX) and LINEs (RIX), whereas DNA transposons (DTX) and MITEs are the main components of rice class II TEs (Matsumoto et al. 2005). There are important structural and mechanistic differences between class I, or retrotransposons, and class II, or DNA transposons. Although both RLX and MITEs target genic regions for integration, their dynamics are very different. While RLXs have a high turnover and RLX TIPs are usually present at a very low frequency in the population, MITEs are maintained in the genome for longer evolutionary periods (Castanera et al. 2021). This suggests that, although both types of TEs can be associated with traits in rice (Akakpo et al. 2020; Castanera et al. 2021), their capacity to predict phenotypes may differ. Certainly, our results show that MITE/DTX are more relevant than RLX/RIX for improving prediction (Table 4.3, Figures 4.2, 4.3). It is finally interesting to note that a single MITE family of ~ 3k TIPs can predict equally well a phenotype as well as 200k SNPs (Figure 4.4). In contrast,

we did not find a consistent or large improvement in prediction when using only gene markers as compared to using all available polymorphisms, as reported also in humans (Visscher et al. 2021).

Some technical considerations should be borne in mind regarding our analyses. Ordinal traits (Supplementary Figure B.2) were treated as continuous. It has been known for decades that a threshold (logistic model) is theoretically a better choice for binary traits than standard linear models (Gianola and Foulley 1983). The logistic model is a class of the so called generalized linear models, where the non-linear relationship between parameters and observations becomes linear after applying a transformation, e.g., logit for binary traits. Despite their theoretical appeal, these models are more difficult to run than linear counterparts and may converge poorly. Empirical evidence generally shows small differences only (Matos et al. 1997; Olesen et al. 1994). Here, we observed (Supplementary Table B.4) that a threshold model may have a small advantage over linear ones but not always. A second issue is the metrics to assess prediction. Here we chose correlation as it has a direct interpretation in terms of response to selection (Falconer and Mackay 1996) and has been widely used, but numerous other metrics exist. For instance, mean square error (RMSE) of prediction is also widely used. We computed RMSE (Supplementary Table B.7, B.8) and we found concordant results regarding the best marker set in 9 (within scenario) or 10 traits (across scenario) out of the 11 traits studied. These issues do not question our main, and most important conclusion regarding that TIPs can improve genomic prediction.

A prerequisite for the inclusion of TIPs in practical breeding programs is to automatize their genotyping. TIP genotyping should primarily target high frequency TIPs in order to be as informative as possible, as it is usually done for SNPs as well. The application of TIP-Chip (Wheelan et al. 2006) or Transposon Insertion Profiling (TIP-seq, Steranka et al. 2019), and TE-sequence capture (Quadrana et al. 2021) to hundreds or thousands of varieties should be cheap, as the sequencing coverage needed per sample is very small. Finally, given the dropping costs of genome sequencing, thousands of rice accessions are being re-sequenced and made public. TIPs could also be included in standard genotyping arrays (Wheelan et al. 2006) as a complement to SNPs. Given that TIPs from a single MITE family can be as efficient as 200k SNPs in some traits (Figure 4.4), perhaps only a small number of TIPs need to be included in the genotyping protocol.

In conclusion, we consistently observed that TIPs can increase predictive accuracy of agronomic traits in rice and do explain a non-negligible fraction of phenotypic variance. Notably, this improvement was larger when prediction was across populations than within Indica. Using markers positioned within genes did not seem to matter too much, although perhaps a more thorough analysis would be needed. In contrast, selecting TIPs from some transposon families did improve prediction. These are important results from a practical point of view and warrants developments to automatize TIP genotyping. From a biological point of view, new studies are needed to understand how TIPs affect complex trait variation. Improving predictive accuracy from molecular data is an important task since even small gains add up over generations and can make a big long-term difference. Assessing the importance of TIPs in other agronomic traits, such as grain yield across different environments, remains also to be studied. Once a plausible set of parameters linking TIPs, SNPs and yield are estimated from real data, simulations can be used to optimize marker genotyping with SNPs and/or TIPs.

Chapter 5

Merging structural and nucleotide genome-wide variation for genomic prediction in rice

(in preparation)

Abstract

Using Bayesian linear models, we have shown that Transposable Insertion Polymorphisms (TIPs) can improve prediction ability in genomic prediction of complex agronomic traits in rice over standard approaches based exclusively on Single Nucleotide Polymorphisms (SNPs). However, TIPs are not the only structural variation in the genome. Structural variations (SVs) such as deletions, inversions, and duplications are prevalent in the genome and they play an important role in plant evolution. Here, we determine the proportion of genetic variance explained by different types of structural variation. Then, we investigate whether merging the structural and nucleotide genome-wide variation can improve prediction ability of traits when compared to using only SNPs. For the purposes of the study, four important agronomic traits were used from 738 rice accessions in total, originated by five different rice population groups (Aus/Boro, Indica, Aromatic, Japonica and Admixed). We assess prediction accuracy by applying cross validation under two different strategies. In the first strategy, we used a k-fold cross validation producing ten partitions from the whole population. In the second strategy, we followed an across population scenarios predicted Aromatic and Admixed accessions from the rest of populations. In each scenario, the performance of BayesC and a Bayesian Reproducible Kernel Hilbert space regressions are compared to Deep Learning networks (DL). We investigated the prediction ability of DL using two different widely used architectures, a Multilayer Perceptron (MLP) and a Convolution Neural Network (CNN). Then we further explored their performance by using various marker input strategies. We found that merging structural and nucleotide variation improves prediction ability on complex traits in rice. Also, our results suggested that DL models outperform in 50% of the studied cases. Finally, DL seems to significantly improve prediction ability of continuous traits against the Bayesian models when training and dataset are distantly related.

Introduction

Rice (*Oryza sativa*) provides a staple food for more than half the word population. However, following the conventional breeding techniques, rice yield cannot meet the high demand caused by the increasing word population and the climate change. Therefore, we need methods that will secure nutritional requirements increasing at the same time the quality and quantity of rice yield. Moreover, the new cultivars must have two important traits: disease resistant and climate resilient. Genomic Prediction (GP) can help achieving all the pre-mentioned requirements, accelerating the breeding progress (Meuwissen et al. 2001). Various studies in plants have shown the effectiveness of GP in increasing breeding speed (Jighly et al. 2019, Tessema et al. 2020, Xu et al. 2020, Krishnappa et al. 2021). GP framework has widely used in rice studies for predicting various quantitative traits, reporting moderate to high predictive performance (Xu et al. 2021). Complex traits are controlled by numerous loci that are difficult to be detected with genetic mapping. GP assumes that quantitative trait loci (QTL) will be in linkage disequilibrium (LD) with at least one molecular marker. Thus, instead of detecting all the QTL associated with a trait an indirect association between marker and trait can be utilized.

Conceptually, since the number of genotyped individuals *n,* is typically smaller than the number of molecular markers *p,* GP faces statistical challenges such as large sampling variance and

increase mean-square error. To overcome this limitation, variables must be selected or restrictions on the solutions must be applied or sometimes both. The main classes of GP methods are the genomic relationship-based method such as Genomic Best Linear Unbiased Prediction (GBLUP, VanRaden 2008) and the SNP effect-based methods such as the Bayesian family (Meuwissen et al. 2001; Habier et al. 2011; Pérez and De Los Campos 2014) and LASSO (Tibshirani 2011). Particularly, Bayesian models don't assume necessarily homogenous across marker effects. They perform variable selection and shrinkage on the effects simultaneously using priors other than Gaussian. BayesC is an example of this category assuming as a prior a normal distribution with constant variance while a fraction of marker has no effect (Habier et al. 2011). On the other hand, methods such as GBLUP involve restriction on the square of solutions (L2 norm), with the effect of the markers assuming to be normally distributed with equal variance.

Deep Learning (DL) networks are a collection of machine learning algorithms that have exhibited excellent performance in some prediction tasks (Min et al. 2017; Pattanayak 2017). The DL models are trained in such a way to find complex relationships between traits. DL networks consist of multiple layers and interconnected nodes. Each layer uses as input the output of the previous layer in order to optimize the prediction or classification. Numerous DL architectures have been proposed such as Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN, Lecun et al. 2015). DL has been around for decades but only recently started to widely be implemented because of the easy implementation framework provided by various online libraries (e.g https://keras.io/; https://pytorch.org/). The performance of the DL networks depends on the accurate hyperparameter choice, which is not an easy task and requires abundant computation resources (Young et al. 2015; Chan et al. 2018).

Despite their features, various works have shown a performance of DL in genomic prediction comparable to linear models (González-Recio et al. 2014; Ma et al. 2017; Bellot et al. 2018; Montesinos-López et al. 2018). Zingaretti et al. (2020) did not find a considerable advantage of DL over linear models, except when epistasis component was important. Note that DL can be used to estimate non-additive effects without the need to partition the effects as in standard linear models. Ehret et al. (2015) found non-relevant differences between a GBLUP and a MLP model. In a wheat study (Ma et al. 2017), DL performed better than GBLUP when used to predict phenotypes from genotypes. Similarly, Gianola et al. (2011) found that MLP performed better than a Bayesian linear model in wheat. In another study in wheat, Pérez-Rodríguez et al. (2012) extensively compared the prediction performance of Radial Basis Function Neural Networks and Bayesian Regularizes Neural Networks against several linear models and semiparametric models such as Reproducible Kernel Hilbert Space. The authors concluded that the non-linear models, such as DL, demonstrated a higher prediction ability than the linear models. For an extensive review in GP using DL models see Montesinos-López et al. 2021.

Most of the studies in GP assume that SNPs are the main source of genetic variability at the whole-genome level in plants. In Vourlaki et al. (2022) we showed that Transposable Insertions Polymorphisms (TIPs) explain a sizable fraction of the genetic variance in the agronomic traits in rice and significantly improved the prediction of phenotypic traits of interest. TIPs account for a major fraction of intraspecific structural variation, as a study in maize recently showed (Haberer et al. 2020). Studies in tomato and in rice found that the use of TIPs can increase association signals compared to

SNPs (Akakpo et al. 2020; Carpentier et al. 2019; Domínguez et al. 2020; Castanera et al. 2021). TIPs play a key role in plant evolution since selection acts on them during local adaptation, speciation, domestication, and breeding (Dublin et al. 2019).

In all, transposable elements (TEs) are not the only type of structural variation in the genome. Structural variations (SVs) such as deletions, inversions, and duplications form an important fraction of genetic variation in plant species. Over the last few years, studies have been focusing on the importance of presence-absence variation and structural variation as a source of phenotypic variability in plants, including in rice. In rice a total of 63 million individual SV calls that grouped into 1.5 million allelic variants, have been identified across the 3,000 Rice Genomes dataset (Fuentes et al. 2019). Fuentes et al. (2019) showed that rice genome regions with frequent SVs were enriched in stress response genes. Here we investigate whether merging all the structural and nucleotide genome-wide variation can improve phenotypic prediction comparing only to SNPs in rice. Finally, we further explore the performance of DL in GP by (i) using multiple marker input strategies, (ii) proposing several approaches to accommodate large scale marker information, (iii) optimizing network architectures. We also provide and document python code based in tensorflow 2 and keras.

Materials and Methods

Rice accessions and traits

In this study we used 738 accessions from the collection conserved at IRRI used for the 3,000-rice genome project (Jackson 1997; Li et al. 2014). Chosen accessions were sequenced at least at 15x depth. The 738 accessions originated by all main rice population groups: Aus/Boro (AUS, N=75), Indica (IND, N=451), Japonica (Jap,N=166), Aromatic (ARO, N=17). The final group is the Admixed (ADM, N=29) consists of accessions that cannot be assigned to a specific rice group. SNP-based group assignment from Sun et al. (2017) was used to identity the different subsets of this study. Studied traits were originally available at IRRI SNP-Seek database (https://snp-seek.irri.org/). From the eleven traits analyzed in Vourlaki et al. (2022) we selected four that span a range of distinct distributions between SNPs and SVs and are either continuous or binary. For continuous traits, grain weight and time to flowering were used, whereas for binary traits, we chose culm diameter and leaf senescence. Binary traits were binned to balance the number of observations per class and time to flowering was log-transformed.

Markers

We used the filtered SNP dataset in Vourlaki et al. (2022). Specifically, a binary ped file format with Core SNP dataset for all chromosomes was downloaded from the SNP-Seek database. The original dataset consisted of 404,399 bi-allelic SNPs from 3,034 rice accessions, including the 739 accessions selected. After filtering (Vourlaki et al. 2022), the final dataset consisted of 228,871 SNPs.

Transposable Elements (TEs) are divided into two main classes, Class I and Class II, based on the mechanism of transposition. In rice, we can find TEs from both classes. Specifically, in rice the most prevalent elements from Class I are RLX (LTR retrotransposons) and RIX (Non-LTR retrotransposons) whereas the most representative superfamilies of Class II are MITEs (Miniature inverted-repeat transposable elements) and the DTX (DNA TEs with terminal inverted repeats) (Mao

et al. 2020). Here we used markers from both classes, making 94% of the TIPs described in Castanera et al. (2021). Class I TIPs were represented by 21,571 RLX and RIX markers. Class II consisted of 52,120 MITE and DTX markers. In contrast to SNPs, TIPs can only be genotyped as presence / absence, recoded consequently as 0/1, and defined as genomic windows with an average size of 1.2 kb. TIP windows were taken from Castanera et al. (2021) and are based on the intersection of the individual TE insertion regions predicted for each accession with genome-wide windows of a fixed size (1kb, merging adjacent windows).

SVs such as insertions (INS), deletions (DEL), tandem duplications (DUP), inversions (INV) and copy number variants (CNVs) were also obtained by Rice SNP-Seek database. SVs genotypes are also recoded as 0/1 and defined as various genomic windows depending on the type. Particularly, minimum window size for INS was 5 bp, for DEL, DUP and INV was 10 bp. CNVs were discarded because of the high missing rate (58 % on average) and INS because of the low accessions availability (present in 390 accessions). Markers with minor allele frequency ≤ 0.01 were filtered out using plink2 (Purcell et al. 2007; Chang et al. 2015). Finally, the dataset used in our analysis consists of 139,229 DEL, 14,638 DUP and 6,083 INV.

Genetic variance inference

To estimate the genetic variance components explained by each marker set, we fitted the following linear model using RKHS (Gianola et al. 2006):

$$\boldsymbol{y = \mu + Z u_1 + Z u_2 + e} \qquad \textbf{(5.1)}$$

where μ is the general mean, y is the phenotype vector of size n (the number of accessions), Z is an identity incidence matrix, u_1 and u_2 are random effects of each of the marker groups and e is the residual. Random effects are assumed to be normally distributed $u_1 \sim N(0, K_1\sigma_1^2)$, $u_2 \sim N(0, K_2\sigma_2^2)$, with constant variance $K_1\sigma_1^2$ and $K_2\sigma_2^2$. Where K_1, K_2 are genomic relationship matrices (GRM) obtained from the markers used in the corresponding model. We fitted the model five times using as K_1 the GRM form SNPs while as K_2, GRM was obtained from MITE-DTX, RLX-RIX, DEL, DUP, INV, successively. The GRM were calculated using AGHMatrix (Amadeu et al. 2016). Model was implemented in BGLR package (Pérez and de Los Campos 2014) using default priors to estimate σ_1^2, σ_2^2.

Genomic Prediction Models

Bayesian Regression Models

Two Bayesian methods are employed in this study: Bayesian RKHS and BayesC. RKHS is a method that does not directly estimate the effect of the markers while using a ridge regression L2 regularization technique like GBLUP. BayesC is a variable selection method that estimates the effect of the markers. Both methods applied to each trait separately. Particularly, for each method, two different models were designed and applied comparing the predictive performance of using all the markers together versus using only SNPs. For RKHS, the models are described as follows:

$$\boldsymbol{y = \mu + Z u_1 + Z u_2 + Z u_3 + Z u_4 + Z u_5 + Z u_6 + e,} \qquad \textbf{(5.2)}$$

$$\boldsymbol{y = \mu + Z u_1 + e} \quad \textbf{(5.3)}$$

Where u_1, u_2, u_3, u_4, u_5, u_6 the six GRM matrices for each marker, SNPs, MITE-DTX, RLX-RIX, DEL, DUP, INV, respectively. For BayesC the complete models were:

$$\boldsymbol{y = \mu + X_{10k} \beta_1 + e} \quad \textbf{(5.4)}$$

$$\boldsymbol{y = \mu + X_{SNPs} \beta_1 + e} \quad \textbf{(5.5)}$$

Where X_{10k} are the standardized genotypic values for the 10,000 most associated markers among SNPs, MITE-DTX, RLX-RIX, DEL, DUP, INV, X_{SNPs} is the standardized genotypic matrix for the 10,000 most associated SNPs, β_1, is the vector of effects for the corresponding matrix. The method based on which the 10,000 most associated markers are selected, is described in more detail in the following sections.

Using either RKHS or BayesC, phenotypes to be predicted were removed from the dataset and the model fitted using the remaining phenotypes. Prediction ability was assessed by computing two different metrics related to the type of the trait. We computed the mean squared error (MSE) between predicted and observed phenotypes for the quantitative traits, whereas the binary cross-entropy was employed for the binary traits. Both models were implemented using BGLR package. BayesC assumes that a proportion of markers will have zero effect with probability sampled from a beta distribution, $\pi \sim Beta(p_0, \pi_0)$. The beta prior is parameterized in a way that the expected value by $E(\pi) = \pi_0$; on the other hand, p_0 can be interpreted as the number of prior counts (prior "successes" plus prior "failures") (Pérez and de Los Campos 2014). Here we chose p_0 = 5 and π_0 = 0.01. For the case of binary traits option "response_type=ordinal" was applied in both methods (RKHS, BayesC). Finally, BGLR was run for 100,000 iterations using default priors for RKHS.

Multilayer Perceptron

One of the most popular DL architectures is the Multilayer Perceptron (MLP). MLP is a fully connected feedforward artificial neural network which transforms any input dimension to the desired dimension. The basic structure consists of an input layer, multiple hidden layers, and an output layer. Each layer consists of neurons, that is, a mathematical function that transforms the data received before passing them forward. Each neuron connects with a weight to every weight to the next layer. All the neurons are connected to every neuron in the previous layer and then connected to every neuron in the next layer. Particularly, let us consider the first hidden layer. Each neuron of this first hidden layer receives the initial inputs multiplied by a corresponding weight coefficient. Then the sum of all inputs multiplied by weight plus a bias, is passed to an activation function which introduces the non-linearity to the network transforming the inputs accordingly. The product of the activation

function is the output of the neuron. We can represent the output of the first hidden layer as (note the transposes):

$$\boldsymbol{Z_1 = f\left(XW^{(0)T} + b^{(0)T}\right)} \quad \textbf{(5.6)}$$

Where $Z_1$is the output of the first layer, $b^{(0)T}$ is the bias vector of the first layer, X is a single matrix of all training examples so that we could compute all the prediction using a single matrix multiplication, $W^{(0)T}$ is the weight matrix and f is a nonlinear activation function. The model is trained successively, that is, the output of neurons from the previous layer will be the input for the next layer. The output of each layer is then formed as follows:

$$\boldsymbol{Z_l = f\left(XW^{(l-1)T} + b^{(l-1)T}\right)} \quad \textbf{(5.7)}$$

Where $\boldsymbol{Z_l}$ is the output of layer l. Figure 5.1 shows a basic workflow of MLP network.

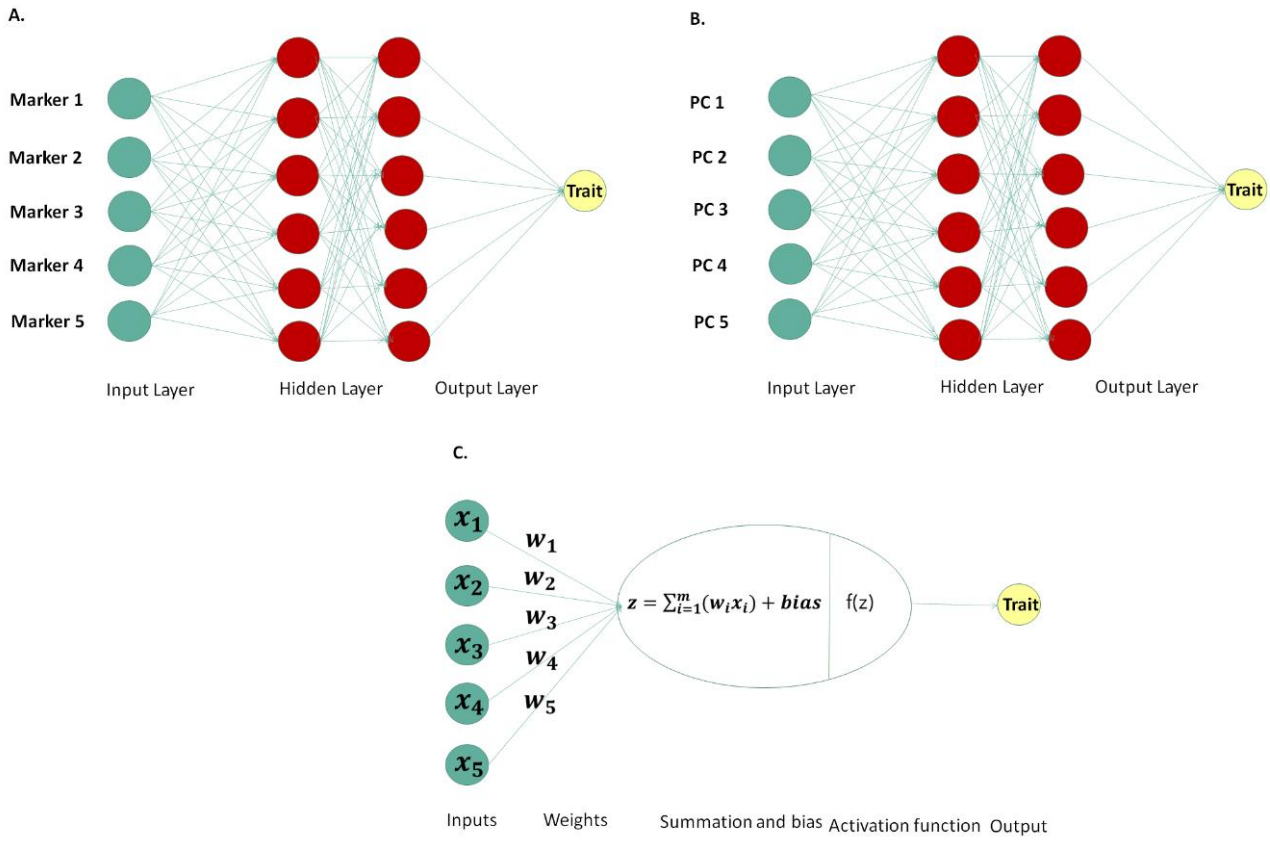

Figure 5.1: Multilayer perceptron (MLP) representation with markers (A) and Principal Component (PC) (B) as input layers; bottom center shows the basic workflow of a perceptron or else neuron (C).

Convolutional Neural Networks

Convolutional neural networks (CNNs) can utilize spatial relationships between nearby variables (e.g., pixels) of the input matrix. In general, CNNs work well with data that has a spatial relationship. This architecture can accommodate situations where input variables are distributed

along a space and are associated with each other such as linkage disequilibrium between nearby markers (Pérez-Enciso and Zingaretti 2019). A CNN is a special case of neural networks which uses convolution instead of a full matrix multiplication in the hidden layers. A CNN has hidden layers which typically consist of convolutional layers, pooling layers, flatten layers and fully connected dense layers. In each convolutional layer, CNN automatically performs a convolution that is a linear operation performed along the input of predefined width and strides by applying kernels or filters. The weights used are the same for all marker windows. The filter moves along windows of same sizes consist of markers performing a multiplication operation (dot product) until the entire matrix is traversed. The filters for CNN are the equivalent of the neurons in a MLP network and they are the learnable parameters of our network representing weights. The output of the convolutional function can be described as an integral transformation (Widder 1954), as follows:

$$\boldsymbol{s(t) = (f * k)(t) = \sum_x k(t-x) f(x)} \qquad \textbf{(5.8)}$$

where k represents the kernel, convolution is the transformation of f into $s(t)$. The operation is performed over an infinite number of copies f resulting in the weighted sum shifting over the kernel. An activation function is applied after each convolution to produce the output layer. After nonlinearity has been applied to the feature map produced by the first layer, a pooling layer usually follows, aiming to reduce the dimensionality and smoothen the representation. Particularly, it merges kernel outputs calculating their mean, maximum or minimum (Figure 5.2). The benefit of using CNN is their ability to develop an internal representation of a two-dimensional matrix extracting the most important features. CNN leverages the fact that nearby inputs variables are more strongly related than the distant ones. The layers in a CNN network are more sparsely than in fully connected MLP. Thus, CNN estimates a smaller number of hyperparameters than MLP which requires too many parameters forming a dense web.

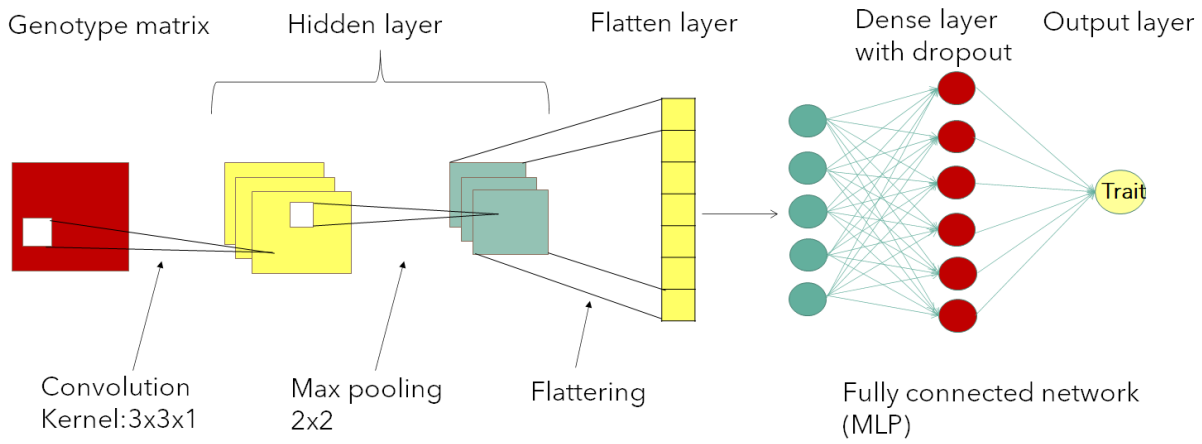

Figure 5.2: Convolutional Neural Network (CNN) representation used in study.

Cross-validation and Independent Prediction

Conventional plant breeding makes trials of new crosses whereas the long growing time of a new cultivar in rice, almost 10 years, results in a very low yield pace. GP can accelerate the breeding

process by predicting new improved varieties from the available genotypes. The prediction accuracy depends on many different genetic factors such as the marker density, LD between markers and QTL, sample size, the relationship between the training and test population and the heritability (Xu et al. 2021). Here we evaluate the prediction accuracy by following two challenging validation scenarios both of high importance in a breeding program: prediction of individuals from two different groups and prediction of randomly selected individuals from the rest ones. For the first strategy, we predicted performance of two distantly related groups, the admixed (ADM, N=29) and aromatic (ARO, N=17) using the rest accessions. Since, accessions to be predicted are not related to the accessions in the training set, it would be expected a low prediction ability from the models for this scenario.

In the second strategy, prediction accuracy was evaluated by implementing a 10-fold CV where training population consisted of 90% of the data and testing set included the 10% of the remaining data. Analysis performed to each of the ten training sets separately assuming ten different breeding scenarios. Since accessions are randomly selected and not based on their origin, samples in the training set might be related to the predicted ones. Note that, in the case of DL application, training population was further split in a validation dataset which included 20% of the training dataset (Figure 5.3). Validation dataset is used during the training process of our network to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. It is important to mention that the model "sees" the data and used them for an evaluation of the process but never "learn from these". After the model is trained, we can retrieve the best hyperparameters and perform prediction using the test dataset. The test dataset provides the gold standard used to make an unbiased final evaluation of the model. It is used only once a model is completely trained using the train and validation sets.

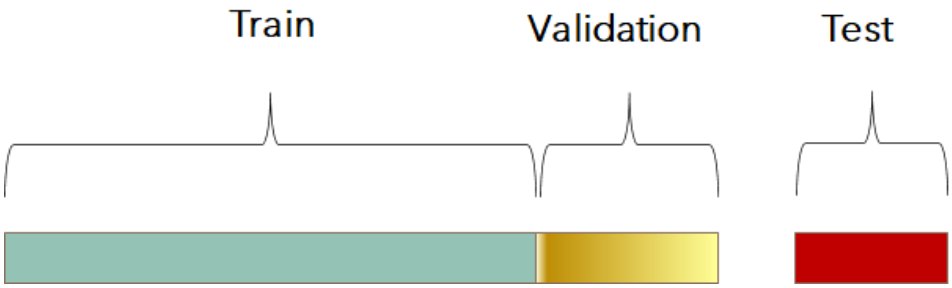

Figure 5.3: A visualization of how the three datasets, training, validation and test are divided.

DL input strategies

In the DL, the input layer consists of a fixed number of neurons where each neuron represents a marker in the training set. Here we explore different marker input strategies aiming to enhance network flexibility and thus improve prediction ability. Three strategies were designed as follows:

1. Most associated markers: In the first input strategy, we merged the structural and nucleotide genomic-wide variation to test whether prediction accuracy can be improved. However, using the whole six genotype matrices (SNPs, MITE-DTX, RLX-RIX, DEL, DUP, INV) would add a high complexity in our network that might cause an overfitting. Thus, from the 462,512 molecular

markers we selected the 10,000 most associated to the traits of interest. Specifically, we performed a genome wide association study (GWAS) fitting a linear model to find associations between each of the six-marker set and each of the four traits (4x6). For each fitted model, a p-value corresponding to each marker was collected. From the collection of the p-values the 10,000 most associated was selected. Note that, since we followed two different cross-validation strategies the process was repeating for each of those, that is for the across population training set and for the ten partitions training sets. This strategy was applied to DL and BayesC models.

2. PCs single matrix: In the second strategy, we exploited the advantages of principal components analysis (PCA) trying to incorporate it in neural networks. Studies have shown that using PCA in DL framework can be particularly advantageous (Seuret et al. 2017). In our study, principal components (PCs) were computed based on eigenvectors for each of the obtained GRM with dimensions [n x n], where n the number of observations. We run the analysis by merging in a single matrix all the six PCs sets introducing as a single layer to the network (Figure 5.1 (B)) testing whether this strategy will enhance the performance.
3. Multiple inputs: Here, we tested whether multiple inputs strategy could improve the prediction of traits. Other works have shown that a multiple inputs strategy can reduce overfitting and computational cost while at the same time exploits mixed data improving prediction (Livieris et al. 2020). Xiong et al. (2021) showed the outperformance of a multiple inputs strategy over the conventional ones, reporting an overall prediction accuracy 79%. Here, we use the six matrix PCs as six inputs feeding to the network in different layers. Thus, the network accepted six different input layers which independently forwards in six different hidden dense layers. Next the six layers are merged by a concatenate layer (Figure 5.4).

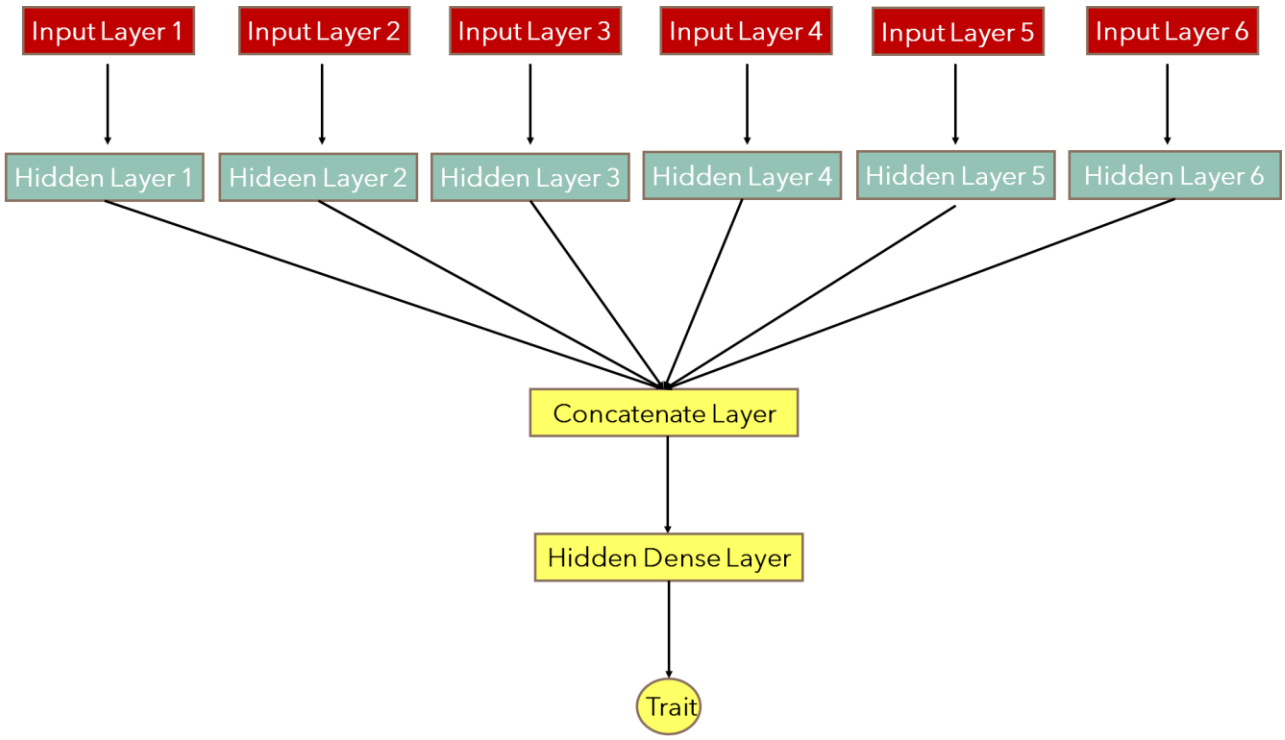

Figure 5.4: Representation of Multiple inputs strategy employed in the present study.

Optimization of Hyperparameters

CNNs, MLP and BayesC were implemented using the 10,000 most associated markers as inputs as well as only SNPs. Additionally, MLP network was employed for using the six PCs as a single input matrix, as six different input layers and for PCs produced by GRM of SNPs. RKHS method was performed under two models, one using as inputs the six GRM from each marker set and the other using the GRM from SNPs. All the models were applied separately to each trait. Table 5.1 shows the different models implemented in our analysis.

Table 5.1: Summary of the analysis.

	MLP, CNN, BayesC		**RKHS**		**MLP**		
	10,000 most associated	**SNPs**	**SIX GRM**	**SNPs GRM**	**PCs single matrix**	**PCs six matrices**	**PCs SNPs**
Culm diameter	x	x	x	x	x	x	x
Leaf senescence	x	x	x	x	x	x	x
Grain weight	x	x	x	x	x	x	x
Time to flowering	x	x	x	x	x	x	x
Apply for 11 scenarios	x	x	x	x	x	x	x

Note that we run the analysis for each of the eleven training sets (10-fold, ARO/ADM). In total we implemented 11 (4 models + different input strategies) x 4 (traits) x 11 (scenarios) runs in our analysis including the Bayesian models. For each of the eleven runs, hyperparameter tuning was performed obtaining the best hyperparameters and then retrained the model with the hyperparameters obtained by the search. Here Keras Tuner (https://www.tensorflow.org/tutorials/keras/keras_tuner) library was used to pick the optimal set. Hyperparameters are the variables that control the training process and the topology of our model. When the model is built for hyperparameter tuning, the search space is also defined in addition to the model architecture. Then a tuner must be selected to determine which hyperparameter combinations should be tested. In our analysis we used the Hyperband tuner. The Hyperband tuning algorithm uses adaptive resource allocation and early stopping to quickly converge on a high-performing model. The algorithm trains a large number of configurations for a few epochs and carries forward only the top-performing half of models to the next round (Li et al. 2018a) evaluating the performance by computing the MSE (for quantitative traits) or the binary cross-entropy (for binary traits) on a held-out validation set. The best model is the one that minimizes the error. After the hyperparameter search was finished, we evaluated the model on the test data and performed prediction computing the pre-mentioned evaluation metrics of interest on the test dataset. Figure 5.5 displays the suggested scheme.

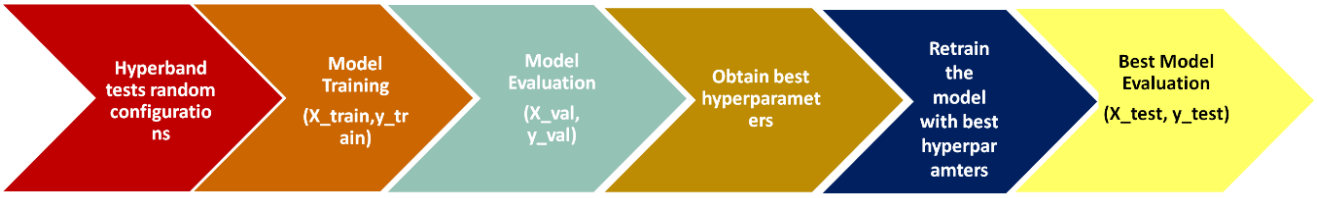

Figure 5.5: Figure depicts the basic scheme performing from hyperband tuner to determine the best configuration towards to the final evaluation of the model.

DL performance is controlled by various parameters and thus the optimization of the hyperparameters is not a trivial step. Here, we designed the tuner search space based on the available literature (Sandhu et al. 2021; Zingaretti et al. 2020). There are two types of hyperparameters, the model hyperparameters and the algorithm parameters. The first ones influence model selection such as the number and width of hidden layers whereas the second ones influence the speed and quality of the learning algorithm as the learning rate for optimizer (e.g., Stochastic Gradient Descent). The hyperparameters chosen to be optimized were: activation function (relu, tanh, linear), number of hidden dense layers (1,2,3,4,5), number of neurons for each hidden layer (10,16,38,50,62,98,112,150), number of filters in CNN (16,32,64,128), optimizers (Adam, RMSprop, SGD), dropout rate (0,0.05,0.1,0.15,0.2,0.25,0.3), L1 and L2 regularizers with optimized weight decay parameter (0.001, 0.01, 0.05,0.1). For the hyperparameter optimization 80% of the training set was used and the remaining 20% used as the validation dataset and applied for inner testing. Training a DL network that can generalize well new dataset is a challenging issue. A model with too little capacity cannot learn from the data, a problem known as underfitting, whereas a model with a large capacity can learn and fit too well to the training dataset results in overfitting. For avoiding and reducing the effects of these two phenomena there are techniques that can be adjusted to a DL network. An approach to reducing generalization error is to use a large model with

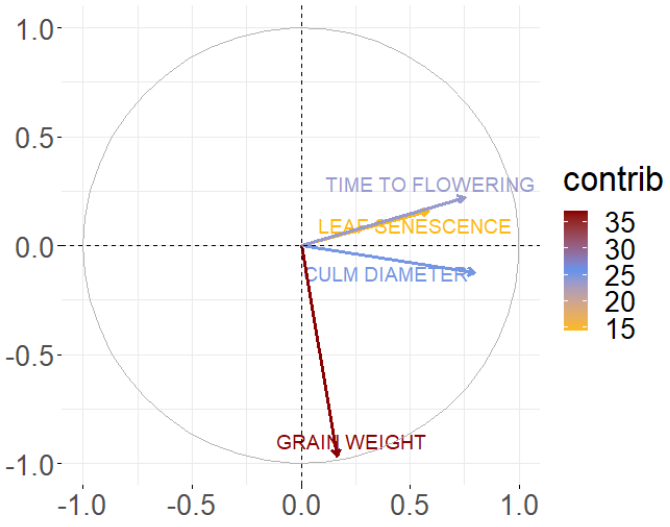

Figure 5.6: PCA loadings of each trait for the two first standardized principal components.

regularization during training that keeps the weights of the model small. These techniques reduce overfitting while at the same time contribute to a faster optimization and overall performance of the model. Here we used two regularization techniques such as L1 and L2 with a weight decay parameter.

These techniques penalize the weight values of the network making values tend to zero, negative values equal to 0 avoiding a parsimonious model. L1 adds "squared absolute value of magnitude" of coefficient as penalty term to the loss function while L2 adds "squared magnitude" of coefficient as penalty term to the loss function. We added L1 and L2 regularizers in the first convolutional layer of CNN model and in the first hidden layer in MLP. Additional to the regularization, dropout and early stopping were applied to reduce the effect of overfitting and underfitting on our models. Dropout is a technique where randomly selected neurons are ignored during the training whereas early stopping is a method that stops the training once the model performance stops improving on the validation set for a number of training epochs. Our analysis was implemented using Tensor Flow 2.8.0 library with Keras 2.8.0 interface and Keras Tuner 1.1.2.

RESULTS

Phenotypic Structure and Genetic Inference

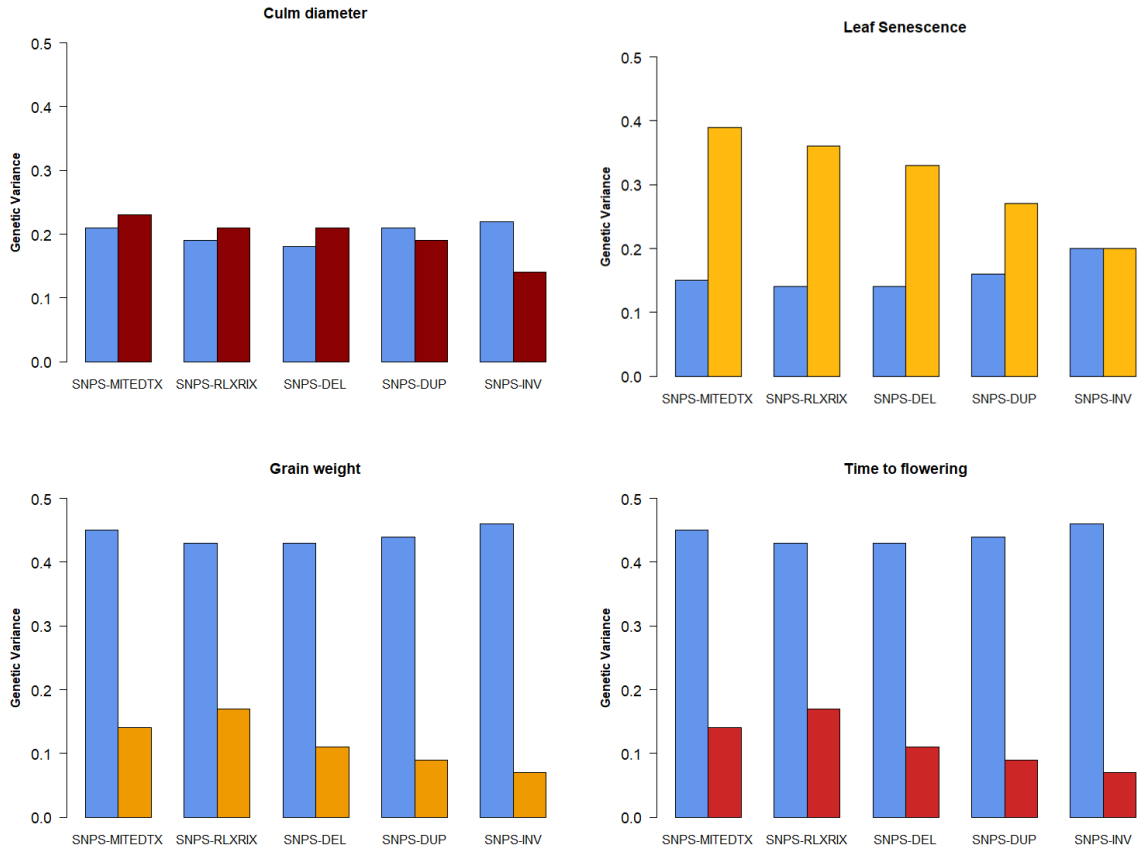

Figure 5.7: Means of posterior distributions of genetic variances explained by each marker set.

PCA gives us the underlying structure of our data and the direction of the maximum variation if we project them in a lower dimension space. Figure 5.6 shows the projections of variables of each trait onto the principal components. The length of the arrow is proportional to trait contribution, whereas the angle between arrows shows whether traits are correlated (pointed out in the same direction) or not. An analysis in two principal components displays that the first component depends on grain weight which contributes the most to the total phenotypic variation. The main contributors to the second component in descending sequence are, the Time to flowering, Culm Diameter and Leaf Senescence.

Genetic variance estimates were obtained for each trait (Figure 5.7). Particularly, we estimated the genetic variance explained by each SVs marker set in comparison to SNPs, in order to understand the relative importance of each set to determine the observed phenotype. As Vourlaki et al. (2022), we chose not to use the term "heritability" since assumes panmixia, a condition not fulfilled here. Figure 5.7 shows that five out of six SVs (MITE-DTX, RLX-RIX, DEL, DUP) can explain a significant fraction of genetic variance, larger than that explained by SNPs in two out of four traits, the Culm diameter and Leaf Senescence.

Comparison of Model performances

The prediction ability of DL implementations is compared to those of Bayesian regression applications using RKHS and BayesC for each trait and under eleven scenarios. Particularly, we assess prediction by following two different validation strategies, prediction using ten randomly selected training sets produced by a 10-fold cross validation strategy and prediction across populations. All the models were applied separately to each of the eleven in total validation scenarios (see Materials and Methods). Figure 5.8 shows the performance of each of the models in terms of an evaluation metric which for binary traits is the binary-cross entropy whereas for quantitative traits is the MSE. The points in each box plot of Figure 5.8 represent the values of the evaluation metric for the 10-fold cross validation strategy whereas the boxplot shows the distribution of the numerical values displaying the data quartiles and average. The value that appears in bold is the median value of each model. The highest prediction ability for culm diameter was obtained using MLP network with multiple PCs inputs strategy. For leaf senescence the optimal prediction ability was reported using MLP with 10,000 most associated markers. Overall, for the binary traits MLP seems to outperform CNN and Bayesian regression models. Note that even a slight improvement in prediction of phenotypes can result in a high genetic gain when accumulated through generations. For the case of quantitative traits, Bayesian Regression models reported higher prediction ability values than those with DL models. Particularly, grain weight was better predicted under RKHS model using SNPs GRM whereas Time to flowering using GRM form 10,000 most associated markers. In addition, the lowest loss values observed in Time to flowering.

In the second cross validation strategy, phenotypes of all ADM and ARO accessions were predicted given the rest of the accessions. Figure 5.9 shows the prediction ability for across population strategy under eleven different models. Here, culm diameter and leaf senescence were better predicted by RKHS using the six GRM as a single input.

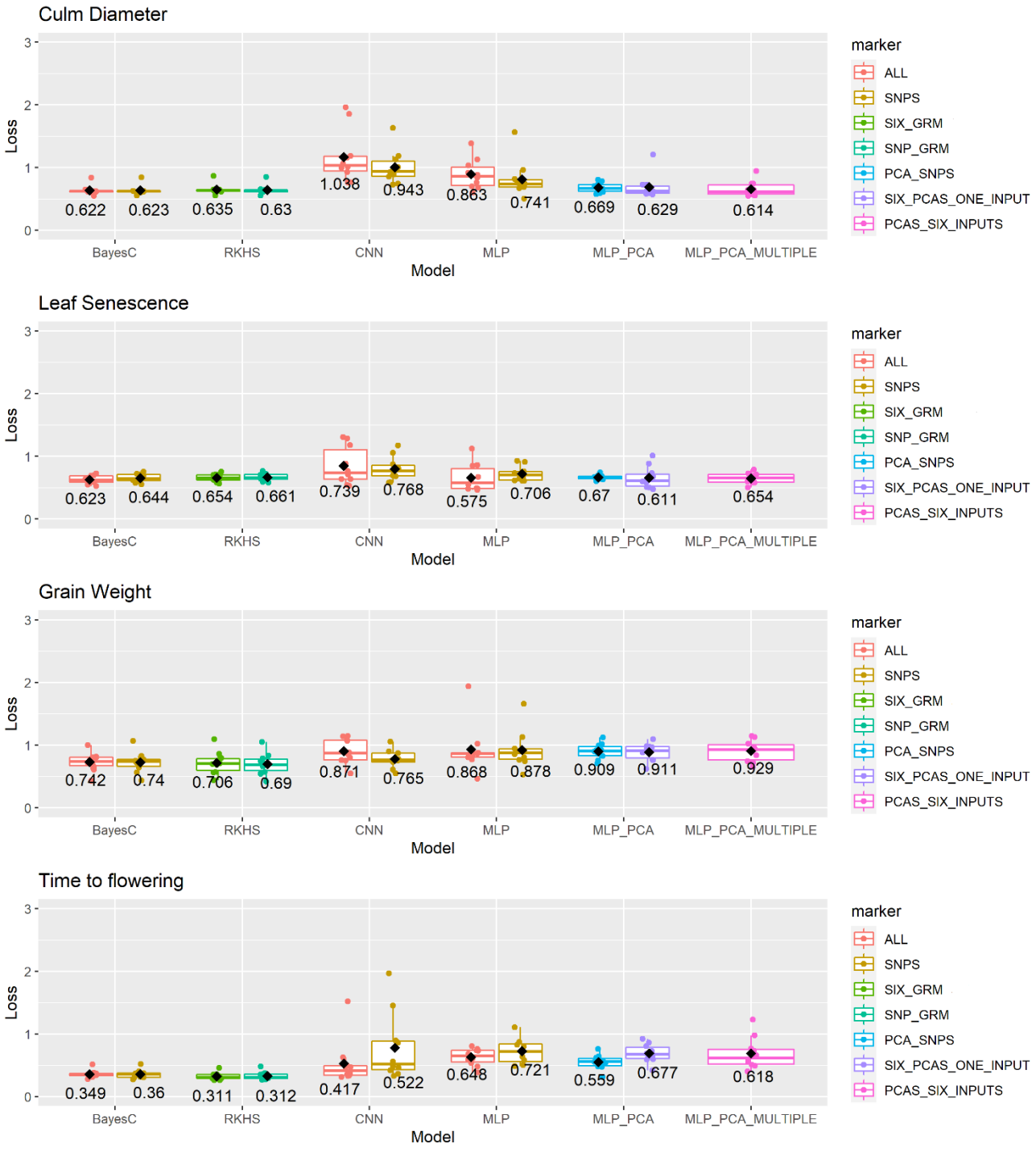

Figure 5.8: Figure shows the performance of each of the eleven tested models under the 10-fold cross validation strategy. Each model was applied separately to each of the ten partitions. Points represent the evaluation metric for each of the ten partitions whereas the displayed numerical value represents the median value. The y-axis shows the loss metric values which for binary traits is the binary-cross entropy and for quantitative traits the MSE. Regarding the legend, where ALL represents the genotype matrix of the 10,000 most associated markers, SNPs is the genotype matrix of 10,000 most associated SNPs, SIX_GRM represents the GRM produced by each of the six marker sets, SNP_GRM, is the GRM produced by SNPs, PCA_SNPS is the PCs produced by only SNPs markers, SIX_PCAs_ONE_INPUT represents the six PCs matrices used as a merged matrix and PCAS_SIX_INPUTS is the six PCs matrices used as six separated layers.

Figure 5.9: Figure shows the performance of each of the eleven tested models under the across population strategy. Points represent the evaluation metric and the corresponding numerical value. The y-axis shows the loss metric values which for binary traits is the binary-cross entropy and for quantitative traits the MSE. Regarding the legend where ALL represents the genotype matrix of the 10,000 most associated markers, SNPs is the genotype matrix of 10,000 most associated SNPs, SIX_GRM represents the GRM produced by each of the six marker sets, SNP_GRM, is the GRM produced by SNPs, PCA_SNPS is the PCs produced by only SNPs markers, SIX_PCAs_ONE_INPUT represents the six PCs matrices used as a merged matrix and PCAS_SIX_INPUTs is the six PCs matrices used as six separated layers.

Table 5.2: Optimized hyperparameters for culm diameter.

Hyperparameter	MLP	CNN	MLP PCs	MLP MULTIPLE
Activation	Tanh	Relu	Relu	Linear
No of hidden layers	3	4	2	3
No of neurons	(16,98,98)	(16,38,112,98)	(98,10)	(150,17,32)
No of filters	-	16	-	-
Optimizer	Adam	Adam	RMSprop	RMSprop
Dropout rate	0.25	0.2	0.1	0.05
Regularization	0.001	0.1	0.001	0.01

Table 5.3: Optimized hyperparameters for leaf senescence.

Hyperparameter	MLP	CNN	MLP PCs	MLP MULTIPLE
Activation	Linear	Relu	Relu	Linear
No of hidden layers	2	2	4	3
No of neurons	(50,62)	(62,62)	(150,62,38,16)	(98,64,64)
No of filters	-	16	-	-
Optimizer	Adam	Adam	RMSprop	RMSprop
Dropout rate	0.2	0	0.1	0.1
Regularization	0.001	0.001	0.01	0.01

Table 5.4: Optimized hyperparameters for grain weight.

Hyperparameter	MLP	CNN	MLP PCs	MLP MULTIPLE
Activation	Tanh	Relu	Tanh	Tanh
No of hidden layers	3	2	3	3
No of neurons	(16,112,150)	(62,50)	(16,112,150)	(10,64,16)
No of filters	-	63	-	-
Optimizer	RMSprop	Adam	RMSprop	RMSprop
Dropout rate	0.1	0	0.1	0.25
Regularization	0.05	0.01	0.05	0.1

Table 5.5: Optimized hyperparameters for time to flowering.

Hyperparameter	MLP	CNN	MLP PCs	MLP MULTIPLE
Activation	Tanh	Tanh	Tanh	Linear
No of hidden layers	4	3	5	3
No of neurons	(112,10,10,62)	(38,10,62,112)	(10,98,112,62,112)	(38,16,16)
No of filters	-	128	-	-
Optimizer	Adam	RMSprop	SGD	RMSprop
Dropout rate	0.05	0.3	0	0.1
Regularization	0.05	0.001	0.05	0.1

It is interesting though that DL models seem to outperform the Bayesian ones in both quantitative traits. More specifically, grain weight is better predicted using MLP with six PCs as single input. The highest prediction ability for time to flowering is reported under a CNN model using all SNPs. In general, time to flowering seems to be better predicted compared to the rest traits since in both cross-validation strategies, the lowest loss values are reported by the best model. On average, prediction across populations was less accurate than in 10-fold scenarios as it was expected because of the distantly related training and test datasets. Note that using all markers instead of only SNPs

improved prediction in six out of eight cases based on results displayed in Figure 5.8 and 5.9. Overall, DL models outperformed Bayesian models four out of eight times. In addition, the improvement in prediction ability of quantitative traits in the across scenario was remarkable using DL models. Specifically, the loss in grain weight was 54% less than in Bayesian models whereas in time to flowering 74% (Figure 5.9). From the four cases where DL models outperformed Bayesian models, MLP was the optimal architecture whereas in the two cases using the PCs as input was the best strategy.

Best Hyperparameters for each trait

The performance of DL models depends on the optimization of hyperparameters with different combinations being essential for particular phenotypic traits as many studies have shown (Cuevas et al. 2019; Montesinos-López et al. 2019; Zingaretti et al. 2020; Sandhu et al. 2021). In this study, we implemented an analysis of 11 (models) x 11 (scenarios) x 4 (traits) runs for each of the four different traits. In each run of the analysis a different set of hyperparameters was optimized but demonstrating all these sets would be extremely complicated. The most frequently selected hyperparameters over four model categories, MLP, CNN, MLP with PCs and MLP with multiple inputs for each of the studied traits are summarized in Tables 5.2-5.5. As Tables 5.2-5.5 depict, the optimal number of layers was three in the 50% of the studied cases, followed by four number of layers in 19% of the cases. In the case of activation function, hyperbolic tangent activation function also knows as Tanh function was the dominant in 44% of the summarized cases. Rectified Linear Unit (Relu) was the second optimal activation function with percentage around 31%. Among the available optimizers, Root Mean Square Propagation (RMSprop), Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam), the most selected was the first one with percentage 56%. Adam was the second most frequently chosen during the hypertuning in 31% of the studied cases. In the case of dropout rate, the most selected value to reduce overfitting in the model was the 0.1 in 37.5 % of the cases. We observed that for regularization parameter it was harder to point out one value since various numbers seem to be selected under different conditions with two be equally frequent, 0.001 and 0.01 in 25% of the cases each. Finally, focusing on different selected parameters between binary and quantitative traits, it is observed that Relu was most selected in binary traits while Tanh in quantitative traits.

Discussion

This study shows that merging structural and nucleotide genome-wide variation for genomic prediction can enhance prediction ability for important agronomic traits in rice. SVs such as MITE-DTX, RLX-RIX, DEL, DUP and INV merged with SNPs improved prediction performance in 75% of the studied cases according to Table 5.6. We found that most of the used SVs sets explain a significant fraction of the phenotypic variation in rice (Figure 5.7).

Studies on plants have shown the association between structural variants and phenotypic traits (Żmieńko et al. 2014). Moreover, SVs are responsible for a diversity of phenotypes across major traits in plants (Sutton et al. 2007; Cook et al. 2012). Late or early flowering on wheat depends on the increase copy number of Vrn-A1 and Ppd-B1 genes respectively (Würschum et al. 2015). In addition, plant height in wheat is associated to a specific tandem duplication (Li et al. 2012). Studies in rice and

tomato have shown that TIPs can reveal association with traits that are not detected with SNPs (Dominguez et al. 2020; Akakpo et al. 2020; Castanera et al. 2021). Transposition activity is not constant over time, and it seems to be strongly associated to stress situation and environmental stimuli (Dubin et al. 2018). The strong association of transposition activity of SVs to stress situations as the adaptation of a crop to a new environment, as in a breeding program, could be an explanation of the high capacity of SVs in the prediction of phenotypic traits. Here we chose to analyze four traits with an economic impact in rice production. Culm diameter, leaf senescence and time to flowering are correlated as Figure 5.6 indicated, whereas grain weight is uncorrelated to them with the highest contribution to the total phenotypic variation yet. Traits such as time to flowering and grain weight are polygenic, controlled by many quantitative trait loci of large effects (Begum et al. 2015; Xu et al. 2015; Chen et al. 2021). Studies in culm diameter have shown that it is controlled by twelve QTLs associated with lodging resistance in dry direct-seeded rice (Yadav et al. 2017). In addition, delayed leaf senescence or stay-green is associated to forty-six QTLs made up the genetic basis of this important trait in rice (Jiang et al. 2004). Genomic prediction of traits such as time to flowering was quite accurate with the loss metrics reported the lowest values across all the study. For leaf senescence the GP ability was lower than that in time to flowering yet accurate.

Table 5.6: Minimum prediction loss and corresponding model with input strategy.

	Scenario	
Traits	10-folds partitions	ARO/ADM accessions
Culm diameter	0.614 (MLP with multiple PCs)	0.633 (RKHS with GRM from all markers)
Leaf senescence	0.575 (MLP with 10,000 most associated markers)	0.652 (RKHS with GRM from all markers)
Grain weight	0.69 (RKHS with GRM from SNPs)	0.637 (MLP with PCs as single input)
Time to flowering	0.311 (RKHS with GRM from all markers)	0.313 (CNN with SNPs)

Increasing the prediction accuracy of traits in plant breeding is challenging but at the same time of highly importance taking into consideration the constant climate change. New methods attempt to improve prediction of agronomic traits promising lower computational cost and better results. DL is a state-of-the-art method applied in many different fields. Many studies have compared DL with standard linear models for genomic prediction (González-Recio et al. 2014; Ma et al. 2017; Bellot et al. 2018; Montesinos-López et al. 2018; Zingaretti et al. 2020). In this study we investigated the performance of DL models for predicting complex traits in rice comparing them to Bayesian regression methods under different input strategies and scenarios. Our results showed that DL can increase prediction accuracy compared to Bayesian methods in about half of the studied cases. Table 5.6 depicts a summary of the study reporting the best models and the lowest values of evaluation error metrics such as MSE and binary-cross entropy. Across DL architectures, MLP was the best one, a result consistent with other studies in plants (Sandhu et al. 2021) but in contrast to previous experience in our group (Bellot et al. 2018; Zingaretti et al. 2020). For the case of Bayesian regression models, RKHS clearly outperformed BayesC.

Another critical and challenging issue in DL models is the optimization of hyperparameters, mainly due to the high computational cost. The tuning of the hyperparameters for each trait depends on the genetic basis and architecture of the trait. As we showed in Tables 5.2,5.3,5.4 and 5.5, different combinations of hyperparameters were selected for the various traits as the prediction ability is highly associated with the interaction of these factors (Bellot et al. 2018; Montesinos-Lopez et al. 2018). We observed that Tanh activation function was the most useful across the analysis. However, for binary traits Relu was best. Different layers in MLP and CNN were selected as well as various number of neurons and filters respectively were chosed to analyze the complex biological connections. DL models can capture interactions of large orders because of the presence of hidden layers (Goodfellow et al. 2016; Lecun et al. 2015). However, RKHS models are also able to capture complex interaction patterns. This ability of both methods can be reflected in our results demonstrating that both are equivalent and can capture complex interactions. Incorporating PCs in the MLP models proved beneficial since in 50% of the studies with DL as the best model, it was the optimal strategy. Using multiple input layers as input strategy was the optimal strategy for culm diameter under the 10-fold scenarios.

It is commonly believed that DL requires a large dataset to be used in order the training of the model to be effective (Min et al. 2017; Alipanahi et al. 2015; Xiong et al. 2015). However, our current results and some of related works (Ma et al. 2017; Sandhu et al. 2021; Zingaretti et al. 2020) support that DL models can be effective even with a smaller dataset for training. Bellot et al. 2018 found that using 100k individuals for prediction did not result in a consistent advantage in DL models. Thus, the training population size can be less important compared to the studied trait (Sandhu et al. 2021). To avoid overfitting that is the biggest issue in a small dataset, regularization and dropout techniques were applied. It is worth mentioning that even though Table 5.6 indicates equivalent prediction values between the two validation scenarios followed here, in across scenario the prediction ability of quantitative traits was improved by DL for 54% in grain weight and 74% in time to flowering. These results might suggest that the association of genetic basis of the studied trait to the accessions used for training can be critical in GP. The fact that training and test dataset were distantly related in ADM/ARO scenario, makes the results even more interesting.

Finally, we would like to mention the challenges and limitations of DL models. Firstly, DL models do not provide clear insights into the genetic architecture of the traits, nor do they give information about the effects of specific markers in the studied traits. Different hyperparameters act on different parts of the data, making it hard to interpret the biological significance and importance of each marker in the model (Bellot et al. 2018; Cuevas et al. 2019). Also, the high computational cost of training models is a significant drawback, especially when multiple hyperparameters must be optimized for each trait separately (Gulli and Pal et al. 2017). It is clear that the outperformance of DL over linear models is not always the case. The prediction ability depends on the studied traits and can be influenced by many factors. There is not a single algorithm that perform better in all species and traits (Perez-Enciso and Zingaretti 2019) since its performance depends on various factors. Therefore, even though the advantage of DL networks against linear methods has not been established yet, their incorporation into plant breeding can be important to improve genetic merit for complex traits.

Chapter 6

Discussion

Many complex traits of interest are highly heritable and yet genetically complex, meaning that their variation arises from differences at numerous loci in the genome. Genetic architecture describes the characteristics of genetic variation that are responsible for heritable phenotypic variability. It refers to the number of genetic variants affecting a trait (or fitness), their frequencies in the population, the magnitude of their effects and their interactions with each other and the environment (Timpson et al. 2018).

This thesis tries to understand and explore the consequences of polygenic variability through two different approaches, the population genetics using inference, and the quantitative genetics using prediction. Two different frameworks, that of domestication and that of plant breeding are used to address the objectives posed by this thesis. While selective breeding is the intentional selection by humans to change the gene pool of population, usually applying truncation selection, domestication is the consequence of all selection pressures coming from selective breeding and from natural selection in a given environment, which modifies the gene pool of a population for life (Kincaid 1993). In both frameworks, humans are interested in reproducing desired traits bearded by the selected or domesticated individuals. However, the way the phenotypic traits vary across a population is still to be determined since their underlying genetic architecture is not easily defined and controlled. The vast majority of the complex traits is controlled by many loci with small effects that experience only subtle changes in their frequency. Because of this complexity, to identify signals of polygenic variability is not a trivial procedure. Nevertheless, studying the way that variability patterns changed under different selective and demographic effects can give us a new insight into the genome of species. At the same time, current studies in GWAS allow us to detect association between phenotypic traits and causative variants that are not only restricted to nucleotide but even structural variation. Accordingly, this thesis was divided into two parts. The first part focuses on the study by simulation of the patterns of variability affected by polygenic adaptation and methodologies for detecting signals of genome variability controlling the polygenic adaptive traits under a domestication process. The second part aims to examine whether the use of different sources of genomic variability can improve the prediction of complex traits under a rice breeding process and evaluating different methodologies.

Many studies in population genetics (e.g., Kim and Stephan 2002; Nielsen 2005) have tried to find signatures of positive directional selection in the genomes of natural (sexually recombining) populations. The final aim has been to find the loci favored by selection and define their associated functions and phenotypes. The predictions of the Neutral Theory (Kimura 1968) have been widely used as a null model in population genetics, which assumes a few loci at which positive selection acts with some occasional extensions to multiloci models. Gillespie (1994) contrasted the Neutral theory with other where the action of natural selection predominates but it is affected by fluctuating environmental changes in the direction of selection, arriving to similar predictions. However, recently, thanks to the effort and advances at genome level sequencing in the last two decades,

polygenic selection has been studied using quantitative genetic theory, which is formulated in terms of allele frequencies. In the quantitative genetic models of adaptation, selection acts on one or more phenotypic traits such that a genotype-phenotype map is assumed to bridge the gap to population genetics theory (de Vladar and Bardon 2014; Stephan 2016). Pritchard et al. 2010 argued that adaptation in natural populations occurs not by sweeps alone, but by subtle allele frequency shifts as well, in many loci controlling polygenic traits. That raises questions around the genomic basis of environmental changes such as domestication.

In Chapter 3 of this thesis, we simulated domestication process under different selective and demographic conditions trying to infer their effect on the genome. We investigated the patterns of variation and the capacity to detect the effect of positive selection when domestication is driven by different scenarios having many loci of medium to weak effects following a polygenic adaptation model. Even though many studies have explored the genomic basis of domestication and the genetic cost of the process (e.g., Ross-Ibarra et al. 2007; Flood and Hancock 2017; Moyers et al. 2018; Flori et al. 2019; Frantz et al. 2020; Leno-Colorado et al. 2020), the genetic architecture that controls the adaptive traits is yet to be revealed. In our study we didn't simulate domestication as a single polygenic trait but as an environmental process that modifies the selective effects of many different loci aiming to study not how genes influence phenotype, but rather to know what evolutionary forces maintain genetic variability**.** Specifically, we assumed two populations diverged after a hypothetical environmental change. The environmental change is simulated assuming a number of deleterious mutations including effectively neutral ones in the wild population to become beneficial in the domesticated population. Conversely, mutations with beneficial effect in the wild population become deleterious or effectively neural in the domesticated population. The impact of migration is also investigated from the wild to the domesticated populations. Firstly, we conducted a comparative study of the patterns of genetic diversity between the two populations.

The age of adaptive mutations depends on demographic events

Table 6.1: Description of the ten simulated scenarios.

Scenarios	Bottleneck Duration	Migration	Strength of Positive Selection (S_b)	Domestication	
				% Change	%Positive
1	Short	Yes	10	0	2.5
2	Short	Yes	1	5	25.0
3	Short	Yes	1	25	10.0
4	Short	No	1	0	2.5
5	Short	No	1	25	25.0
6	Short	No	10	25	2.5
7	Long	No	10	0	2.5
8	Long	No	1	5	2.5
9	Long	No	1	25	25.0
10	Long	No	10	5	10.0

In Chapter 3, we simulated ten different scenarios (Table 6.1). We quantified the fraction of adaptive substitutions (α) across scenarios and populations, observing that for wild population the

highest values of α are seen in scenarios with strong selection (scenarios 1, 6, 7 and 10, Figure 6.1) whereas for domesticated population, the largest α values are observed in scenarios 1,3, 5 and 9 (Figure 6.1). These scenarios (except scenario 1) assumed a high fraction of deleterious mutations in wild population that become beneficial in domesticated population (Table 6.1). A large fraction of adaptive amino acid substitutions (25-60%) in these scenarios can be explained by initially deleterious polymorphisms that become beneficial (m_7 sites) in the process of domestication (Figure 6.2, Table 3.4, Supplementary Table A.3).

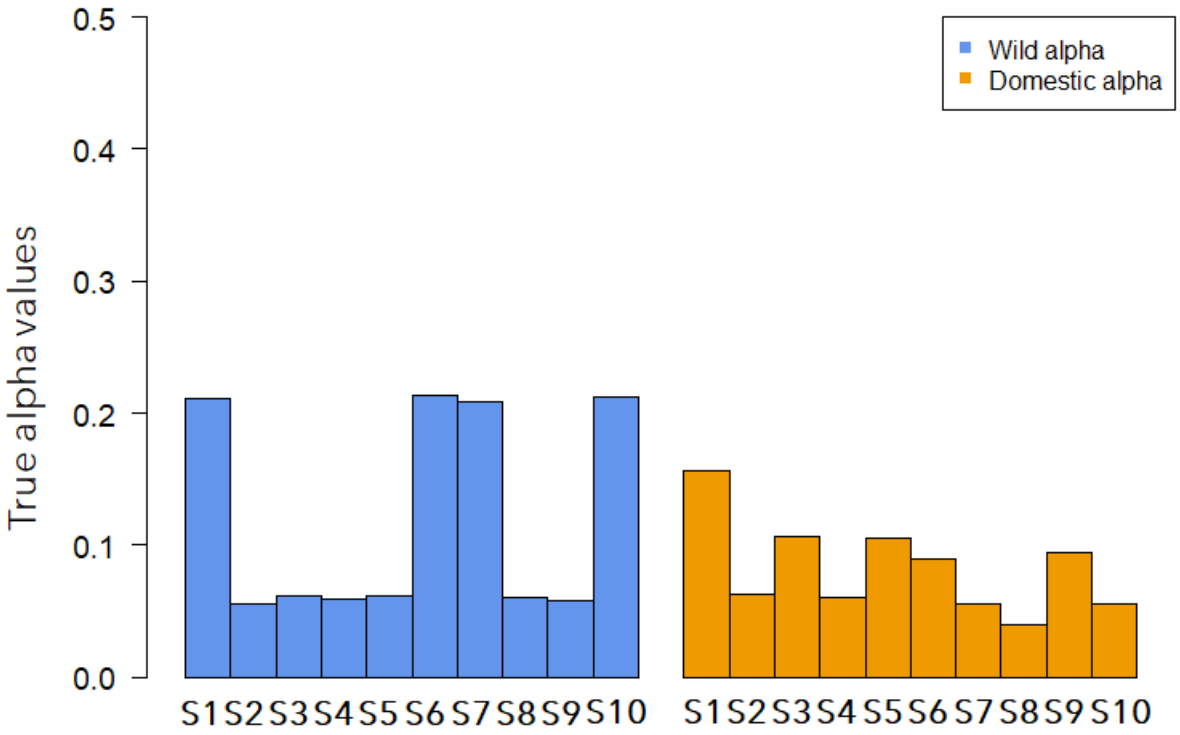

Figure 6.1: Figure shows the true alpha for total mutations in relation to all nonsynonymous fixations per scenario and for each population. S number refers to scenarios in Table 6.1.

In all simulated scenarios the largest proportion of adaptive substitutions came from polymorphisms that have been beneficial before domestication started (Figure 6.2, Supplementary Table A.3). We observed that there is an association between the bottleneck duration and the number of beneficial variants that are exclusive in domesticated population. Under a short bottleneck, new beneficial variants segregating in domesticated population don't have the time to reach fixation whereas the opposite is noticed when a long bottleneck is simulated. In fact, for longer bottlenecks, "de novo" adaptive amino acid substitutions reach fixation in higher proportion than other scenarios, but at the same time a high proportion of fixed deleterious mutations increase as well (Figure 6.2. Supplementary Table A.3).

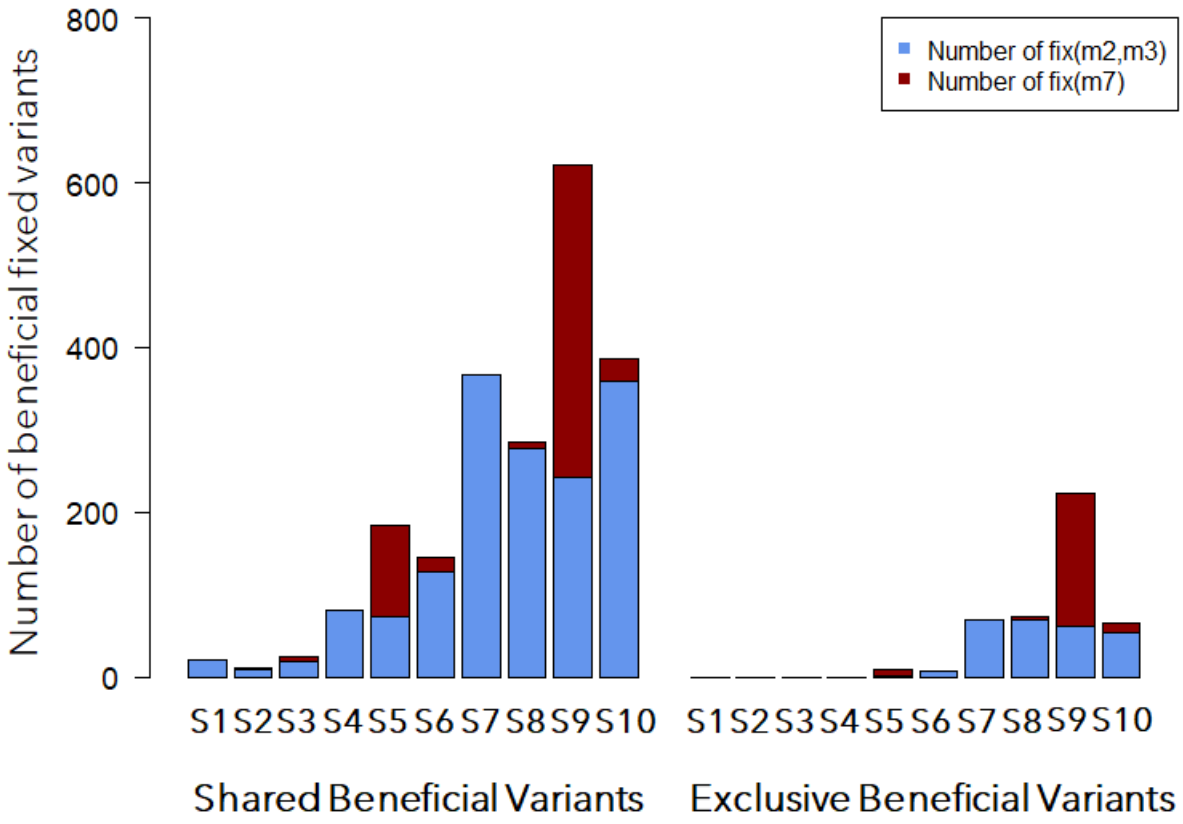

Figure 6.2: Figure shows the absolute number of shared and exclusive beneficial variants per each scenario for the domesticated population. Red color indicates variants that change their effect from deleterious to beneficial after domestication split (m_7). Blue color shows the variants that change their effect after domestication split but remain beneficial (m_2, m_3). S number refers to scenarios in Table 6.1.

Distortion of the site frequency spectrum (SFS) caused by demography and linked selection

Comparing the 1D-SFS of synonymous and nonsynonymous variants against the expected 1D-SFS for only neutral mutations under the same demographic model revealed the action of linked selection in the domesticated populations (Figure 3.4). During a long bottleneck (scenarios 7, 8, 9 and 10, Table 6.1), haplotypes with deleterious variants can increase in frequency under the effect of a small effective population size as well. These deleterious variants will drag with them neutral variants which are in linkage resulting in an excess at low frequency variants and a lack of variation in the rest frequency classes as observed in Figure 3.4. Linked selection seems to be taking place also in short bottleneck scenarios and particularly in scenarios where a high fraction of deleterious segregating mutations become beneficial (Figure 3.4, scenarios 2, 3, 5, 6, reflected as a subtle increase in the number of synonymous variants at higher frequencies). It is interesting though to study more in depth if linked beneficial selection is affecting the 1D-SFS in short bottleneck scenarios. We found that in scenarios 3, 5 and 6 (Figure 3.4) the excess of synonymous variants at high frequency is larger than the excess of nonsynonymous variants. We assume that the shift in the selection coefficients in the domesticated population is setting the ideal conditions for the emergence of soft sweeps (Hermisson

and Pennings 2005; Stetter et al. 2018). Therefore, soft sweeps might be the reason that caused this distortion of the synonymous 1D-SFS in short bottleneck scenarios.

Quantifying the signal of domestication

Two algorithms were employed in this study for inferring the distribution of fitness effects (DFE) parameters under domestication: polyDFE and *dadi*. Using polyDFE, the DFE of each population is compared individually, whereas with *dadi* a new joint DFE model is inferred. We found that polyDFE provides more precise estimates of the deleterious section of the distribution than the new joint DFE algorithm from *dadi* (Figure 3.5). On the other hand, the new joint DFE model can distinguish the strength of positive selection between weak and strong even though tends to overestimate the strength for scenarios with a high proportion of positive change (Table 3.7). However, the new *dadi* algorithm is able to estimate the fraction of mutations that changed their selective effect in the domesticated population (p_c, see Figure 3.9). Nevertheless, we were not able to infer accurately the deleterious mutations that have become beneficial (p_{c+}) since our estimates were not significantly different from zero.

The process of adaptation is of fundamental importance in evolutionary biology. Advancing our understanding of the genomic basis of domestication is essential not only to understand this phenomenon but overall to understand how evolutionary forces act and shape the genetic diversity of populations. The knowledge of the full DFE can give us an insight into how these evolutionary forces interact and change the effect of mutations on fitness under an environmental change. Finally, despite the limitations encountered in this study for detecting a polygenic signal of domestication, we were able to quantify it through the analysis of a joint DFE.

Genome-wide variation and association with traits

Modern genotyping technologies has made it possible to identify quantitative trait loci (QTL), the regions of a chromosome or individual sequence variants that are responsible for trait variation (Barton and Keightley, 2002). This has accelerated the delivery of new crop varieties with improved yield and quality. Genomic prediction (GP) methods can increase breeding speed as they are particularly effective to predict complex traits affected by many genes, overcoming the Marker-assisted selection (MAS) limitations. Most studies in plant breeding assume that the genetic variability at the whole-genome level is due to SNPs. In Chapter 4 we incorporated TIPs as genetic markers in GP models to study whether they can improve prediction of traits in rice compared to using only SNPs. However, TIPs are not the only structural variation in the genome. Structural variation such as deletions, inversions, and duplications are prevalent in the genome and they play an important role in plant evolution. In Chapter 5, we investigated if merging the structural and nucleotide genome-wide variation can improve prediction ability of traits. It is important to note that genetic differences caused by SVs and TIPs can lead to phenotypic variations in a species.

SVs explain a significant fraction of total genetic variation

The prediction ability of GP methods depends on the proportion of phenotypic variance explained by the markers. In Chapters 4 and 5 we showed that SVs can explain a high fraction of

genetic variance, even larger than that explained by SNPs (Table 4.1, Figure 5.7). Overall, SVs contributed more to the phenotypic variation in traits such as culm diameter, culm strength, leaf senescence, salt injury and panicle threshability. Figure 6.3 shows a summary of the results (Table 4.1, Figure 5.7) found in Chapters 4 and 5, indicating which type or variation explains the largest fraction of genetic variance in each phenotypic trait.

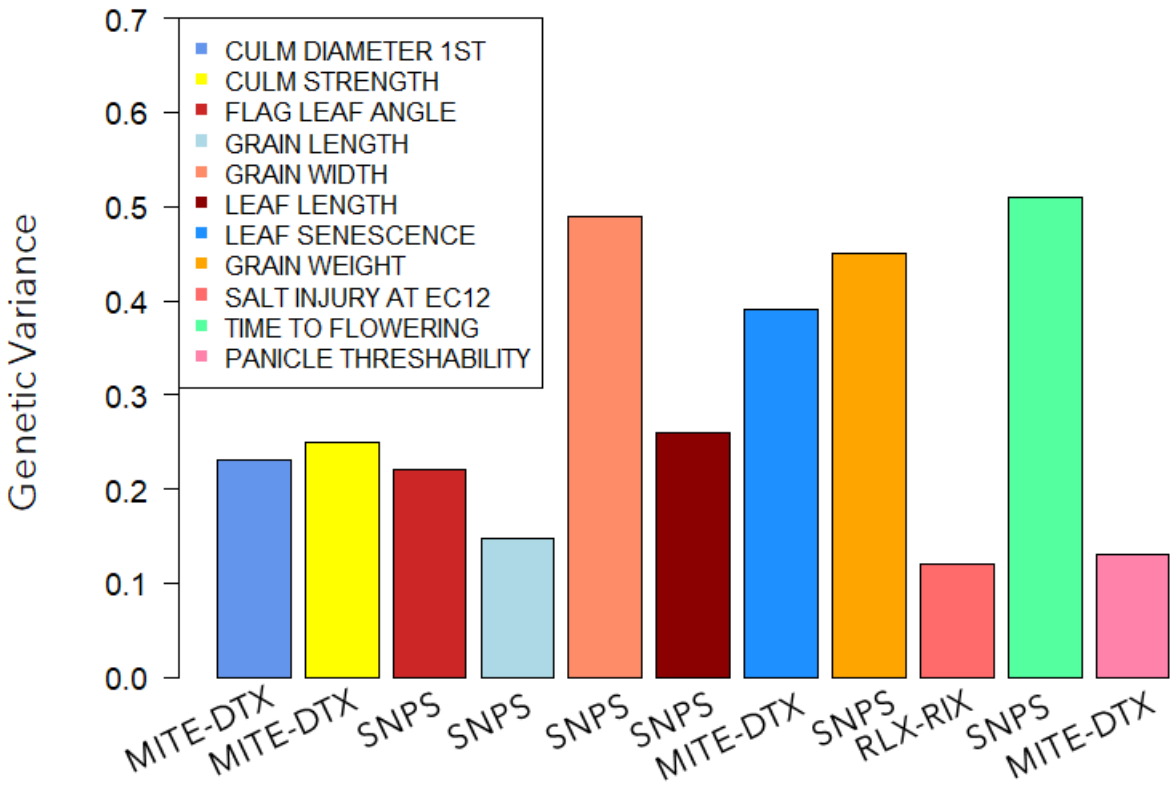

Figure 6.3: Bar plot displays the type of variation explains the highest fraction of genetic variance for each phenotypic trait. Plot is summarized by results demonstrated in Table 4.1 and Figure 5.7.

SVs improve prediction of phenotypic traits

Our results suggested that SVs improve prediction ability for important agronomic traits in rice (Figure 4.2, 4.3, 5.8, 5.9). Table 6.2 indicates which type of variation resulted in the best predictions. Summarized results were extracted by Table 4.3 and Table 5.6 evaluating only the highest value per each trait for each Table, independently of the cross-validation strategy. Based on the results it is interesting to examine the reasons behind the high capacity of TIPs and SVs to predict phenotypes better than SNPs.

Studies on plants have shown that SVs are responsible for phenotypic variation in many important traits in plants whereas they demonstrate associations with traits that are not detected with SNPs (Dominguez et al. 2010; Akakpo et al. 2020; Castanera et al. 2021; Żmieńko et al. 2014; Sutton et al. 2007; Cook et al. 2012; Fuentes et al. 2019). TEs have been proposed as an agent of rapid adaptation because they can produce abundant genetic variation in a limited time whereas they affect phenotypic variation (Stapley et al. 2015; Schrader et al. 2014; Casacuberta et al. 2013; Chuong

et al. 2017; Barrón et al. 2014). Nevertheless, the extent to which TEs contribute to rapid phenotypic variation and the mechanisms by which they influence phenotypes are unknown (Niu et al. 2019). Fuentes et al. (2019) showed that rice genome regions with frequent SVs were enriched in stress response genes. In contrast to SNP mutation rate, transposition activity is not constant over time with bursts of transposition associated with stress situation or environmental stimuli. Hence, the adaptation of a crop to a new environment, as in breeding could be a period particularly prone to transposition activity (Dubin et al. 2018; Baduel and Quadrana 2021).

Table 6.2: Type of variation results in the highest prediction ability for each phenotypic trait.

Trait	Type of Variation
Culm Diameter	MITE-DTX, ALL
Culm strength	SNPs
Flag leaf angle	SNPs
Grain length	MITE-DTX
Grain width	SNPs, ALL
Leaf length	RLX-RIX
Leaf senescence	MITE-DTX, RLX-RIX, ALL
Grain weight	MITE-DTX, ALL
Salt injury	SNPs, MITE-DTX
Time to flowering	ALL
Panicle threshability	SNPs

*ALL: Structural and nucleotide variation applied together.

TE insertions are highly deleterious with the strongly ones rapidly removed from the population. Insertions that have little or no effects on genome function and host fitness may reach fixation according to the efficiency of selection and drift at purging these insertions from the population, which vary among species (Lynch 2007). Also, TE insertions mostly segregate in small frequencies whereas the fitness effect varies across the different superfamilies. Natural selection and genetic drift seem to shape the distribution and accumulation of TEs (Lynch 2007). The action of evolutionary forces can explain why some elements are maintained in certain genomic location than others (Campos et al. 2016). Moreover, SVs can demonstrate stronger effects than SNPs as some transposon types tend to localize near genes. Unlike SNPs that are randomly distributed along the genome, most TE families show strong insertion preferences towards genes. However, because TE typically have major functional impacts, they are more rapidly purged by natural selection than SNPs (Baduel and Quadrana 2021). Particularly, TEs families associated to environmental cues, often integrate within genes involved in the environmental response generating large-effect mutations, some of which are potentially adaptive (Baduel and Quadrana 2021).

Therefore, SVs could be causative mutations linked to a specific phenotypic trait. Indeed, SVs seem to play a major role in plant genome evolution both in wild and under a breeding framework with examples of SVs as causative mutations have been reported for various agricultural traits (Lisch 2013; Dubin et al 2018). Given the fact that climate change has already shown its consequences, understanding how organisms adapt to new environments is of major importance. Since standing genetic variation is generally considered as the main source of rapid adaptation to environmental changes, the majority of genomic studies mostly focused on SNPs. However, SVs seems to contribute with rare and typically large-effect alleles supplying species with adaptive de novo variants in response to the environment (Capblancq et al. 2020; Baduel and Quadrana 2021). Understanding this major role of SVs as a source of genetic variation and incorporating them in plant breeding programs could provide us with a high advantage on predicting new phenotypes adapted to drastic environmental changes. Hence, we will be able to sustain and improve the crop yield for global population in response to new environmental conditions.

Polygenic traits

All the traits studied were of particular importance for a rice breeding program. Among them, time to flowering is a highly critical trait since an early or late growing can affect the yield (Wang and Li 2019). Morphological traits such as grain weight, grain length and grain width as well as stay-green traits as leaf senescence, are of great importance for crop production. It was interesting to observe that using GP either with Bayesian methods or DL networks, the highest and more accurate prediction ability values were reported for time to flowering, grain length, grain width and leaf senescence (Table 4.3 & 5.6). Some of those traits are uncorrelated as time to flowering with grain width (Figure 4.1). However, all these pre-mentioned traits are polygenic. Traits such as time to flowering, grain length and grain width are controlled by large effect quantitative trait loci (Begum et al. 2015; Xu et al. 2015; Chen et al. 2021). Numerous QTL have been detected in leaf senescence (Jiang et al. 2004).

We found that GP was quite accurate for these traits, with SVs often outperforming SNPs. As we have already mentioned, TE insertions have a major impact on phenotypes associated with rapid adaptation. Thus, a question arises in our study is whether there is an association between specific studied traits in rice and SVs. The role of SVs in producing major effect alleles and causative variants associated with phenotypic traits, could be a potential reason of the precise prediction of these traits in our study. Since plant breeding is a cumulative progress acts over generation, even a small improvement on the prediction of a trait will be of high importance in medium to long term.

From domestication to breeding

Andersson and Purugganan (2022) reviewed the progress that has been made in the last years for revealing the underlying genetic variation controls for the phenotypic diversity in crops and domesticated animals. Their study suggested that no obvious domesticated genes are involved in animal domestication. Particularly, a polygenic background is observed. The domestication of plant and animal species causes genetic differentiation between the domesticated populations and their wild ancestors, but also increases the phenotypic variation within species as new traits appeared to be selected. The selection pressure acts adapting species associated to a given environment,

increasing the fitness of the species, and consequently modifying the phenotypic diversity by selecting adaptive traits. Therefore, selection by domestication can be classified in different types. Usually, when variation in the selected traits of the domesticated population decreases, directional selection is on action. If the variation in selected traits is increased, then diversifying selection that can lead to speciation is driving the process (Meyer and Puruggannan 2013). However, domestication process can be defined as a coevolutionary mutualistic relationship between humans and plants or animals (Zeder et al. 2006). In some domestication pathways the process was unconscious at first followed by human acting in an intentional manner and apply conscious (artificial) selection to the plants or animals they interacted with. Assuming that conscious directional selection has occurred later in the domestication process, it could result in genes underlying traits that show a reduction in diversity due to the fixation of a favored allele (Ross-Ibarra et al. 2007).

Strong selection can lead to the fast fixation of advantageous variants in domesticated population as studies in plants have shown (Andersson and Purugganan 2022). In contrast to animal domestication, various genes important to the domestication process have been detected in plants. It is assumed that animal domestication occurred by a gradual stage at many loci affecting tameness. That could explain why no specific domestication genes have been revealed. On the other hand, in plants, few genes with major effect have been identified. Among the numerous plant loci that have been identified and linked to specific crop domesticated phenotypes, structural variation seems to underlie a high fraction of variation. More specifically, in a study of 60 known crop genes validated or putative as causative mutations, 41% were SNPs, 38% were insertions/deletions, 15% are TE insertions and 5% duplications or chromosomal rearrangements (Meyer and Puruggannan 2013). In animals, various large-scale deletions, inversions, and translocations have been linked to alleles with major phenotypic effects (Andersson and Puragganan 2022). The role of TE insertions in the evolution of domesticated species is of great importance. Studies have reported how TE insertions have been affected the phenotypic variation. For example, allelic variation has been derived as result of coding region transposon insertions in wrinkled R phenotype in Mendel´s peas and in color polymorphism of date palm (Bhattacharyya et al. 1990; Hazzouri et al. 2015). In general, the activity of TEs in plant evolution has been demonstrated more broadly than in animal genomes. However, TIPs associated with phenotypic traits of domesticated animal have been reported as in case of henny feathering trait in chicken (Matsumine et al. 1991; Li et al. 2019) and in short-legged phenotype in several dog breeds (Parker et al. 2009). As we showed in Chapters 4 and 5 of this thesis, SVs and TIPs can explain a high fraction of phenotypic variation in rice agronomic traits (Table 4.1, Figure 5.7). At the same time, the remarkable performance of the prediction of particular phenotypic traits and the outperformance over SNPs in many cases (Tables 4.3 & 5.6), makes the incorporation of them in a breeding program imperative. The incorporation of SVs and TIPs in breeding programs could be a key factor for achieving the high production challenges. To further benefit breeding, SVs and TIPs genotyping automatization is needed along with extensive databases of crop genome sequences. To sum up, it has been shown that polygenic adaptation, strong selection, deleterious mutations, large effect loci alleles and structural variation are some of the reasons genetic architecture underlies phenotypic variation varies so much across forms of life.

The study of the genetic basis of diversity in domesticated species is inextricably linked to the improvement achieved during the last century in food and nutritional security. In this direction,

quantitative genetics theory has been the main tool for developing new and effective breeding strategies in crops and animals. Sustainable food production along with the development of disease resistant and climate resilient crops are some of the requirements for food security. However, the constantly increasing word population and the climate change, have made extremely challenging to meet the high demands in rice yield. It has been projected that production has to be increased by 60% by 2050 amid climate change consequences (Budhlakoti et al. 2022), such as increased heat, drought, and insect outbreaks. The conventional breeding strategies of hybridization and phenotypic selection don't satisfy the demanded genetic gain. The improvement in the genetic gain based on the Lush equation (Lush 1943) can be secured through i) better phenotypic selection via high-throughput phenotypic and ii) exploiting a broad genetic information representing a diverse eco-geography in breeding program. The progress in genomics approaches leads to the availability of huge resources like genome sequence information. This information has been extensively used for the identification of loci associated with complex and important traits. Genomic prediction has emerged as a significant tool which can exploit genetic information for modeling the crop yield, accelerating the breeding progress under different environmental conditions. Various GP models such as Bayesian models and Deep Learning networks can be implemented to achieve a better prediction ability in breeding programs. However, our results (Figures 5.8 & 5.9) suggested that there is not a single method that performs better in all species and traits. Finally, genetics studies of crops and livestock species are important to highlight the genetic architecture of phenotypic variation, and to help us understand better how species are adapted and how to secure global food production.

Chapter 7

Conclusions

1. Weak beneficial polygenic effects are usually difficult to detect. Nevertheless, these can be detectable if they are affecting a large proportion of the genome, say >10%.

2. Some aspects of the full distribution of fitness effects (DFE), such as the shape and strength of the deleterious DFE, are accurately estimated in both wild and domestic populations when using only the 1D-SFS. However, using the new joint DFE model contrasting the 2D-SFS we were able to quantify a signal of domestication, expressed as the fraction of mutations with divergent selective effects.

3. In highly polygenic models of domestication, the main source of adaptive mutations (i.e., shared ancestral versus new exclusive variants) is highly dependent on the demographic patterns, such as the strength of the bottleneck.

4. In scenarios of polygenic domestication, the SFS of neutral sites deviated from the expected neutral pattern, which points to the action of linked selection as the process to explain the deviation from neutrality in the SFS of neutral sites.

5. Structural variation explains a sizable fraction of the total genetic variation in agronomic traits in rice.

6. Structural variation can improve significantly genomic prediction of complex polygenic traits.

7. There is not a single genomic prediction method that performs better in all species and traits. However, Deep learning methods could be beneficial for plant breeding programs.

Bibliography

3KRGP, (2014) The 3,000 rice genomes project. Gigascience 3:7. doi: 10.1186/2047-217X-3-7.

Akakpo R., Carpentier M.C., Ie Hsing Y., Panaud O. (2020) The impact of transposable elements on the structure, evolution, and function of the rice genome. New Phytologist 226:44–49. https://doi.org/10.1111/nph.16356.

Alipanahi B., Delong A., Weirauch M. T., Frey B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. 33, 831–838. doi: 10.1038/nbt.3300.

Alkan C., Coe B.P. and Eichler E.E. (2011) Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376. doi: 10.1038/nrg2958.

Amadeu R.R., Cellon C., Olmstead J.W., et al. (2016) AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. The Plant Genome 9: https://doi.org/10.3835/plantgenome2016.01.0009.

Amills M., Megens H. J., Manunza A., Ramos-Onsins S. E., and Groenen M. A. M. (2017) A genomic perspective on wild boar demography and evolution, pp. 376–387 in Ecology, Conservation and Management of Wild Pigs and Peccaries, Cambridge University Press.

Andersson L., (2012) Genetics of animal domestication, pp. 260–274 in Biodiversity in Agriculture: Domestication, Evolution, and Sustainability, Cambridge University Press. http://dx.doi.org/10.1017/cbo9781139019514.014.

Andersson L., Purugganan M. (2022) Molecular genetic variation of animals and plants under domestication. Proc Natl Acad Sci U S A. 26;119(30): e2122150119. doi: 10.1073/pnas.2122150119.

Andersson, L. (2001). Genetic dissection of phenotypic diversity in farm animals. Nat. Rev. Genet. 2: 130-138.

Andolfatto P. (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Res. 17(12):1755-62. doi: 10.1101/gr.6691007.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 408(6814):796-815. doi: 10.1038/35048692.

Arunkumar R., Ness RW, Wright SI, Barrett SC. (2015) The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. Genetics. 199(3):817-29. doi: 10.1534/genetics.114.172809.

Avni R., M. Nave, O. Barad, K. Baruch, S. O. Twardziok, et al., (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357: 93–97. https://doi.org/10.1126/science.aan0032

Baduel P., Quadrana L. (2021) Jumpstarting evolution: How transposition can facilitate adaptation to rapid environmental changes. Curr Opin Plant Biol. 61:102043. doi: 10.1016/j.pbi.2021.102043.

Barghi N., R. Tobler, V. Nolte, A. M. Jakšić, F. Mallard, et al., (2019) Genetic redundancy fuels polygenic adaptation in Drosophila. PLoS Biology 17. https://doi.org/10.1371/journal.pbio.3000128.

Barghi N., Schlötterer C. (2020) Distinct Patterns of Selective Sweep and Polygenic Adaptation in Evolve and Resequence Studies. Genome Biol Evol. 12(6):890-904. doi: 10.1093/gbe/evaa073.

Barrón M.G., Fiston-Lavier A.S., Petrov D.A., González J. (2014) Population genomics of transposable elements in Drosophila. Annu Rev Genet. 48:561-81. doi: 10.1146/annurev-genet-120213-092359.

Barton H. J., and Zeng K., (2018) New methods for inferring the distribution of fitness effects for INDELs and SNPs. Molecular Biology and Evolution 35: 1536–1546. https://doi.org/10.1093/molbev/msy054

Barton N.H., Etheridge A.M., Veber A. (2016) The infinitesimal model. bioRxiv 039768.

Barton N.H., Keightley P.D. (2002) Understanding quantitative genetic variation. Nat Rev Genet. 3(1):11-21. doi: 10.1038/nrg700.

Barton, N.H. (1995) Linkage and the limits to natural selection. Genetics, 140(2):821–841, ISSN 0016-6731, 1943-2631. doi: 10.1093/genetics/140.2.821.

Bataillon T., Bailey SF. (2014) Effects of new mutations on fitness: insights from models and data. Ann N Y Acad Sci. 1320(1):76-92. doi: 10.1111/nyas.12460.

Begum H., Spindel J.E., Lalusin A., et al (2015) Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). PLoS ONE 10: https://doi.org/10.1371/journal.pone.0119873.

Bellot P., de Los Campos G., Pérez-Enciso M. (2018). Can deep learning improve genomic prediction of complex human traits? Genetics 210, 809–819. doi: 10.1534/genetics.118.301298.

Berg J. J., and Coop G., (2014) A Population Genetic Signal of Polygenic Adaptation. PLoS Genetics 10. https://doi.org/10.1371/journal.pgen.1004412.

Bhattacharyya M. K., Smith A. M., Ellis T. H. N., Hedley C., Martin C., (1990) The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell 60, 115–122. doi: 10.1016/0092-8674(90)90721-p.

Bierne N., Eyre-Walker A. (2004) The genomic rate of adaptive amino acid substitution in Drosophila. Mol Biol Evol. 21(7):1350-60. doi: 10.1093/molbev/msh134.

Birky C.W. Jr., Walsh J.B. (1988) Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci U S A. 85(17):6414-8. doi: 10.1073/pnas.85.17.6414.

Booker T.R. (2020) Inferring Parameters of the Distribution of Fitness Effects of New Mutations When Beneficial Mutations Are Strongly Advantageous and Rare.G3:Genes, Genomes, Genetics, 10(7):2317–2326. ISSN 2160-1836. doi: 10.1534/g3.120.401052.

Boucher J. I., P. Cote, J. Flynn, L. Jiang, A. Laban et al., (2014) Viewing protein fitness landscapes through a next-gen lens. Genetics 198: 461–471. doi: 10.1534/genetics.114.168351.

Bourque G., Burns K.H., Gehring M., Gorbunova V., Seluanov A., Hammell M., Imbeault M., Izsvák Z., Levin H.L., Macfarlan T.S., Mager D.L., Feschotte C. (2018) Ten things you should know about transposable elements. Genome Biol. 19(1):199. doi: 10.1186/s13059-018-1577-z.

Boyko A.R., Williamson S.H., Indap A.R., et al. (2008) Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. PLOS Genetics, 4(5):e1000083. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000083.

Boyko A. R., Boyko R. H., Boyko C. M., Parker H. G., Castelhano M., et al., (2009) Complex population structure in African village dogs and its implications for inferring dog domestication history. Proceedings of the National Academy of Sciences of the United States of America 106: 13903–13908.

Boyle E.A., Li Y.I., Pritchard J.K. (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 169(7):1177-1186. doi: 10.1016/j.cell.2017.05.038.

Britten R. J., and Davidson E. H. (1969) Gene regulation for higher cells: A theory. Science 165, 349–357. doi: 10.1126/science.165.3891.349.

Brown P.O., Bowerman B., Varmus H.E., Bishop J.M. (1987) Correct integration of retroviral DNA in vitro. Cell. 49:347–56. doi: 10.1016/0092-8674(87)90287-x.

Browning B.L., Zhou Y., Browning S.R. (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. American Journal of Human Genetics 103:338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Budhlakoti N., Kushwaha A.K., Rai A., Chaturvedi K.K., Kumar A., Pradhan A.K., Kumar U., Kumar R.R., Juliana P., Mishra D.C., Kumar S. (2022) Genomic Selection: A Tool for Accelerating the Efficiency of Molecular Breeding for Development of Climate-Resilient Crops. Front Genet. 13:832153. doi: 10.3389/fgene.2022.832153.

Bulmer M. G. (1980) The mathematical theory of quantitative genetics. Oxford, UK: Oxford University Press.

Bustamante C.D., Nielsen R., and Hartl D.L. (2002) A Maximum Likelihood Method for Analyzing Pseudogene Evolution: Implications for Silent Site Evolution in Humansand Rodents. Molecular Biology and Evolution, 19(1):110–117. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a003975.

Bustos-Korts D., Boer M.P., Malosetti M., et al (2019) Combining Crop Growth Modeling and Statistical Genetic Modeling to Evaluate Phenotyping Strategies. Frontiers in Plant Science 10: https://doi.org/10.3389/fpls.2019.01491.

Butelli E., Licciardello C., Zhang Y., Liu J., Mackay S., Bailey P., Reforgiato-Recupero G., Martin C. (2012). Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24:1242–1255. doi: 10.1105/tpc.111.095232.

Campos-Sánchez R., Cremona M.A., Pini A., Chiaromonte F., Makova K.D. (2016) Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. PLoS Comput Biol. 12(6):e1004956. doi: 10.1371/journal.pcbi.1004956.

Capblancq T., Fitzpatrick M.C., Bay R.A., Exposito-Alonso M., Keller S.R. (2020) Genomic prediction of (mal)adaptation across current and future climatic landscapes. Annu Rev Ecol Evol Syst. 245–69.

Carpentier M.C., Manfroi E., Wei F.J., et al (2019) Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nature Communications 10: https://doi.org/10.1038/s41467-018-07974-5.

Casacuberta E., González J. (2013) The impact of transposable elements in environmental adaptation. Mol Ecol. 22(6):1503-17. doi: 10.1111/mec.12170.

Casillas, S. and Barbadilla, A. (2017) Molecular Population Genetics. Genetics, 205(3):1003–1035. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.116.196493.

Castanera R., Vendrell-Mir P., Bardil A., et al (2021) Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. Plant Journal 107:118–135. https://doi.org/10.1111/tpj.15277.

Castellano D., Macià M. C., Tataru P., Bataillon T., and Munch K., (2019) Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. Genetics 213: 953–966. https://doi.org/10.1534/genetics.119.302494.

Chan M., Scarafoni D., Duarte R., Thornton J., Skelly L. (2018). "Learning network architectures of deep CNNs under resource constraints," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. (IEEE), 1784–1791. doi: 10.1109/CVPRW.2018.00222.

Chang C.C., Chow C.C., Tellier L.C.A.M., et al (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4: https://doi.org/10.1186/s13742-015-0047-8.

Charlesworth B. (2010) Molecular population genomics: a short history. Genet Res (Camb). 92(5-6):397-411. doi: 10.1017/S0016672310000522.

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

Charlesworth B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetics Research, 63(3):213–227. ISSN 1469- 5073, 0016-6723. doi: 10.1017/S0016672300032365.

Chen K., Łyskowski A., Jaremko Ł., Jaremko M. (2021) Genetic and Molecular Factors Determining Grain Weight in Rice. Frontiers in Plant Science 12. doi: 10.3389/fpls.2021.605799.

Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T. et al. (2017) The impact of structural variation on human gene expression. Nat. Genet. 49, 692–699. doi: 10.1038/ng.3834.

Chuong E.B., Elde N.C., Feschotte C. (2017) Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 18(2):71-86. doi: 10.1038/nrg.2016.139.

Cockerham C.C. (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39,859–882. doi: 10.1093/genetics/39.6.859.

Cook L.M., Saccheri I.J. (2013) The peppered moth and industrial melanism: evolution of a natural selection case study. Heredity (Edinb). 110(3):207-12. doi: 10.1038/hdy.2012.92.

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., et al. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science 338, 1206–1209. doi: 10.1126/science.1228746.

Coop G. (2020), Population and Quantitative Genetics.

Cooper M., Technow F., Messina C., et al (2016) Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. Crop Science 56. https://doi.org/10.2135/cropsci2015.08.0512.

Crow J. F., and Kimura M., (1979) Efficiency of truncation selection. Proceedings of the National Academy of Sciences of the United States of America 76. https://doi.org/10.1073/pnas.76.1.396.

Cuevas J., Montesinos-López O., Juliana P., Guzmán C., Pérez-Rodríguez P., González-Bucio J., Burgueño J., Montesinos-López A., Crossa J. (2019) Deep Kernel for Genomic and Near Infrared Predictions in Multi-environment Breeding Trials. G3 (Bethesda). 9(9):2913-2924. doi: 10.1534/g3.119.400493.

Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O.A., Burgueno, J., Bandeira, M.B.E., Crossa, J. (2018) Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. G3-Genes Genomes Genet. 8, 1347–1365.

Cui Y., Li R., Li G., Zhang F., Zhu T, Zhang Q., Ali J., Li Z, Xu S. (2020) Hybrid breeding of rice via genomic selection. Plant Biotechnol J. 18(1):57-67. doi: 10.1111/pbi.13170.

Daborn P.J., Yen J.L., Bogwitz M.R., Le Goff G., Feil E., Jeffers S., Tijet N., Perry T., Heckel D., Batterham P., et al. (2002). A single p450 allele is associated with insecticide resistance in Drosophila. Science 297:2253–2256. doi: 10.1126/science.1074170.

Daetwyler H.D., Bansal U.K., Bariana H.S., Hayden M.J., Hayes B.J. (2014) Genomic prediction for rust resistance in diverse wheat landraces. Theor Appl Genet. 127(8):1795-803. doi: 10.1007/s00122-014-2341-8.

Daetwyler H.D., Pong-Wong R., Villanueva B., Woolliams J.A. (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: https://doi.org/10.1534/genetics.110.116855.

Darwin C. (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray.

Darwin C. (1868) The variation of animals and plants under domestication. London: John Murray.

Darwin C. (1958) The autobiography of Charles Darwin 1809–1882. London: Collins.

Daub J.T., Hofer T., Cutivet E., Dupanloup I., Quintana-Murci L., Robinson-Rechavi M., Excoffier L. (2013) Evidence for polygenic adaptation to pathogens in the human genome. Mol Biol Evol. 30(7):1544-58. doi: 10.1093/molbev/mst080.

Dayan T., (1994) Early domesticated dogs of the near east. Journal of Archaeological Science 21. https://doi.org/10.1006/jasc.1994.1062.

De los Campos G. (2018), Analysis & Prediction of Complex Traits using Whole-Genome Regression Methods

De los Campos G.m Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L. (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193, 327–345. doi: 10.1534/genetics.112.143313.

de Vladar H. P. de, and Barton N., (2014) Stability and response of polygenic traits to stabilizing selection and mutation. Genetics 197: 749–767. https://doi.org/10.1534/genetics.113.159111.

Desta Z.A., Ortiz R. (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. 19(9):592-601. doi: 10.1016/j.tplants.2014.05.006.

Dobzhansky Th. (1937) Genetics and the Origin of Species (Columbia Univ. Press, New York); 2nd Ed., 1941; 3rd Ed., 1951.

Domínguez M., Dugas E., Benchouaia M., et al (2020) The impact of transposable elements on tomato diversity. Nature Communications 11: https://doi.org/10.1038/s41467-020-17874-2.

Driscoll C. A., Macdonald D. W., and O'Brien S. J., (2009) From wild animals to domestic pets, an evolutionary view of domestication. Proceedings of the National Academy of Sciences of the United States of America 106 Suppl 1: 9971–9978. https://doi.org/10.1073/pnas.0901586106.

Dubin M.J., Mittelsten Scheid O., Becker C. (2018) Transposons: a blessing curse. Current Opinion in Plant Biology 42:23–29. doi: 10.1016/j.pbi.2018.01.003.

Ehret A., Hochstuhl D., Krattenmacher N., Tetens J., Klein M.S., Gronwald W., Thaller G. (2015) Short communication: Use of genomic and metabolic information as well as milk performance records for prediction of subclinical ketosis risk via artificial neural networks. J Dairy Sci. 98(1):322-9. doi: 10.3168/jds.2014-8602.

Evans S.N., Shvets Y., and Slatkin M. (2007) Non-equilibrium theory of the allele frequency spectrum. Theoretical Population Biology, 71(1):109–119. ISSN 0040-5809. doi:10.1016/j.tpb.2006.06.005.

Ewens W.J., (1965), Two diffusion distributions in genetics, Annals of Humans Genetics, 29(1), p.1-4, https://doi.org/10.1111/j.1469-1809.1965.tb00496.x

Eyre-Walker A., and Keightley P.D., (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26: 2097–2108. doi: 10.1093/molbev/msp119.

Eyre-Walker A., Keightley P.D. (2007) The distribution of fitness effects of new mutations. Nat Rev Genet. 8(8):610-8. doi: 10.1038/nrg2146.

Eyre-Walker A., Woolfit M., and Phelps T. (2006) The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. Genetics, 173(2):891–900. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.106.057570.

Falconer D. S. & Mackay T. F. C. (1996) Introduction to quantitative genetics, 4th edn. Harlow, UK:Longman.

Fay J.C., Wyckoff G.J., and Wu C.I. (2001) Positive and Negative Selection on the Human Genome. Genetics, 158(3):1227–1234. ISSN 0016-6731, 1943-2631. doi: 10.1093/genetics/158.3.1227.

Fernando R.L., Grossman M. (1989) Marker assisted selection using best linear unbiased prediction. Genet Sel Evol. 21(4):467–77. doi: 10.1186/1297-9686-21-4-467.

Fisher R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edin. 52, 399 –433.

Fisher R.A. (1930) The genetical theory of natural selection. The genetical theory of natural selection. Clarendon Press, Oxford, England. doi: 10.5962/bhl.title.27468.Pages: xiv, 272.

Flood P. J., and Hancock A. M., (2017) The genomic basis of adaptation in plants. Current Opinion in Plant Biology 36. doi: 10.1016/j.pbi.2017.02.003.

Flori L., Moazami-Goudarzi K., Alary V., Araba A., Boujenane I., et al., (2019) A genomic map of climate adaptation in Mediterranean cattle breeds. Molecular Ecology 28. https://doi.org/10.1111/mec.15004 for genomic selection. Livest. Prod. Sci. 166:54-65

Fowler D. M., C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany et al., (2010) High-resolution mapping of protein sequence-function relationships. Nat. Methods 7: 741–746.

Franssen S.U., Nolte V., Tobler R., Schlötterer C. (2015) Patterns of linkage disequilibrium and long-range hitchhiking in evolving experimental Drosophila melanogaster populations. Mol Biol Evol. 32(2):495-509. doi: 10.1093/molbev/msu320.

Frantz L. A. F., Bradley D. G., Larson G., and Orlando L., (2020) Animal domestication in the era of ancient genomics. Nature Reviews Genetics 21.

Frantz L. A. F., Schraiber J. G., Madsen O., Megens H. J., Cagan A., et al., (2015) Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nature Genetics 47. https://doi.org/10.1038/ng.3394.

Fuentes R.R., Chebotarov D., Duitama J., et al., (2019) Structural variants in 3000 rice genomes. Genome Research 29:870–880. https://doi.org/10.1101/gr.241240.118.

Galtier N. (2016) Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet. 12(1):e1005774. doi: 10.1371/journal.pgen.1005774.

Galtier N., and Rousselle M. (2020) How Much Does Ne Vary Among Species? Genetics,216(2):559–572. ISSN 1943-2631. doi: 10.1534/genetics.120.303622.

Gao L., Gonda I., Sun H., Ma Q., Bao K., Tieman D., M, Burzynski-Chang E., A, Fish T.L., Stromberg K.A., Sacks G.L., Thannhauser T.W., Foolad M.R., Diez M.J., Blanca J., Canizares J., Xu Y., van der Knaap E., Huang S., Klee H.J., Giovannoni J.J., Fei Z. (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. 51(6):1044-1051. doi: 10.1038/s41588-019-0410-2.

Garud N.R., Messer P.W., Buzbas E.O., Petrov D.A. (2015) Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 11(2):e1005004. doi: 10.1371/journal.pgen.1005004.

Gianola D., Fernando R.L., Stella A. (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics. 173(3):1761-76. doi: 10.1534/genetics.105.049510.

Gianola D., Foulley J. (1983) Sire evaluation for ordered categorical data with a threshold model. Genetics Selection Evolution 15:201. https://doi.org/10.1186/1297-9686-15-2-201.

Gianola D., de los Campos G., Hill W.G., Manfredi E., Fernando R. (2009) Additive genetic variability and the Bayesian alphabet. Genetics. 183(1):347-63. doi: 10.1534/genetics.109.103952.

Gianola D., Okut H., Weigel K.A., Rosa G.J. (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 12:87. doi: 10.1186/1471-2156-12-87.

Gilbert C., Feschotte C. (2018) Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr Opin Genet Dev. 49:15-24. doi: 10.1016/j.gde.2018.02.007.

Gillespie J.H. (1989) Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol. 6(6):636-47. doi: 10.1093/oxfordjournals.molbev.a040576.

Gillespie J.H. (2000) Genetic drift in an infinite population. The pseudohitchhiking model. Genetics. 155(2):909-19. doi: 10.1093/genetics/155.2.909.

Gillespie J.H. (1994). The Causes of Molecular Evolution. Oxford University Press. ISBN 0-19-509271-6.

Goddard M.E., Hayes B.J. (2007) Genomic selection. Journal of Animal Breeding and Genetics 124:323–330. https://doi.org/10.1111/j.1439-0388.2007.00702.x.

Golicz A.A., Bayer PE., Barker G.C., Edger P.P., Kim H., Martinez P.A., Chan C.K., Severn-Ellis A., McCombie W.R., Parkin I.A., Paterson A.H., Pires J.C., Sharpe A.G., Tang H., Teakle G.R., Town C.D., Batley J., Edwards D. (2016) The pangenome of an agronomically important crop plant Brassica oleracea. Nat Commun. 7:13390. doi: 10.1038/ncomms13390.

González-Recio O., Rosa G. J. M., Gianola, D. (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest. Sci. 166, 217–231. doi: 10.1016/j.livsci.2014.05.036.

Goodfellow I. J., Bengio Y., and Courville, A. C. (2016). Deep learning. Adaptive computation and machine learning. Cambridge: MIT Press.

Gorkovskiy A., Verstrepen K.J. (2021) The Role of Structural Variation in Adaptation and Evolution of Yeast and Other Fungi. Genes (Basel). 12(5):699. doi: 10.3390/genes12050699.

Grabundzija I., Messing S.A., Thomas J., Cosby R.L., Bilic I., Miskey C., et al. A (2016) Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat Commun. 7:10716.

Granleese T., Clark S. A., and van der Werf J. H. J., (2019) Genotyping strategies of selection candidates in livestock breeding programs. Journal of Animal Breeding and Genetics 136: 91–101. https://doi.org/10.1111/jbg.12381.

Greenblatt I.M., Brink R.A. (1963) Transpositions of modulator in maize into divided and undivided chromosome segments. Nature. 197:412–3.

Gregory T.R. (2009) Artificial Selection and Domestication: Modern Lessons from Darwin's Enduring Analogy. *Evo Edu Outreach* **2**, 5–27. https://doi.org/10.1007/s12052-008-0114-z.

Grenier C., Cao T.V., Ospina Y., Quintero C., Châtel M.H., Tohme J., Courtois B., Ahmadi N. (2015) Accuracy of Genomic Selection in a Rice Synthetic Population Developed for Recurrent Selection Breeding. PLoS One. 10(8):e0136594. doi: 10.1371/journal.pone.0136594. Erratum in: PLoS One. 2016;11(5):e0154976.

Groen S.C., Joly-Lopez Z., Platts A.E., Natividad M., Fresquez Z., Mauck W.M., Quintana M.R., Cabral C.L.U., Torres R.O., Satija R., Purugganan M.D., Henry A. (2022) Evolutionary systems biology reveals patterns of rice adaptation to drought-prone agro-ecosystems. Plant Cell. 34(2):759-783. doi: 10.1093/plcell/koab275.

Groenen M. A. M., Archibald A. L., Uenishi H., Tuggle C. K., Takeuchi Y., et al., (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491. https://doi.org/10.1038/nature11622.

Gulli A., and Pal S. (2017). Deep learning with Keras. Birmingham: Packt Publishing Ltd.

Guo T., Yu X., Li X., Zhang H., Zhu C., Flint-Garcia S., McMullen M.D., Holland J.B., Szalma S.J., Wisser R.J., Yu J. (2019) Optimal Designs for Genomic Selection in Hybrid Crops. Mol Plant. 12(3):390-401. doi: 10.1016/j.molp.2018.12.022.

Gutenkunst R. N., Hernandez R. D., Williamson S. H., and Bustamante C. D. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5. https://doi.org/10.1371/journal.pgen.1000695.

Haberer G., Kamal N., Bauer E., et al. (2020) European maize genomes highlight intraspecies variation in repeat and gene content. Nature Genetics 52:950–957. https://doi.org/10.1038/s41588-020-0671-9.

Habier D., Fernando R.L., Kizilkaya K., Garrick D.J. (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 12:186. doi: 10.1186/1471-2105-12-186.

Haldane J.B.S. (1932) The causes of evolution. Macmillan.

Haller B. C., and Messer P. W. (2016) SLiM: An Evolutionary Simulation Framework. http://benhaller.com/slim/SLiM_Manual.pdf 3.4: 1–600.

Haller B. C., and Messer P. W., (2017) AsymptoticMK: A web-based tool for the asymptotic McDonald-Kreitman test. G3: Genes, Genomes, Genetics 7. https://doi.org/10.1534/g3.117.039693.

Halligan D.L, Kousathanas A., Ness R.W., Harr B., Eöry L., Keane T.M., Adams D.J., Keightley P.D. (2013) Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9(12): e1003995. doi: 10.1371/journal.pgen.1003995.

Halligan D.L. and Keightley P. D. (2009), Spontaneous Mutation Accumulation Studies in Evolutionary Genetics, Annual Review of Ecology, Evolution, and Systematics 40:1, 151-172, https://doi.org/10.1146/annurev.ecolsys.39.110707.173437.

Harris K., Nielsen R. (2016) The Genetic Cost of Neanderthal Introgression. Genetics. 203(2):881-91. doi: 10.1534/genetics.116.186890.

Harrisson K. A., Amish S. J., Pavlova A., Narum S. R., Telonis-Scott M., et al. (2017) Signatures of polygenic adaptation associated with climate across the range of a threatened fish species with high genetic connectivity. Molecular Ecology 26. https://doi.org/10.1111/mec.14368.

Hartfield M., Otto S.P. (2011) Recombination and hitchhiking of deleterious alleles. Evolution. 65(9):2421-34. doi: 10.1111/j.1558-5646.2011.01311.x.

Hartl D. L., Moriyama E. N., and Sawyer S. A., (1994) Selection intensity for codon bias. Genetics 138. https://doi.org/10.1093/genetics/138.1.227.

Hartl D. L., Clark A. G. (2007) Principles of Population Genetics., Écoscience, 14:4, 544-545.

Hassen B.M., Cao T.V., Bartholomé J., Orasen G., Colombi C., Rakotomalala J., Razafinimpiasa L., Bertone C., Biselli C., Volante A., Desiderio F., Jacquin L., Valè G., Ahmadi N. (2018) Rice diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel. Theor Appl Genet. 131(2):417-435. doi: 10.1007/s00122-017-3011-4.

Hazzouri K.M., Flowers J.M., Visser H.J., Khierallah H.S.M., Rosas U., Pham G.M., Meyer R.S., Johansen C.K., Fresquez Z.A., Masmoudi K., Haider N., El Kadri N., Idaghdour Y., Malek J.A., Thirkhill D., Markhand G.S., Krueger R.R., Zaid A., Purugganan M.D. (2015) Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. Nat Commun. 6:8824. doi: 10.1038/ncomms9824.

Heffner E.L., Sorrells M.E., Jannink, J.L. (2009) Genomic selection for crop improvement. Crop Sci. 49, 1–12. https://doi.org/10.2135/cropsci2008.08.0512.

Henn B.M., Botigué L.R., Peischl S., Dupanloup I., Lipatov M., Maples B.K., Martin A.R., Musharoff S., Cann H., Snyder M.P., Excoffier L., Kidd J.M., Bustamante C.D. (2016) Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci U S A. 113(4):E440-9. doi: 10.1073/pnas.1510805112.

Herbrich R., Graepel T., Campbell C. (1999) Bayes Point Machines: Estimating the Bayes Point in Kernel Space. In: Proceedings of IJCAI Workshop Support Vector Machines. Stockholm, pp 23–27.

Hermisson J. and Pennings P. S., (2017), Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation, Methods in Ecology and Evolution, 8(6):700-716, https://doi.org/10.1111/2041-210X.12808.

Hermisson J., and Pennings P. S., (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169. https://doi.org/10.1534/genetics.104.036947.

Heslot, N., Yang, H.P., Sorrells, M.E., Jannink, J.L. (2012) Genomic selection in plant breeding: A comparison of models. Crop Sci. 52, 146–160. https://doi.org/10.2135/cropsci2011.06.0297.

Hickey L.T.N, Hafeez A., Robinson H., Jackson S.A., Leal-Bertioli S.C.M., Tester M., Gao C., Godwin I.D., Hayes B.J., Wulff B.B.H. (2019) Breeding crops to feed 10 billion. Nat Biotechnol. 37(7):744-754. doi: 10.1038/s41587-019-0152-9.

Hietpas R.T., Jensen J.D., Bolon D.N. (2011) Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A. 108(19):7896-901. doi: 10.1073/pnas.1016024108.

Hill W.G. (2010) Understanding and using quantitative genetic variation. Philos Trans R Soc Lond B Biol Sci. 365(1537):73-85. doi: 10.1098/rstb.2009.0203.

Höllinger I., Pennings P. S., and Hermisson J. (2019) Polygenic adaptation: From sweeps to subtle frequency shifts. PLoS Genetics 15. https://doi.org/10.1371/journal.pgen.1008035.

Huang M., Balimpoya E.G., Mgonja E.M, McHale L.K., Luzi-Kihupi A., Wang G.L., Sneller C.H., (2019) Use of genomic selection in breeding rice (oryza sativa L.) for resistance to rice blast (magnaporthe oryzae), Mol. Breed. 39 114. https://dx.doi.org/10.1007/s11032-019-1023-2.

Huang X., Fortier A. L., Coffman A. J., Struck T. J., Irby M. N., et al. (2021) Inferring Genome-Wide Correlations of Mutation Fitness Effects between Populations. Molecular Biology and Evolution 38. https://doi.org/10.1093/molbev/msab162.

Huang X., Kurata N., Wei X., Wang Z. X., Wang A., et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490. https://doi.org/10.1038/nature11532

Huber C.D., Kim B.Y., Marsden C.D., Lohmueller K.E. (2017) Determining the factors driving selective effects of new nonsynonymous mutations. Proc Natl Acad Sci U S A. 114(17):4465-4470. doi: 10.1073/pnas.1619508114.

Hudson R.R., (2001) Two-locus sampling distributions and their application. Genetics 159. https://doi.org/10.1093/genetics/159.4.1805.

Hudson R.R., Kaplan N.L. (1995) Deleterious background selection with recombination. Genetics. 141(4):1605-17. doi: 10.1093/genetics/141.4.1605.

Isidro J., Jannink J.L., Akdemir D., Poland J., Heslot N., Sorrells M.E. (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet. 128(1):145-58. doi: 10.1007/s00122-014-2418-4.

Iwata H., Ebana K., Uga Y., Hayashi T. (2015) Genomic prediction of biological shape: elliptic Fourier analysis and kernel partial least squares (PLS) regression applied to grain shape prediction in rice (Oryza sativa L.). PLoS One. 10(3):e0120610. doi: 10.1371/journal.pone.0120610.

Jackson M.T. (1997) Conservation of rice genetic resources: The role of the International Rice Genebank at IRRI. Plant Molecular Biology 35:61–67. https://doi.org/10.1007/978-94-011-5794-0_6.

Jain K., and Stephan W. (2015) Response of polygenic traits under stabilizing selection and mutation when loci have unequal effects. G3: Genes, Genomes, Genetics 5. https://doi.org/10.1534/g3.115.017970.

Jain K., and Stephan W. (2017a) Modes of Rapid Polygenic Adaptation. Mol Biol Evol.34(12):3169-3175. doi: 10.1093/molbev/msx240.

Jain K., and Stephan W. (2017b) Rapid adaptation of a polygenic trait after a sudden environmental shift. Genetics 206. https://doi.org/10.1534/genetics.116.196972.

Jasinska A. J., and Freimer N. B. (2009) The complex genetic basis of simple behavior. Journal of Biology 8. https://jbiol.biomedcentral.com/articles/10.1186/jbiol172.

Jerison E. R., Kryazhimskiy S., Desai M. S. (2014) Pleiotropic consequences of adaptation across gradations of environmental stress in budding yeast. arXiv:104.09.7839. https://doi.org/10.48550/arXiv.1409.7839.

Jia Z. (2017) Controlling the Overfitting of Heritability in Genomic Selection through Cross Validation. Sci Rep. 7(1):13678. doi: 10.1038/s41598-017-14070-z.

Jiang G.H., He Y.Q., Xu C.G., Li X.H., Zhang Q. (2004) The genetic basis of stay-green in rice analyzed in a population of doubled haploid lines derived from an indica by japonica cross. Theor. Appl. Genet., 108, 688–698.

Jighly A., Lin Z., Pembleton L.W., et al. (2019) Boosting Genetic Gain in Allogamous Crops via Speed Breeding and Genomic Selection. Frontiers in Plant Science 10: https://doi.org/10.3389/fpls.2019.01364.

Johnson T., Barton N. (2005). Theoretical models of selection and mutation on quantitative traits. Philos Trans R Soc Lond B Biol Sci. 360:1411–1425. http://doi:10.1098/rstb.2005.1667.

Kaler A.S., Purcell L.C., Beissinger T., Gillman J.D. (2022) Genomic prediction models for traits differing in heritability for soybean, rice, and maize. BMC Plant Biology 22:. https://doi.org/10.1186/s12870-022-03479-y.

Kassambara A., and Mundt F. (2020) Package 'factoextra': Extract and visualize the results of multivariate data analyses. CRAN- R Package 84.

Kawahara Y., de la Bastide M., Hamilton J.P., et al (2013) Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. Rice 6:3–10. https://doi.org/10.1186/1939-8433-6-4.

Keightley P.D., and Eyre-Walker A. (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177. https://doi.org/10.1534/genetics.107.080663.

Keightley P.D., and Jackson B.C. (2018) Inferring the Probability of the Derived vs. the Ancestral Allelic State at a Polymorphic Site. Genetics. 209(3):897-906. doi: 10.1534/genetics.118.301120.

Kempthorne O. (1954) The correlation between relatives in a random mating population. Proc. R. Soc. Lond. B 143, 102 –113. doi:10.1098/rspb.1954.0056.

Kim B. Y., Huber C. D., and Lohmueller K. E. (2017) Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics 206. https://doi.org/10.1534/genetics.116.197145.

Kim Y., Stephan W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 160(2):765-77. doi: 10.1093/genetics/160.2.765.

Kimura M. and Ohta, T. (1971) Protein Polymorphism as a Phase of Molecular Evolution. Nature, 229(5285):467–469. ISSN 1476-4687.doi: 10.1038/229467a0.

Kimura M. (1964) Diffusion models in population genetics. Journal of Applied Probability,1(2):177–232. ISSN 0021-9002, 1475-6072. doi: 10.2307/3211856.

Kimura M. (1968) Evolutionary Rate at the Molecular Level. Nature, 217(5129):624–626. ISSN 1476-4687. doi: 10.1038/217624a0.

Kimura M. (1955) Stochastic Processes and Distribution of Gene Frequencies Under Natural Selection. Cold Spring Harbor Symposia on Quantitative Biology, 20:33–53.ISSN 0091-7451, 1943-4456. doi: 10.1101/SQB.1955.020.01.006. Publisher: ColdSpring Harbor Laboratory Press.

Kincaid H.L. (1993) Selective Breeding and Domestication. In: Cloud, J.G., Thorgaard, G.H. (eds) Genetic Conservation of Salmonid Fishes. NATO ASI Series, vol 248. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-2866-1_27.

Kofler R., Nolte V., Schlötterer C. (2015) Tempo and mode of transposable element activity in Drosophila. PLOS Genet. 11: e1005406. doi: 10.1371/journal.pgen.1005406.

Kondrashov A. S. (1988) Deleterious mutations and the evolution of sexual reproduction. Nature 336.

Koufopanou V., Lomas S., Tsai I.J., Burt A. (2015) Estimating the Fitness Effects of New Mutations in the Wild Yeast Saccharomyces paradoxus. Genome Biol Evol. 7(7):1887-95. doi: 10.1093/gbe/evv112.

Kousathanas, A. and Keightley, P.D. (2013) A Comparison of Models to Infer the Distribution of Fitness Effects of New Mutations. Genetics, 193(4):1197–1208. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.112.148023.

Krishnappa G., Savadi S., Tyagi B.S., et al (2021) Integrated genomic selection for rapid improvement of crops. Genomics 113:1070–1086. doi: 10.1016/j.ygeno.2021.02.007.

Lande R., and Thompson R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124, 743–756. doi: 10.1093/genetics/124.3.743.

Larson G. and Fuller D.Q. (2014) The Evolution of Animal Domestication, Annual Review of Ecology, Evolution, and Systematics 45:1, 115-136. https://doi.org/10.1146/annurev-ecolsys-110512-135813.

Larson G., and Burger J. 2013 A population genetics view of animal domestication. Trends in Genetics 29. doi: 10.1016/j.tig.2013.01.003.

Larson G., Piperno D. R., Allaby R. G., Purugganan M. D., Andersson L., et al. (2014) Current perspectives and the future of domestication studies. Proceedings of the National Academy of Sciences of the United States of America 111. https://doi.org/10.1073/pnas.1323964111.

Lecun Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521, 436–444. doi: 10.1038/nature14539.

Lee K.M., Coop G. (2017) Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data. Genetics. 207(4):1591-1619. doi: 10.1534/genetics.117.300417.

Legarra A., Chistensen O. F., Aguilar I., and Misztal I. (2014) Single step, a general approach. Livestock Science 166:54-65. https://doi.org/10.1016/j.livsci.2014.04.029.

Leno-Colorado J., Guirao-Rico S., Pérez-Enciso M., and Ramos-Onsins S. E. (2020) Pervasive selection pressure in wild and domestic pigs. bioRxiv. doi: https://doi.org/10.1101/2020.09.09.289439.

Li G., Tang J., Zheng J., Chu C. (2021) Exploration of rice yield potential: Decoding agronomic and physiological traits. Crop Journal 9:577–589. https://doi.org/10.1016/j.cj.2021.03.014.

Li J., Davis B.W., Jern P., Dorshorst B.J., Siegel P.B., Andersson L. (2019) Characterization of the endogenous retrovirus insertion in *CYP19A1* associated with henny feathering in chicken. Mob DNA. 10:38. doi: 10.1186/s13100-019-0181-4.

Li L., Jamieson K., DeSalvo G., Rostamizadeh A. and Talwalkar A. (2018a) Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. Journal of Machine Learning Research 18 1-52. http://jmlr.org/papers/v18/16-558.html.

Li N., Xu R., Duan P., and Li Y. (2018b) Control of grain size in rice. Plant Reproduction 31(3):237-251. doi: 10.1007/s00497-018-0333-6.

Li Z., Fu B.Y., Gao Y.M., et al. (2014) The 3,000 rice genomes project. Gigascience 3:8. doi: 10.1186/2047-217X-3-8.

Li X., Guo K., Zhu X., Chen P., Li Y., Xie G., Wang L., Wang Y., Persson S., Peng L. (2017) Domestication of rice has reduced the occurrence of transposable elements within gene coding regions. BMC Genomics. 18(1):55. doi: 10.1186/s12864-016-3454-z.

Li Y., Vinckenbosch N., Tian G., Huerta-Sanchez E., Jiang T., Jiang H., Albrechtsen A., Andersen G., Cao H., Korneliussen T., Grarup N., Guo Y., Hellman I., Jin X., Li Q., Liu J., Liu X., Sparsø T., Tang M., Wu H., Wu R., Yu C., Zheng H., Astrup A., Bolund L., Holmkvist J., Jørgensen T., Kristiansen K., Schmitz O., Schwartz T.W., Zhang X., Li R., Yang H.., Wang J, Hansen T., Pedersen O., Nielsen R., Wang J. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet. 42(11):969-72. doi: 10.1038/ng.680.

Li Y., Xiao J., Wu J., Duan J., Liu Y., Ye X., Zhang X., Guo X., Gu Y., Zhang L., Jia J., Kong X. (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. New Phytol. 196(1):282-291. doi: 10.1111/j.1469-8137.2012.04243.x.

Lisch D. (2013) How important are transposons for plant evolution? Nature Reviews Genetics 14(1):49-61. doi: 10.1038/nrg3374.

Livieris I.E., Dafnis S.D.,Papadopoulos G.K., Kalivas D.P. (2020) A Multiple-Input Neural Network Model for Predicting Cotton Production Quantity: A case study, Algorithms, 13(11), 273. https://doi.org/10.3390/a13110273.

Lorenz A.J., Smith K.P. (2015) Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. Crop Sci., 55, 2657–2667. https://doi.org/10.2135/cropsci2014.12.0827.

Lourenco D.A.L, Fragomeni, B.O., Bradford, H.L., et al. (2017). Implications of SNP weighting on single-step genomic predictions for different reference population sizes. J. Anim. Breed. Genet. 134:46-471. doi: 10.1111/jbg.12288.

Lu J., Tang T., Tang H., Huang J., Shi S., Wu C.I. (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 22(3):126-31. doi: 10.1016/j.tig.2006.01.004.

Lu L., Chen J., Robb S.M.C, et al (2017) Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. Proc Natl Acad Sci U S A 114. https://doi.org/10.1073/pnas.1716459114.

Luan D.D., Korman M.H., Jakubczak J.L., Eickbush T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 72:595–605. doi: 10.1016/0092-8674(93)90078-5.

Lush J. L. (1937) Animal Breeding Plans. Ames, Iowa: Iowa State Press.

Lush J. L. (1943). Animal Breeding Plans. Edn 2. Charleston, South Carolina: Bibliolife DBA of Bibilio Bazaar.

Lynch M. (2007) The origins of genome architecture. 1st ed. Sunderland: Sinauer Associates.

Lynch. M., Walsh B. (1998) Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer Associates.

Ma W., Qiu Z., Song J., Cheng Q., Ma C. (2017) DeepGS: predicting phenotypes from genotypes using deep learning. bioRxiv 241414. doi: 10.1101/241414.

Maltecca C., Parker K.L., Cassady J.P. (2012) Application of multiple shrinkage methods to genomic predictions. J. Anim. Sci. 90, 1777–1787. doi: 10.2527/jas.2011-4350.

Mansueto L., Fuentes R.R., Borja F.N., et al. (2017) Rice SNP-seek database update: New SNPs, indels, and queries. Nucleic Acids Research 45: D1075–D1081. https://doi.org/10.1093/nar/gkw1135.

Mao L., Wood T.C., Yu Y., et al (2000) Rice Transposable elements: A survey of 73,000 sequence-tagged-connectors. Genome Research 10: https://doi.org/10.1101/gr.10.7.982.

Marsden C. D., del Vecchyo D. O., O'Brien D. P., Taylor J. F., Ramirez O., et al. (2016) Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proceedings of the National Academy of Sciences of the United States of America 113. https://doi.org/10.1073/pnas.1512501113.

Mather K.A., Caicedo A.L., Polato N.R., et al. (2007) The extent of linkage disequilibrium in rice (Oryza sativa L.). Genetics 177:2223–2232. https://doi.org/10.1534/genetics.107.079616.

Matos C.A.P., Thomas D.L., Gianola D., et al. (1997) Genetic Analysis of Discrete Reproductive Traits in Sheep Using Linear and Nonlinear Models: II. Goodness of Fit and Predictive Ability. Journal of Animal Science 75:88–94. https://doi.org/10.2527/1997.75188x.

Matsumine H., Herbst M. A., Ou S. H., Wilson J. D., McPhaul M. J. (1991) Aromatase mRNA in the extragonadal tissues of chickens with the henny-feathering trait is derived from a distinctive promoter structure that contains a segment of a retroviral long terminal repeat. Functional organization of the Sebright, Leghorn, and Campine aromatase genes. J. Biol. Chem. 266, 19900–19907.

Matsumoto T., Wu J., Kanamori H., et al. (2005) The map-based sequence of the rice genome. Nature 436:793–800. https://doi.org/10.1038/nature03895.

McClintock B. Components of action of the regulators Spm and Ac. Carnegie Institution of Washington Year Book 64, 527–536 (1965).

McDonald J.H., Kreitman M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature. 351(6328):652-4. doi: 10.1038/351652a0.

McManus K.F., Kelley J.L., Song S., Veeramah K.R., Woerner A.E., Stevison L.S., Ryder O.A., Ape Genome Project G, Kidd J.M., Wall J.D., Bustamante C.D., Hammer M.F. (2015) Inference of gorilla demographic and selective history from whole-genome sequence data. Mol Biol Evol. 32(3):600-12. doi: 10.1093/molbev/msu394.

Messer P.W., and Petrov D.A., (2013) Frequent adaptation and the McDonald-Kreitman test. Proceedings of the National Academy of Sciences of the United States of America 110. https://doi.org/10.1073/pnas.1220835110.

Meuwissen T. H. E., Hayes B. J., and Goddard M. E (2016). Genomic selection: A paradigm shift in animal breeding. Anim. Front. 6: 6-14. https://doi.org/10.2527/af.2016-0002.

Meuwissen T. H. E., Hayes B. J., and Goddard, M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157, 1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Meyer R. S., and Purugganan M. D. (2013). Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics 14: 840– 852. https://doi.org/10.1038/nrg3605.

Min S., Lee B., Yoon S. (2017). Deep learning in bioinformatics. Brief. Bioinform. 18, 851–869. doi: 10.1093/bib/bbw068.

Miyata T., Yasunaga T., Nishida T. (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci U S A. 77(12):7328-32. doi: 10.1073/pnas.77.12.7328.

Montenegro J.D., Golicz A.A., Bayer P.E., Hurgobin B., Lee H.T., Chan C.K., Visendi P. et al. (2017) The pangenome of hexaploid bread wheat. Plant J. 90, 1007–1013. https://doi.org/10.1111/tpj.13515.

Montesinos-López A., Montesinos-López O. A., Hernández-Suárez C. M., Gianola D., Crossa J. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3. Genes Genomes Genet. 8 (12), 3813–3828. doi: 10.1534/g3.118.200740.

Montesinos-López O.A., Martín-Vallejo J., Crossa J., Gianola D., Hernández-Suárez C. M., Montesinos-López A., et al. (2019). New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. G3. Genes Genomes Genet. 9, 1545–1556. doi: 10.1534/g3.119.300585.

Montesinos-López O.A., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J.A., Martini J.W.R., Fajardo-Flores S.B., Gaytan-Lugo L.S., Santana-Mancilla P.C., Crossa J. (2021) A review of deep learning applications for genomic selection. BMC Genomics. 22(1):19. doi: 10.1186/s12864-020-07319-x.

Moon S., Akey J.M. (2016) A flexible method for estimating the fraction of fitness influencing mutations from large sequencing datasets. Genome Res. 26(6):834-43. doi: 10.1101/gr.203059.115.

Moran P.a.P. (1962) The statistical processes of evolutionary theory. The statistical processes of evolutionary theory, Publisher: Clarendon Press; Oxford University Press.

Moutinho A.F., Bataillon T., Dutheil J.Y. (2020) Variation of the adaptive substitution rate between species and within genomes. *Evol Ecol* 34, 315–338. https://doi.org/10.1007/s10682-019-10026-z.

Moyers B.T., Morrell P.L., McKay J.K. (2018) Genetic Costs of Domestication and Improvement. J Hered. 109(2):103-116. doi: 10.1093/jhered/esx069.

Murga Moreno J. (2022), Cataloguing the shape and strength of positive selection on 1000 Genomes Project data, PhD thesis, Autonomous University de Barcelona.

Nachimuthu V.V., Raveendran M., Duraialaguraja S., et al. (2015) Analysis of Population Structure and Genetic Diversity in Rice Germplasm Using SSR Markers: An Initiative Towards Association Mapping of Agronomic Traits in Oryza Sativa. Rice 8. https://doi.org/10.1186/s12284-015-0062-5.

Nielsen N.H., Jahoor A., Jensen D., Orabi J., Cericola F., Edriss V., Jensen, J. (2016) Genomic prediction of seed quality traits using advanced barley breeding lines. PLoS ONE. 11(10):e0164494. doi: 10.1371/journal.pone.0164494.

Nielsen R. (2005) Molecular Signatures of Natural Selection. Annual Review of Genetics,39(1):197–218. doi: 10.1146/annurev.genet.39.073003. 112420.

Niu X.M., Xu Y.C., Li Z.W., Bian Y.T., Hou X.H., Chen J.F., Zou Y.P., Jiang J., Wu Q., Ge S., Balasubramanian S., Guo Y.L. (2019) Transposable elements drive rapid phenotypic variation in Capsella rubella. Proc Natl Acad Sci U S A. 116(14):6908-6913. doi: 10.1073/pnas.1811498116.

Nordborg M., Charlesworth B., Charlesworth D. (1996) The effect of recombination on background selection. Genet Res. 67(2):159-74. doi: 10.1017/s0016672300033619.

Norman A., Taylor J., Edwards J., Kuchel H. (2018) Optimising genomic selection in wheat: Effect of marker density, population size and population structure on prediction accuracy. G3-Genes Genomes Genet. 8, 2889–2899. doi: 10.1534/g3.118.200311.

Novo I., Santiago E., Caballero A. (2022) The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. PLoS Genet. 18(1):e1009764. doi: 10.1371/journal.pgen.1009764.

Ohta T. and Gillespie, J.H. (1996) Development of Neutral and Nearly Neutral Theories. Theoretical Population Biology, 49(2):128–142. ISSN 0040-5809. doi: 10.1006/tpbi.1996.0007.

Ohta T. (1973) Slightly Deleterious Mutant Substitutions in Evolution. Nature, 246(5428):96–98. ISSN 1476-4687. doi: 10.1038/246096a0.

Ohta T. (1989) The mutational load of a multigene family with uniform members. Genetical Research 53. https://doi.org/10.1017/S0016672300028020.

Olesen I., Perez-Enciso M., Gianola D., Thomas D.L. (1994) A comparison of normal and nonnormal mixed models for number of lambs born in Norwegian sheep. J Anim Sci 72. https://doi.org/10.2527/1994.7251166x.

Onogi A., Ideta O., Inoshita Y., Ebana K., Yoshioka T., Yamasaki M., Iwata H. (2015) Exploring the areas of applicability of whole-genome prediction methods for Asian rice (Oryza sativa L.). Theor Appl Genet. 128(1):41-53. doi: 10.1007/s00122-014-2411-y.

Orr H.A., Coyne J.A. (1992). The genetics of adaptation: a reassessment. Am Nat. 140(5):725-42. doi: 10.1086/285437.

Parker H. G. et al. (2009) An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science 325, 995–998. doi: 10.1126/science.1173275.

Pattanayak S. (2017). Unsupervised Learning with Restricted Boltzmann Machines and Auto-encoders. Pro Deep Learning with TensorFlow (Berkeley, CA: Apress), 279–343. doi: 10.1007/978-1-4842-3096-1_5.

Pérez P., de Los Campos G. (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. https://doi.org/10.1534/genetics.114.164442.

Pérez-Enciso M., Rincón J.C., Legarra A. (2015) Sequence- vs. chip-assisted genomic selection: Accurate biological information is advised. Genetics Selection Evolution 47: https://doi.org/10.1186/s12711-015-0117-5.

Pérez-Enciso M., Zingaretti L. (2019). A guide on deep learning for complex trait genomic prediction. Genes (Basel). 10, 553. doi: 10.3390/genes10070553

Pérez-Rodríguez P., Gianola D., González-Camacho J.M., Crossa J., Manès Y., Dreisigacker S. (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2(12):1595–605. doi: 10.1534/g3.112.003665.

Pray L. (2008) Transposons, or jumping genes: Not junk DNA? Nature Education 1(1):32.

Pritchard J. K., Pickrell J. K., and Coop G., (2010) The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology 20(4):R208-15. doi: 10.1016/j.cub.2009.11.055.

Purcell S., Neale B., Todd-Brown K., et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559–575. https://doi.org/10.1086/519795.

Purugganan M. D., and Fuller D. Q. (2009) The nature of selection during plant domestication. Nature. 2009 Feb 12;457(7231):843-8. doi: 10.1038/nature07895.

Purugganan M.D. (2022) What is domestication? Trends Ecol Evol. 37(8):663-671. doi: 10.1016/j.tree.2022.04.006.

Quadrana L., Silveira A.B., Caillieux E., Colot V. (2021) Detection of Transposable Element Insertions in Arabidopsis Using Sequence Capture. Methods Mol Biol. 2250:141-155. doi: 10.1007/978-1-0716-1134-0_14.

Ragsdale A. P., and Gutenkunst R. N. (2017) Inferring demographic history using two-locus statistics. Genetics 206. https://doi.org/10.1534/genetics.117.201251.

Ragsdale A. P., Coffman A. J., Hsieh P., Struck T. J., Gutenkunst R. N. (2016) Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations. Genetics 203. http://doi.org/10.1534/genetics.115.184812.

Ramírez O., Burgos-Paz W., Casas E., Ballester M., Bianco E., et al. (2015) Genome data from a sixteenth century pig illuminate modern breed relationships. Heredity 114. https://doi.org/10.1038/hdy.2014.81.

Raybaut P. (2009) Spyder Documentation. Spyder Project.

Redding R. W. (2015) The Pig and the Chicken in the Middle East: Modeling Human Subsistence Behavior in the Archaeological Record Using Historical and Animal Husbandry Data. Journal of Archaeological Research 23. https://doi.org/10.1007/s10814-015-9083-2.

Reinoso-Peláez E.L., Gianola D., González-Recio O. (2022) Genome enabled prediction methods based on machine learning. Methods Mol Biol 2467:189–218. https://doi.org/10.1007/978-1-0716-2205-6_7.

Reznick D. N. (2009) The Origin Then and Now: An Interpretive Guide to the Origin of Species. Princeton University Press, Princeton, NJ.

Robertsen C.D., Hjortshøj R.L., Janss L.L. (2019) Genomic selection in cereal breeding. Agronomy 9(2):95. https://doi.org/10.3390/agronomy9020095.

Rose N. H., Bay R. A., Morikawa M. K., and Palumbi S. R., (2018) Polygenic evolution drives species divergence and climate adaptation in corals. Evolution. 2018 Jan;72(1):82-94. doi: 10.1111/evo.13385.

Rosegrant M.W., Cline S.A. (2003) Global Food Security: Challenges and Policies. Science. 2003 Dec 12;302(5652):1917-9. doi: 10.1126/science.1092958.

Ross-Ibarra J., Morrel P. L. l, and Gaut B. S. (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proceedings of the National Academy of Sciences of the United States of America 104. https://doi.org/10.1073/pnas.0700643104.

Rossum G. van, and Drak F. L. (2006) Python Reference Manual. October 22. http://docs.python.org/ref/ref.html.

Rubin C. J., Zody M. C., J. Eriksson, J. R. S. Meadows, E. Sherwood, et al. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464. https://doi.org/10.1038/nature08832.

Sandhu K.S., Lozada D.N., Zhang Z., Pumphrey M.O. and Carter A.H. (2021) Deep Learning for Predicting Complex Traits in Spring Wheat Breeding Program.Front. PlantSci. 11: 613325.doi: 10.3389/fpls.2020.613325.

Sanseverino W., Hénaff E., Vives C., Pinosio S., Burgos-Paz W., Morgante M., Ramos-Onsins S.E., Garcia-Mas J., Casacuberta J.M. (2015) Transposon Insertions, Structural Variations, and SNPs Contribute to the Evolution of the Melon Genome. Mol Biol Evol. 32(10):2760-74. doi: 10.1093/molbev/msv152.

Santiago N., Herráiz C., Ramón Goñi J., et al. (2002) Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana. Molecular Biology and Evolution 19: https://doi.org/10.1093/oxfordjournals.molbev.a004052.

Sawyer S.A., Hartl D.L. (1992) Population genetics of polymorphism and divergence. Genetics. 132(4):1161-76. doi: 10.1093/genetics/132.4.1161.

Sawyer S.A., Kulathinal R.J., Bustamante C.D., et al. (2003) Bayesian Analysis Suggeststhat Most Amino Acid Replacements in Drosophila Are Driven by Positive Selection. Journal of Molecular Evolution, 57(1): S154–S164. doi:10.1007/s00239-003-0022-3.

Saxena R.K., Edwards D., Varshney R.K. (2014). Structural variations in plant genomes. Brief Funct Genomics. 13(4):296-307. doi: 10.1093/bfgp/elu016.

Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S., et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112-5. doi: 10.1126/science.1178534. Erratum in: Science. 2012 Aug 31;337(6098):1040.

Schneider A., Charlesworth B., Eyre-Walker A., Keightley P.D. (2011) A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics. 189(4):1427-37. doi: 10.1534/genetics.111.131730.

Schork N. J., Fallin D., and Lanchbury S. (2000) Single nucleotide polymorphisms and the future of genetic epidemiology. Clin Genet. 2000 Oct;58(4):250-64. doi: 10.1034/j.1399-0004.2000.580402.x.

Schrader L., Kim J.W., Ence D., Zimin A., Klein A., Wyschetzki K., Weichselgartner T., Kemena C., Stökl J., Schultner E., Wurm Y., Smith C.D., Yandell M., Heinze J., Gadau J., Oettler J. (2014) Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat Commun. 5:5495. doi: 10.1038/ncomms6495.

Sethupathy P. and Hannenhalli S. (2008) A Tutorial of the Poisson Random Field Model in Population Genetics. Advances in Bioinformatics, Volume 2008, Article ID 257864, 9 pages. doi:10.1155/2008/257864.

Seuret M., Alberti M., Liwicki M., Ingold R. (2017) PCA-Initialized Deep Neural Networks Applied to Document Image Analysis. IAPR International Conference on Document Analysis and Recognition (ICDAR) 01: 877-882. https://doi.ieeecomputersociety.org/10.1109/ICDAR.2017.148.

Shomura A., Izawa T., Ebana K., Ebitani T., Kanegae H., et al. (2008) Deletion in a gene associated with grain size increased yields during rice domestication. Nature Genetics 40(8):1023-8. https://doi.org/10.1038/ng.169.

Slowikowski K. (2020) ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2”.

Smith J.M., Haigh J. (1974) The hitch-hiking effect of a favorable gene. Genet Res. 23(1):23-35.

Smith N.G.C., and Eyre-Walker A. (2002) Adaptive protein evolution in Drosophila. Nature 415(6875):1022-4. doi: 10.1038/4151022a.

Sohail M., Maier R.M., Ganna A., Bloemendal A., Martin A.R., Turchin M.C., Chiang C.W., Hirschhorn J., Daly M.J., Patterson N., Neale B., Mathieson I., Reich D., Sunyaev S.R. (2019) Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife e39702. doi: 10.7554/eLife.39702.

Soller M., and Beckmann J. S. (1983). Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. 67(1):25-33. doi: 10.1007/BF00303917.

Song X., Cao X. (2017) Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. Curr Opin Plant Biol.; 36:111-118. doi: 10.1016/j.pbi.2017.02.004.

Sousa M.B.E., Cuevas J., Couto E.G.D., Perez-Rodriguez P., Jarquin D., Fritsche-Neto R., Burgueno J., Crossa J. (2017) Genomic-enabled prediction in maize using kernel models with genotype x environment interaction. G3-Genes Genomes Genet. 7(6):1995-2014. doi: 10.1534/g3.117.042341.

Spindel J.E., Begum H., Akdemir D., et al. (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity (Edinb) 116:395–408. doi: 10.1038/hdy.2015.113.

Spindel J., Begum H., Akdemir D., Virk P., Collard B., Redoña E., Atlin G., Jannink J.L., McCouch S.R. (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet. 11(2):e1004982. doi: 10.1371/journal.pgen.1004982. Erratum in: PLoS Genet. 2015 Jun;11(6):e1005350.

Stapley J., Santure A.W., Dennis S.R. (2015) Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol. 24(9):2241-52. doi: 10.1111/mec.13089.

Stephan W. (2016) Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Molecular Ecology 25. https://doi.org/10.1111/mec.13288.

Stephan W., and John S. (2020) Polygenic adaptation in a population of finite size. Entropy 22(8):907. doi: 10.3390/e22080907.

Steranka J.P., Tang Z., Grivainis M., et al. (2019) Transposon insertion profiling by sequencing (TIPseq) for mapping LINE-1 insertions in the human genome. Mob DNA 10:. https://doi.org/10.1186/s13100-019-0148-5.

Stetter M. G., Thornton K., and Ross-Ibarra J. (2018) Genetic architecture and selective sweeps after polygenic adaptation to distant trait optima. PLoS Genetics 14. https://doi.org/10.1371/journal.pgen.1007794.

Stoletzki N., Eyre-Walker A. (2011) Estimation of the neutrality index. Mol Biol Evol. 28(1):63-70. doi: 10.1093/molbev/msq249.

Sun C., Hu Z., Zheng T., et al. (2017) RPAN: Rice pan-genome browser for ∼3000 rice genomes. Nucleic Acids Research 45: https://doi.org/10.1093/nar/gkw958.

Sutton T., Baumann U., Hayes J., Collins N. C., Shi B.J., Schnurbusch T., et al. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318, 1446–1449. doi: 10.1126/science.1146853.

Tataru P. and Bataillon T. (2019) PolyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species. Bioinformatics 35(16):2868-2869. doi: 10.1093/bioinformatics/bty1060.

Tataru P., Mollion M., Glémin S., and Bataillon T. (2017) Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207(3):1103-1119. doi: 10.1534/genetics.117.300323.

Team R.C. (2021) R: A language and environment for statistical computing v. 3.6. 1 (R Foundation for Statistical Computing, Vienna, Austria, 2019). Scientific Reports 11:12957.

Tehseen M.M., Kehel Z., Sansaloni C.P., et al. (2021) Comparison of genomic prediction methods for yellow, stem, and leaf rust resistance in wheat landraces from afghanistan. Plants 10(3):558. https://doi.org/10.3390/plants10030558.

Tessema B.B., Liu H., Sørensen A.C., et al (2020) Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat. Frontiers in Genetics 11: https://doi.org/10.3389/fgene.2020.578123.

The International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928-33. doi: 10.1038/35057149.

Thornton K. R. (2019) Polygenic adaptation to an environmental shift: Temporal dynamics of variation under Gaussian stabilizing selection and additive effects on a single trait. Genetics 213(4):1513-1530. https://doi.org/10.1534/genetics.119.302662.

Tibshirani R. (2011) Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society Series B: Statistical Methodology 73:273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x.

Timpson N.J., Greenwood C.M.T., Soranzo N., Lawson D.J., Richards J.B. (2018) Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 19(2):110-124. doi: 10.1038/nrg.2017.101.

Trut L., Oskina I., and Kharlamova A. (2009) Animal evolution during domestication: The domesticated fox as a model. BioEssays 31(3):349-60. doi: 10.1002/bies.200800070.

Uricchio L.H., Kitano H.C., Gusev A., Zaitlen N.A. (2019a) An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evol Lett. 3(1):69-79. doi: 10.1002/evl3.97.

Uricchio L.H., Petrov D.A., Enard D. (2019b) Exploiting selection at linked sites to infer the rate and strength of adaptation. Nat Ecol Evol. 3(6):977-984. doi: 10.1038/s41559-019-0890-6.

Uricchio L.H., Zaitlen N.A., Ye C.J., Witte J.S., Hernandez R.D. (2016) Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res. 26(7):863-73. doi: 10.1101/gr.202440.115.

Van Laere A.S., Nguyen M., Braunschweig M., Nezer C., Collette C., Moreau L., Archibald A.L., Haley C.S., Buys N., Tally M., Andersson G., Georges M., Andersson L. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature. 425(6960):832-6. doi: 10.1038/nature02064.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. Journal of Dairy Science 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

van't Hof A.E., Edmonds N., Dalíková M., Marec F., Saccheri I.J. (2011) Industrial melanism in British peppered moths has a singular and recent mutational origin. Science. 332(6032):958-60. doi: 10.1126/science.1203043.

Vendrell-Mir P., Barteri F., Merenciano M., et al. (2019) A benchmark of transposon insertion detection tools using real data. Mob DNA 10:. https://doi.org/10.1186/s13100-019-0197-9.

Vignieri S. N., Larson J. G., Hoekstra H. E. (2010) The selective advantage of crypsis in mice. Evolution 64: 2153–2158. doi: 10.1111/j.1558-5646.2010.00976.x.

Visscher P.M., Yengo L., Cox N.J., Wray N.R. (2021) Discovery and implications of polygenicity of common diseases. Science (1979) 373:1468–1473. doi: 10.1126/science.abi8206.

Voss-Fels K.P., Cooper M., Hayes B.J. (2019) Accelerating crop genetic gains with genomic selection. Theor Appl Genet. 132(3):669-686. doi: 10.1007/s00122-018-3270-8.

Vourlaki I.T., Castanera R., Ramos-Onsins S.E., Casacuberta J.M., Pérez-Enciso M. (2022) Transposable element polymorphisms improve prediction of complex agronomic traits in rice. Theor Appl Genet. Sep;135(9):3211-3222. doi: 10.1007/s00122-022-04180-2.

Vu V.Q. (2011) ggbiplot: A ggplot2 based biplot. R package version.

Wang B., Li J. (2019) Understanding the molecular bases of agronomic trait improvement in rice. Plant Cell 31:1416–1417. https://doi.org/10.1105/tpc.19.00343.

Wang X., Xu Y., Xu Z., Zu C. (2018) Genomic selection methods for crop improvement: Current status and prospects, Crop J., 6(4):330-340. https://doi.org/10.1016/j.cj.2018.03.001.

Wang X., Li L., Yang Z., Zheng X., Yu S., Xu C., Hu Z. (2017) Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity (Edinb). 118(3):302-310. doi: 10.1038/hdy.2016.87.

Watterson G.A. (1962) Some Theoretical Aspects of Diffusion Theory in Population Genetics. Ann. Math. Statist. 33 (3) 939 – 957. https://doi.org/10.1214/aoms/1177704463.

Watterson G.A. (1975) On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7(2):256–276. ISSN 0040-5809. doi: 10.1016/0040-5809(75)90020-9.

Wellenreuther M., Merot C., Berdan E. and Bernatchez L. (2019) Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Mol. Ecol. 28(6):1203-1209. doi: 10.1111/mec.15066.

Wheelan S.J., Scheifele L.Z., Martinez-Murillo F., et al. (2006) Transposon insertion site profiling chip (TIP-chip). Proc Natl Acad Sci U S A. 103(47):17632-7. https://doi.org/10.1073/pnas.0605450103.

Widder D. V. (1954). The convolution transforms. Bull. Am. Math. Soc. 60, 444–456. doi: 10.1090/S0002-9904-1954-09828-2.

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 Genet. 3(6):e90. doi: 10.1371/journal.pgen.0030090.

Wright S. (1921) Correlation and Causation. Journal of Agricultural Research, 20, 557-585.

Wright S. (1931) Evolution in Mendelian Populations. Genetics. 16(2):97-159. doi: 10.1093/genetics/16.2.97.

Wright S. (1938) The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. USA 24(7):253-9. doi: 10.1073/pnas.24.7.253.

Wright S. (1945) The Differential Equation of the Distribution of Gene Frequencies. Proc Natl Acad Sci U S A. 31(12):382-9. doi: 10.1073/pnas.31.12.382.

Wright S.I., Bi I. V., Schroeder S. C., Yamasaki M., Doebley J. F., et al. (2005) Evolution: The effects of artificial selection on the maize genome. Science 308. https://doi.org/10.1126/science.1107891.

Würschum T., Boeven P.H., Langer S.M., Longin C.F., Leiser W.L. (2015) Multiply to conquer copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. BMC Genet 16: 96. doi:10.1186/s12863-015-0258-0.

Xiong H. Y., Alipanahi B., Lee L. J., Bret Schneider H., Merico D., Yuen R. K. C., et al. (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806. doi: 10.1126/science.1254806.

Xiong C., Zheng J., Xu L., Cen C., Zheng R., Li Y. (2021) Multiple-Input Convolutional Neural Network Model for Large-Scale Seismic Damage Assessment of Reinforced Concrete Frame Buildings. Appl. Sci., 11, 8258. https://doi.org/10.3390/app11178258.

Xu F., Sun X., Chen Y., et al. (2015) Rapid identification of major QTLs associated with rice grain weight and their utilization. PLoS ONE. 10(3):e0122206. https://doi.org/10.1371/journal.pone.0122206.

Xu S., Zhu D., Zhang Q. (2014) Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc Natl Acad Sci U S A. 111(34):12456-61. doi: 10.1073/pnas.1413750111.

Xu Y., Liu X., Fu J., Wang H., Wang J., Huang C., Prasanna B.M., Olsen M.S., Wang G., Zhang A. (2020) Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. Plant Commun. 1(1):100005. doi: 10.1016/j.xplc.2019.100005.

Xu Y., Ma K., Zhao Y. et al. (2021) Genomic selection: a breakthrough technology in rice breeding. Crop Journal 9:669–677. https://doi.org/10.1016/j.cj.2021.03.008.

Xu Y., Wang X., Ding X., et al. (2018) Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice. 11(1):32. https://doi.org/10.1186/s12284-018-0223-4.

Xue S., Bradbury P. J., Casstevens T., and Holland J. B. (2016) Genetic architecture of domestication-related traits in maize. Genetics 204(1):99-113. https://doi.org/10.1534/genetics.116.191106.

Yabe S., Yoshida H., Kajiya-Kanegae H., Yamasaki M., Iwata H., Ebana K., Hayashi T., Nakagawa H. (2018) Description of grain weight distribution leading to genomic selection for grain-filling characteristics in rice. PLoS One. 13(11):e0207627. doi: 10.1371/journal.pone.0207627.

Yadav S., Singh U.M., Naik S.M., Venkateshwarlu C., Ramayya P.J., Raman K.A., Sandhu N., Kumar A. (2017) Molecular Mapping of QTLs Associated with Lodging Resistance in Dry Direct-Seeded Rice (Oryza sativa L.). Front Plant Sci. Aug 21;8:1431. doi: 10.3389/fpls.2017.01431.

Yang Z., Bielawski J.P. (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15(12):496-503. doi: 10.1016/s0169-5347(00)01994-7.

Yang Z., Nielsen R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19(6):908-17. doi: 10.1093/oxfordjournals.molbev.a004148.

Young S.R., Rose D.C., Karnowski T.P., Lim S.H., Patton R.M. (2015) Optimizing deep learning hyper-parameters through an evolutionary algorithm. Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments - MLHPC '15, (New York, New York, USA: ACM Press), 1–5. doi: 10.1145/2834892.2834896.

Yu J., Pressoir G., Briggs W.H., Vroh Bi. I., Yamasaki M., Doebley J.F., McMullen M.D., Gaut B.S., Nielsen D.M., Holland J.B., Kresovich S., Buckler E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38(2):203-8. doi: 10.1038/ng1702.

Yuan Y., Bayer P.E., Batley J., Edwards D. (2021) Current status of structural variation studies in plants. Plant Biotechnol J. 19(11):2153-2163. doi: 10.1111/pbi.13646.

Zeder M.A. (2012) The domestication of animals. Journal of Anthropological Research 68. https://doi.org/10.3998/jar.0521004.0068.201.

Zeder M.A. (2015) Core questions in domestication research. Proceedings of the National Academy of Sciences of the United States of America 112. https://doi.org/10.1073/pnas.1501711112.

Zeder M.A., Emshwiller E., Smith B. D., and Bradley D. G. (2006) Documenting domestication: The intersection of genetics and archaeology. Trends in Genetics 22(3):139-55. doi: 10.1016/j.tig.2006.01.007.

Zeder M.A. (2011) The Origins of Agriculture in the Near East, The Univeristy of Chicago press journals.

Zhang M., Yang Q., Ai H., Huang L. (2022) Revisiting the evolutionary history of pigs via De Novo mutation rate estimation in a three-generation pedigree. Genomics Proteomics Bioinformatics. S1672-0229(22)00014-6. doi: 10.1016/j.gpb.2022.02.001.

Zhao H., Mitra N., Kanetsky P.A., et al. (2018) A practical approach to adjusting for population stratification in genome-wide association studies: Principal components and propensity scores (PCAPS). Statistical Applications in Genetics and Molecular Biology 17(6):/j/sagmb.2018.17.issue-6/sagmb-2017-0054/sagmb-2017-0054.xml. doi: 10.1515/sagmb-2017-0054.

Zhen Y., Huber C.D., Davies R.W., Lohmueller K.E. (2018) Stronger and higher proportion of beneficial amino acid changing mutations in humans compared to mice and flies. https://doi.org/10.1101/427583.

Zhen Y., Huber C.D., Davies R.W., Lohmueller K.E. (2021) Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and *Drosophila melanogaster*. Genome Res. 31(1):110-120. doi: 10.1101/gr.256636.119.

Zingaretti L. M., Gezan S. A., Ferrão L. F. V., Osorio L. F., Monfort A., Muñoz P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. Front. Plant Sci. 11:25. doi: 10.3389/fpls.2020.00025.

Żmieńko A., Samelak A., Kozłowski P., Figlerowicz M. (2014) Copy number polymorphism in plant genomes. Theor Appl Genet. 127(1):1-18. doi:10.1007/s00122-013-2177-7.

Appendices

Appendix A

Detection of domestication signals through the analysis of the full distribution of fitness effects using forward simulations and polygenic adaptation

Results

A. Summary statistics of the simulated populations

DOMESTIC POPULATION

Table A.0: Proportion of type of nonsynonymous mutations (m2-m7) per site and Observed number of fixed and polymorphic mutations per scenario (s1-s10)

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
SITE PROPORTION										
m2	0.0250	0.0238	0.0188	0.0250	0.0188	0.0188	0.0250	0.0238	0.0188	0.0238
m3	0.0000	0.0003	0.0006	0.0000	0.0016	0.0002	0.0000	0.0000	0.0016	0.0001
m7	0.0000	0.0122	0.0244	0.0000	0.0609	0.0061	0.0000	0.0013	0.0609	0.0049
m4	0.0000	0.0009	0.0056	0.0000	0.0047	0.0061	0.0000	0.0012	0.0047	0.0011
m6	0.0000	0.0366	0.2194	0.0000	0.1828	0.2377	0.0000	0.0475	0.1828	0.0439
m5	0.9750	0.9262	0.7312	0.9750	0.7312	0.7312	0.9750	0.9262	0.7312	0.9262
NUMBER FIXATIONS										
m2	6905	1352	1191	1553	1259	5333	7180	1807	1390	7018
m3	0	13	43	0	106	35	0	3	120	39
m7	0	288	614	0	1642	193	0	37	2017	180
m4	0	59	325	0	283	1686	0	80	317	296
m6	0	894	5282	0	4746	6540	0	1593	6127	1483
m5	25825	23013	17856	25019	19334	20260	33222	30906	24276	31765
NUMBER POLYMORPHISMS										
m2	1843	1714	1286	1331	891	1021	63	68	49	81
m3	0	29	46	0	84	10	0	1	6	1
m7	0	627	1233	0	2419	244	0	4	162	13
m4	0	57	353	0	183	265	0	3	16	0
m6	0	1851	10664	0	6200	7422	0	120	511	123
m5	39822	46254	35925	37619	25604	23661	2877	2616	2113	2634
FIXED SUBSTITUTIONS IN DOMESTIC / POLYMORPHIC OR ABSENT IN WILD										
m2	21	8	18	81	67	135	437	348	283	407
m3	0	0	1	0	8	0	0	1	21	5
m7	0	3	6	0	117	19	0	11	540	39
m4	0	0	1	0	18	21	0	17	65	26
m6	0	3	43	0	314	349	0	447	1574	321
m5	114	162	167	1264	1303	1212	7525	8169	6462	7370
NUMBER OF EXCLUSIVE POLYMORPHISMS										
m2	276	652	482	780	559	629	59	68	48	80
m3	0	13	20	0	53	3	0	1	6	1
m7	0	307	606	0	1766	204	0	4	162	13
m4	0	14	122	0	115	141	0	3	16	0
m6	0	843	4967	0	4561	5728	0	120	510	123
m5	10415	21003	16824	25984	18609	17845	2874	2610	2104	2628

Table A.1A: Number of fixed and polymorphic mutations at Wild, Domestic and both

Scenario	FW	FD	FBoth	PW	PD	PBoth
1	55727	55657	55446	65732	67901	82552
2	48458	47989	47645	72312	82962	108832
3	47958	47692	47193	73534	81549	110484
4	48282	49786	47377	74009	63839	117157
5	48615	51191	47785	71426	57637	112422
6	55449	57608	54503	64650	52329	103552
7	55230	68436	54909	65935	4364	70295
8	47548	62701	47221	72736	4240	76973
9	47869	63002	47582	72139	4383	76518
10	55805	68992	55340	65371	4310	69680

FW: Fixed variants at Wild. FD: Fixed variants at Domestic. FBoth: Fixed variants at the species. PW: Polymorphic variants at Wild. PD: Polymorphic variants at Domestic. PBoth: Polymorphic variants in the species

Table A.1B: Number of fixed, exclusive and shared synonymous variants observed in the Domestic populations for each scenario

Scenarios	SfWD	SfW	SfD	SxW	SxD	Ssh	SfWxD	SfDxW
1	22851	0	0	5018	5848	20269	119	76
2	22202	0	0	9091	12875	19178	377	168
3	22118	0	0	10220	13164	18545	333	263
4	22150	3	8	19082	15484	8983	422	1056
5	22242	16	53	19452	14550	7355	351	1526
6	22192	22	57	17868	13468	5847	391	1312
7	22469	125	1547	21571	1418	3	3	4018
8	21788	149	1704	23724	1422	3	3	4783
9	22280	131	1586	23554	1516	2	8	4889
10	22727	181	1519	21278	1448	1	9	3965

SfWD: Fixed variant in the species in relation to the outgroup. SfW: Exclusive fixed variant in Wild. SfD: Exclusive fixed variant in Domestic. SxW: Exclusive polymorphism in Wild, SxD: Exclusive polymorphisms in Domestic. Ssh: Shared polymorphic variants. SfWxD: Fixed variants in Wild and polymorphic in Domestic. SfDxW: Fixed in Domestic and polymorphic in Wild. S: Scenarios.

Table A.2: Ratios of Polymorphisms and Divergence at functional versus neutral positions per population

Scenarios	Pisyn		Ksyn		Pinsyn/Pisyn		Knsyn/Ksyn	
	WILD	**DOM**	**WILD**	**DOM**	**WILD**	**DOM**	**WILD**	**DOM**
1	0.00114	0.00114	0.01032	0.01032	0.720	0.720	0.706	0.706
2	0.00129	0.00136	0.01028	0.01028	0.703	0.705	0.581	0.582
3	0.00132	0.00124	0.01028	0.01027	0.693	0.698	0.581	0.580
4	0.00132	0.00113	0.01027	0.01028	0.701	0.710	0.584	0.584
5	0.00130	0.00098	0.01030	0.01030	0.688	0.713	0.585	0.585
6	0.00113	0.00084	0.01021	0.01019	0.716	0.744	0.716	0.715
7	0.00114	0.00003	0.01021	0.01021	0.722	1.012	0.715	0.719
8	0.00130	0.00003	0.01011	0.01012	0.703	0.985	0.593	0.598
9	0.00131	0.00003	0.01024	0.01024	0.699	0.890	0.583	0.589
10	0.00113	0.00003	0.01029	0.01027	0.726	1.057	0.707	0.713

Pisyn: Nucleotide diversity at neutral positions

Ksyn: Nucleotide divergence at neutral positions

Pinsyn/Pisyn: Ratio of nonsynonymous to synonymous polymorphisms

Knsyn/Ksyn: Ratio of nonsynonymous to synonymous divergence

B. True alpha

Table A.3: Number of Total, Shared and Exclusive beneficial fixations and Nonsynonymous fixed mutations

	All Variants						**Shared Variants**					**Exclusive Variants**				
	Wild	**Domestic**			**Wild**	**Domestic**	**Wild**	**Domestic**		**Wild**	**Domestic**	**Wild**	**Domestic**		**Wild**	**Domestic**
S	**Nben***	**Nben†**	**Nm2+m3§**	**Nm7¶**	**Nsynß**	**Nsynß**	**Nben***	**Nben†**	**Nm7¶**	**Nsynß**	**Nsynß**	**Nben***	**Nben†**	**Nm7¶**	**Nsynß**	**Nsynß**
1	6915	21	21	0	32757	135	6915	21	0	32757	135	0	0	0	0	0
2	1442	11	8	3	25879	176	1442	11	3	25879	176	0	0	0	0	0
3	1563	25	19	6	25507	236	1563	25	6	25507	236	0	0	0	0	0
4	1504	81	81	0	25707	1345	1504	81	0	25705	1329	0	0	0	0	0
5	1589	192	75	117	26006	1827	1586	184	110	25975	1748	3	8	7	31	79
6	6996	154	135	19	32844	1736	6989	147	19	32804	1654	7	7	0	40	82
7	6788	437	437	0	32633	7962	6747	367	0	32446	5242	41	70	0	187	2720
8	1543	360	349	11	25608	8993	1524	287	8	25439	6145	19	73	3	169	2848
9	1469	844	304	540	25450	8945	1459	622	380	25311	6169	10	222	160	139	2776
10	6984	451	412	39	32888	8168	6916	386	28	32620	5386	68	65	11	268	2782

* Nben = Number of fix (m2, m3, m4)); † Nben = Number of fix (m2, m3, m7); § Number of fix(m2,m3)); ¶ Number of fix(m7), ß Nsyn = Number of fix (nonsynonmous). s: scenario.

C. Estimation of the proportion of adaptive substitutions (alpha)

Table A.4: Proportion of adaptive variants with 95% confidence intervals (CI) using Asymptotic McDonald-Kreitman Test by 100 bootstraps sets with replacement

	Wild			Domestic		
Scenario	obs	CI05	CI95	obs	CI05	CI95
s1	0.090	0.040	0.149	0.112	0.081	0.173
s2	-0.068	-0.116	0.019	-0.048	-0.108	0.040
s3	-0.034	-0.093	0.087	-0.032	-0.087	0.088
s4	-0.104	-0.118	0.025	-0.085	-0.107	0.051
s5	0.039	-0.042	0.069	0.053	-0.046	0.137
s6	0.122	0.082	0.146	0.092	0.083	0.226
s7	0.211	0.141	0.248	-0.898	-2.932	-0.941
s8	-0.061	-0.086	0.084	-0.506	-1.400	0.434
s9	-0.016	-0.020	0.123	-1.404	-2.601	-0.736
s10	0.163	0.062	0.170	-0.883	-2.234	-0.514

CI05: 5% distribution from bootstrap analysis. CI95: 95% distribution from bootstrap analysis

Table A.5: Proportion of adaptive variants with 95% confidence intervals (CI) using standard McDonald-Kreitman Test by 100 bootstraps sets with replacement

	Wild			**Domestic**		
Scenario	**obs**	**CI05**	**CI95**	**obs**	**CI05**	**CI95**
s1	0.013	0.001	0.052	0.008	0.009	0.054
s2	-0.209	-0.218	-0.152	-0.196	-0.212	-0.156
s3	-0.198	-0.206	-0.148	-0.194	-0.198	-0.145
s4	-0.203	-0.199	-0.148	-0.209	-0.217	-0.172
s5	-0.174	-0.179	-0.133	-0.202	-0.208	-0.145
s6	0.030	0.022	0.067	-0.003	-0.004	0.045
s7	0.023	0.026	0.072	-0.372	-0.550	-0.206
s8	-0.187	-0.183	-0.135	-0.638	-0.850	-0.450
s9	-0.210	-0.212	-0.153	-0.380	-0.600	-0.263
s10	0.010	0.002	0.046	-0.647	-0.869	-0.338

CI05: 5% distribution from bootstrap analysis. CI95: 95% distribution from bootstrap analysis

Table A.6: Proportion of adaptive variants with 95% confidence intervals (CI) inferred by polyDFE

	Wild			Domestic		
Scenario	obs	CI05	CI95	obs	CI05	CI95
s1	0.165	0.122	0.203	0.167	0.136	0.207
s2	-0.008	-0.049	0.030	-0.008	-0.043	0.020
s3	0.031	-0.005	0.074	-0.011	-0.049	0.021
s4	-0.024	-0.053	0.014	-0.012	-0.053	0.021
s5	0.043	0.010	0.087	0.023	-0.023	0.065
s6	0.189	0.162	0.217	0.173	0.145	0.208
s7	0.176	0.149	0.205	-0.371	-0.396	-0.099
s8	0.021	-0.009	0.047	-0.313	-0.643	-0.056
s9	0.001	-0.035	0.046	-0.307	-0.524	-0.110
s10	0.147	0.118	0.181	-0.375	-0.399	-0.166

CI05: 5% distribution from bootstrap analysis. CI95: 95% distribution from bootstrap analysis

D. Estimation of the DFE and the detecting the differences in DFE between Wild and Domestic

PolyDFE

Table A.7: AIC weighted parameters and 95% confidence intervals

	Wild											
Scenario	**b**	**CI05**	**CI95**	**pb**	**CI05**	**CI95**	**Sb**	**CI05**	**CI95**	**Sd**	**CI05**	**CI95**	**Ne**
1	0.28	0.15	0.96	0.05	0.00	0.35	2.99	0.62	8.34	-5.44	-8.17	-4.17	335
2	0.17	0.15	0.22	0.00	0.00	0.05	0.21	0.06	2.75	-7.44	-9.05	-6.07	360
3	0.20	0.16	0.26	0.00	0.00	0.01	0.06	0.01	0.24	-6.42	-8.25	-4.91	358
4	0.20	0.15	0.96	0.03	0.00	0.37	0.16	0.05	0.67	-8.66	-11.79	-6.31	375
5	0.21	0.18	0.24	0.00	0.00	0.01	0.18	0.03	2.40	-6.26	-8.41	-5.25	345
6	0.37	0.22	1.28	0.07	0.02	0.24	5.22	1.41	12.21	-4.78	-6.32	-3.85	327
7	0.16	0.14	0.28	0.00	0.00	0.11	0.21	0.08	4.38	-6.76	-9.27	-5.17	340
8	0.20	0.16	0.64	0.01	0.00	0.26	0.22	0.04	1.92	-6.70	-8.30	-5.42	358
9	0.19	0.16	0.22	0.00	0.00	0.02	0.10	0.03	2.43	-6.79	-8.32	-5.53	349
10	0.18	0.13	1.32	0.02	0.00	0.48	5.14	0.24	18.62	-7.35	-9.75	-5.52	328
	Domestic												
Scenario	**b**	**CI05**	**CI95**	**pb**	**CI05**	**CI95**	**Sb**	**CI05**	**CI95**	**Sd**	**CI05**	**CI95**	**Ne**
1	0.26	0.16	0.78	0.05	0.00	0.33	3.39	0.53	9.61	-5.49	-7.90	-4.06	373
2	0.17	0.15	0.22	0.00	0.00	0.04	0.18	0.04	2.20	-7.92	-10.44	-6.43	479
3	0.16	0.13	0.18	0.00	0.00	0.01	0.04	0.01	0.20	-10.00	-13.64	-8.10	535
4	0.22	0.16	0.76	0.01	0.00	0.33	0.13	0.03	0.30	-6.76	-9.87	-4.76	327
5	0.24	0.20	0.34	0.00	0.00	0.02	0.17	0.03	2.06	-4.46	-5.41	-3.46	313
6	0.58	0.26	2.78	0.18	0.01	0.43	2.06	0.22	7.08	-3.94	-5.47	-2.88	292
7	1.46	0.55	3.59	0.00	0.00	0.17	0.57	0.09	10.12	-0.04	-0.87	0.00	54
8	2.03	0.66	4.96	0.01	0.00	0.21	0.35	0.07	2.76	-0.52	-1.09	-0.02	56
9	0.24	0.10	1.53	0.01	0.00	0.09	0.25	0.07	3.05	-3.15	-27.00	-0.53	57
10	1.96	0.77	4.66	0.06	0.00	0.58	4.12	0.33	30.58	-0.02	-0.45	-0.01	55

E. Demography and Joint DFE Inference Using dadi

Table A.8: Estimated demographic parameters

Scenario	N_a	N_{ew}	N_{e1d}	N_{e2d}	m	t_1	t_{bot}	t_3
1	301.4	0.99	0.08	1.39	94.6	8.62	0.00	0.14
2	671.3	0.50	0.28	3.24	15.0	4.71	0.12	0.01
3	490.4	0.69	0.11	1.88	6.73	17.9	0.24	0.04
4	268.7	1.27	0.07	1.32	0.00	6.24	0.01	0.29
5	389.3	0.86	0.08	0.89	0.00	9.81	0.03	0.18
6	470.1	0.62	0.10	0.69	0.00	6.26	0.05	0.13
7	511.5	0.65	0.01	2.30	0.00	13.5	0.17	0.00
8	405.3	0.83	0.01	3.37	0.00	12.9	0.21	0.00
9	572.8	0.58	0.01	1.85	0.00	16.5	0.14	0.00
10	463.1	0.64	0.01	3.55	0.00	7.60	0.17	0.01

S: Scenarios. Expected N_{ew} = 1. Expected N_{e1d} = 0.1, Expected N_{e2d} = 1, Expected m = 200 for scenarios 1-3, and m=0 for scenarios 4-10. Expected t_1 = 10**,** Expected t_{bot} = 2Ne*0.016 for scenarios 1-6 and t_{bot} = 2Ne*0.161 for scenarios 7-10, Expected t_3 =2Ne* 0.15 for scenarios 1-6 and t_3 = 2Ne*0.005 for scenarios 7-10.

Table A.9: Confidence Intervals of estimated demographic parameters in relation to ancestral N_e

	95%CI (N_{ew})	95%CI (N_{e1d})	95%CI (N_{e2d})	95%CI (m)	95%CI (t_1)	95%CI (t_{bot})	95%CI (t_3)
1	[0.924,1.052]	[0.054,0.098]	[0.669,2.125]	[67.97,121.2]	[4.444,12.789]	[0.001,0.002]	[0.139,0.140]
2	[0.412,0.587]	[0.228,0.331]	[1.000,6.000]	[12.67,17.32]	[3.252, 6.176]	[0.080,0.148]	[0.001,0.001]
3	[0.673,0.727]	[0.099,0.112]	[1.638,2.115]	[6.619,6.834]	[17.248,18.61]	[0.206,0.268]	[0.038,0.039]
4	[0.808,2.025]	[0.057,0.081]	[0.529,2.116]	[0.000,0.040]	[1.000,14.402]	[0.013,0.016]	[0.122,0.456]
5	[0.802,0.909]	[0.030,0.119]	[0.822,0.951]	[0.000,0.023]	[8.617,10.996]	[0.0116,0.05]	[0.169,0.200]
6	[0.519,0.725]	[0.063,0.137]	[0.571,0.809]	[0.000,0.039]	[4.439,8.0820]	[0.027,0.067]	[0.108,0.155]
7	[0.623,0.665]	[0.009,0.011]	[0.000,3.933]	[0.000,0.004]	[0.000,18.339]	[0.127,0.188]	[0.004,0.007]
8	[0.034,1.621]	[0.000,0.019]	[0.000,7.691]	[0.000,0.008]	[7.310,18.608]	[0.009,0.412]	[0.005,0.013]
9	[0.564,0.606]	[0.008,0.012]	[0.787,2.929]	[0.000,0.002]	[3.752,20.281]	[0.132,0.154]	[0.003,0.005]
10	[0.428,0.740]	[0.001,0.019]	[1.178,3.658]	[0.000,0.004]	[6.876,11.158]	[0.134,0.172]	[0.000,0.010]

S: Scenarios. Expected N_{ew} = 1. Expected N_{e1d} = 0.1, Expected N_{e2d} = 1, Expected m = 200 for scenarios 1-3, and m=0 for scenarios 4-10. Expected t_1 = 10**,** Expected t_{bot} = 2Ne*0.016 for scenarios 1-6 and t_{bot} = 2Ne*0.161 for scenarios 7-10, Expected t_3 = 2Ne*0.15 for scenarios 1-6 and t_3 = 2Ne*0.005 for scenarios 7-10.

Table A.10: Estimated selective parameters

	Negative DFE parameters			Positive DFE parameters			
S	**shape**	**scale**	**s_d**	**p_{+w}**	**p_c**	**p_{c+}**	$p_c * p_{c+}$.
1	0.464	4.884	0.015	0.031	0.000	0.000	0.000
2	0.828	8.257	0.020	0.281	0.202	0.587	0.118
3	4.116	0.882	0.015	0.292	0.235	0.000	0.000
4	0.409	6.110	0.019	0.095	0.001	0.000	0.000
5	0.380	9.981	0.019	0.103	0.198	0.124	0.024
6	0.462	7.723	0.015	0.036	0.157	0.020	0.003
7	0.282	26.99	0.030	0.017	0.000	0.000	0.000
8	0.553	6.498	0.018	0.163	0.034	0.000	0.000
9	0.364	16.27	0.021	0.138	0.000	0.000	0.000
10	0.408	8.313	0.014	0.029	0.000	0.000	0.000

S: Scenarios.

The real shape of the negative gamma distribution is equal in all scenarios and populations and is fixed to 0.2. The real negative scale of the gamma distribution is equal for all simulated scenarios and populations and is fixed to 10. The real selection coefficient s_d is 0.01 in the homozygote. Let p_{+w} be the fraction of mutations that are positively selected in the Wild population, p_c be the fraction of mutations that change selection coefficient in the Domestic population, and p_{c+} be the fraction of those mutations that become beneficial in the Domestic population.

Table A.11: Confidence Intervals for estimated selective parameters

	Negative DFE parameters			Positive DFE parameters		
S	**95%CI shape**	**95%CI scale**	**95%CI s_d**	**95%CI p_{+w}**	**95%CI p_c**	**95%CI $p_c * p_{c+}$.**
1	[0.378,0.587]	[3.534,6.831]	[0.013,0.017]	[0.025,0.037]	[0.000, 0.118]	[0.000,0.000]
2	[0.481,1.668]	[3.804,14.79]	[0.018,0.023]	[0.199,0.348]	[0.025,0.385]	[0.000,0.251]
3	[1.634,8.548]	[0.438,2.233]	[0.014,0.016]	[0.243,0.321]	[0.169,0.309]	[0.0,0.00008]
4	[0.329,0.517]	[4.552,8.077]	[0.017,0.020]	[0.066,0.119]	[0.000,0.078]	[0.00,0.0098]
5	[0.277,0.634]	[5.487,14.96]	[0.017,0.022]	[0.047,0.161]	[0.082,0.365]	[0.00,0.1055]
6	[0.397,0.536]	[6.158,9.690]	[0.014,0.017]	[0.030,0.042]	[0.071,0.284]	[0.000,0.224]
7	[0.242,0.325]	[21.36,36.22]	[0.026,0.034]	[0.011,0.021]	[0.000,0.053]	[0.000,0.000]
8	[0.336,0.899]	[3.943,11.73]	[0.016,0.021]	[0.093,0.224]	[0.000,0.731]	[0.00,0.0174]
9	[0.190,0.888]	[6.061,34.09]	[0.018,0.025]	[0.000,0.265]	[0.000,0.362]	[0.000,0.002]
10	[0.354,0.481]	[6.791,10.55]	[0.013,0.016]	[0.023,0.036]	[0.000,0.602]	[0.000,0.217]

S: Scenarios.

The real shape of the negative gamma distribution is equal in all scenarios and populations and is fixed to 0.2. The real negative scale of the gamma distribution is equal for all simulated scenarios and populations and is fixed to 10. The real selection coefficient s_d is 0.01 in the homozygote. Let p_{+w} be the fraction of mutations that are positively selected in the Wild population, p_c be the fraction of mutations that change selection coefficient in the Domestic population, and p_{c+} be the fraction of those mutations that become beneficial in the Domestic population.

F. Description of the types of mutations and allele frequency distributions across populations and scenarios

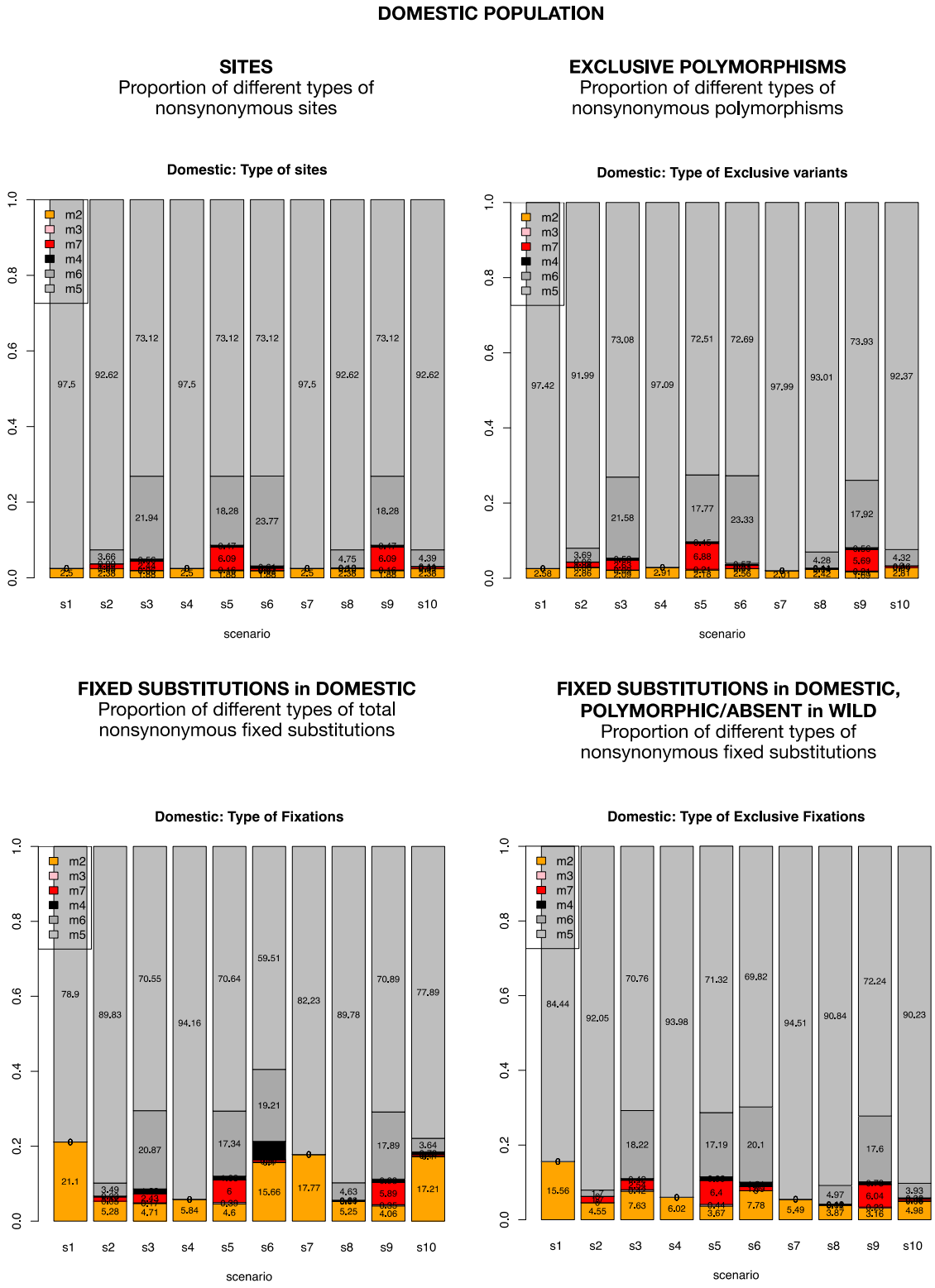

Figure A.0: Proportion of the different types of nonsynonymous sites, polymorphisms and fixed variants at the Domestic population. The y axis is shown in log10 scale. The x-axis indicates in each of the ten scenarios.

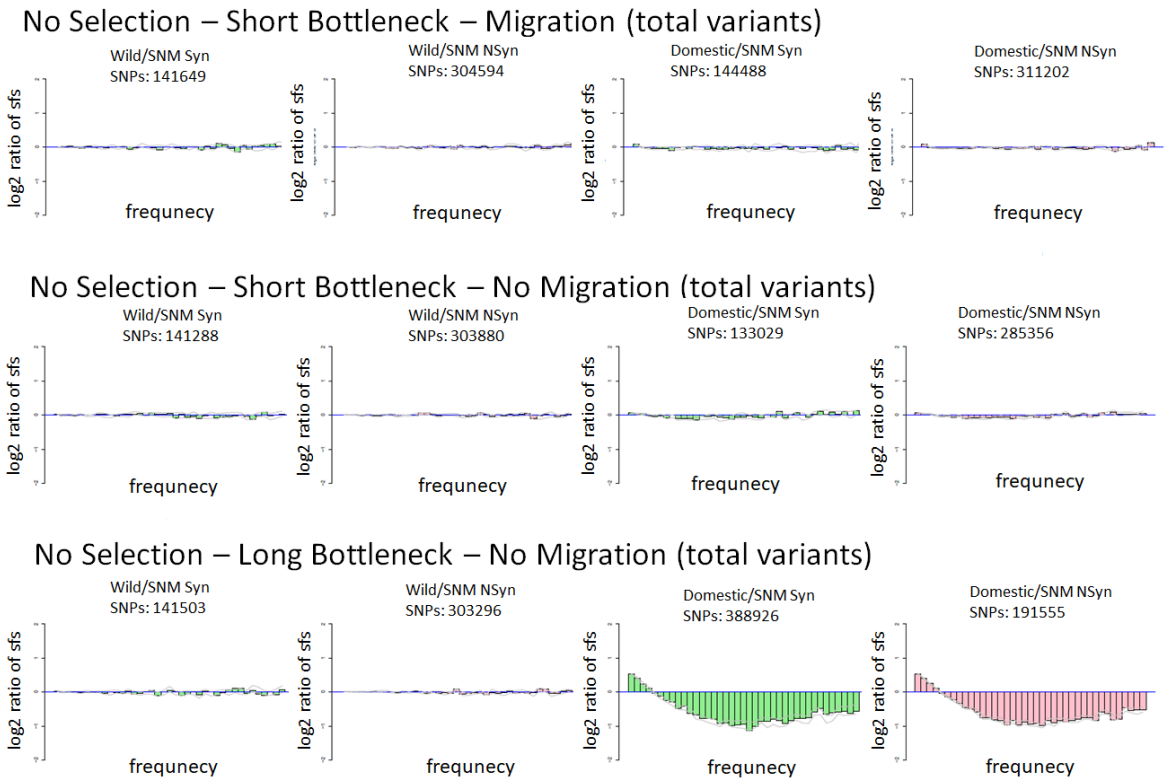

Figure A.1: log2 ratio of the expected SFS under Standard Neutral Models vs the simulated one under three scenarios: Short bottleneck with migration (first row), short bottleneck without migration (second row) and long bottleneck without migration (third row).

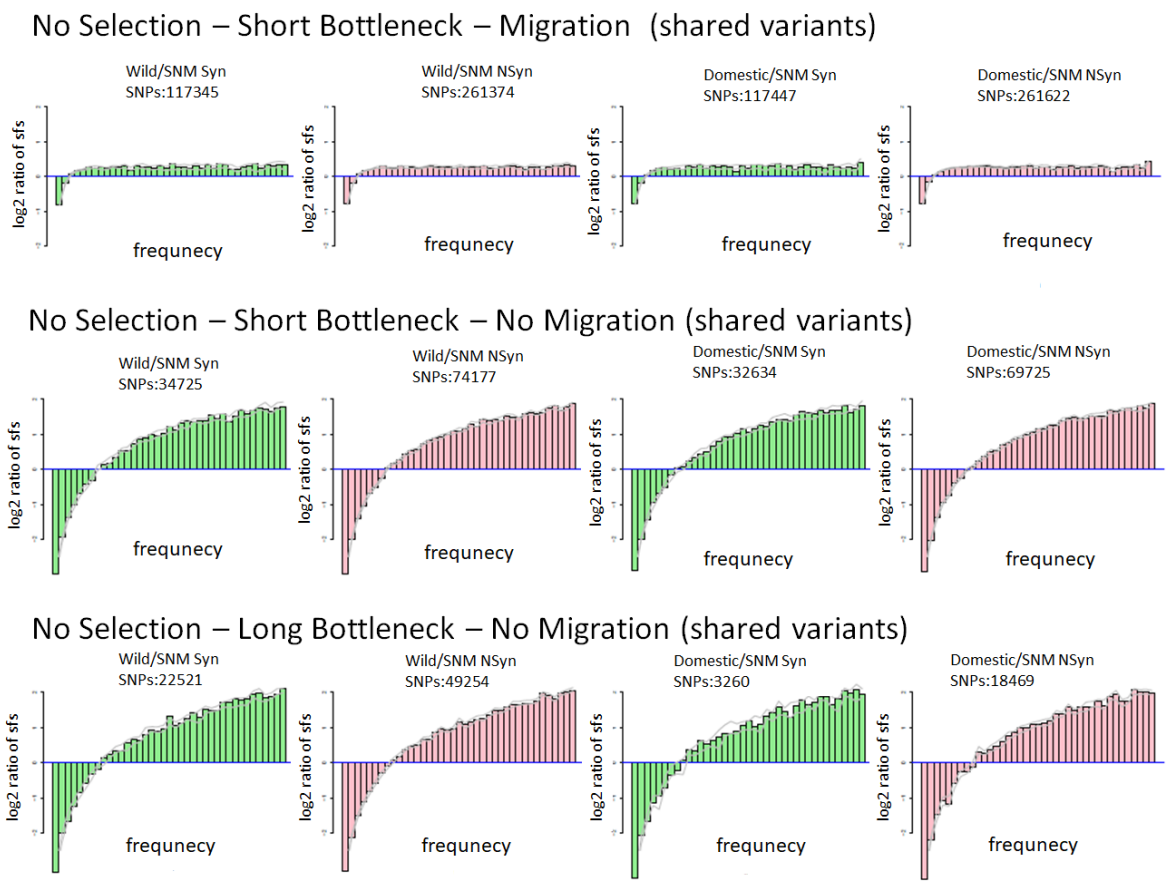

Figure A.2: log2 ratio of the expected SFS under Standard Neutral Models vs the simulated one using shared variants under three scenarios: Short bottleneck with migration (first row), short bottleneck without migration (second row) and long bottleneck without migration (third row).

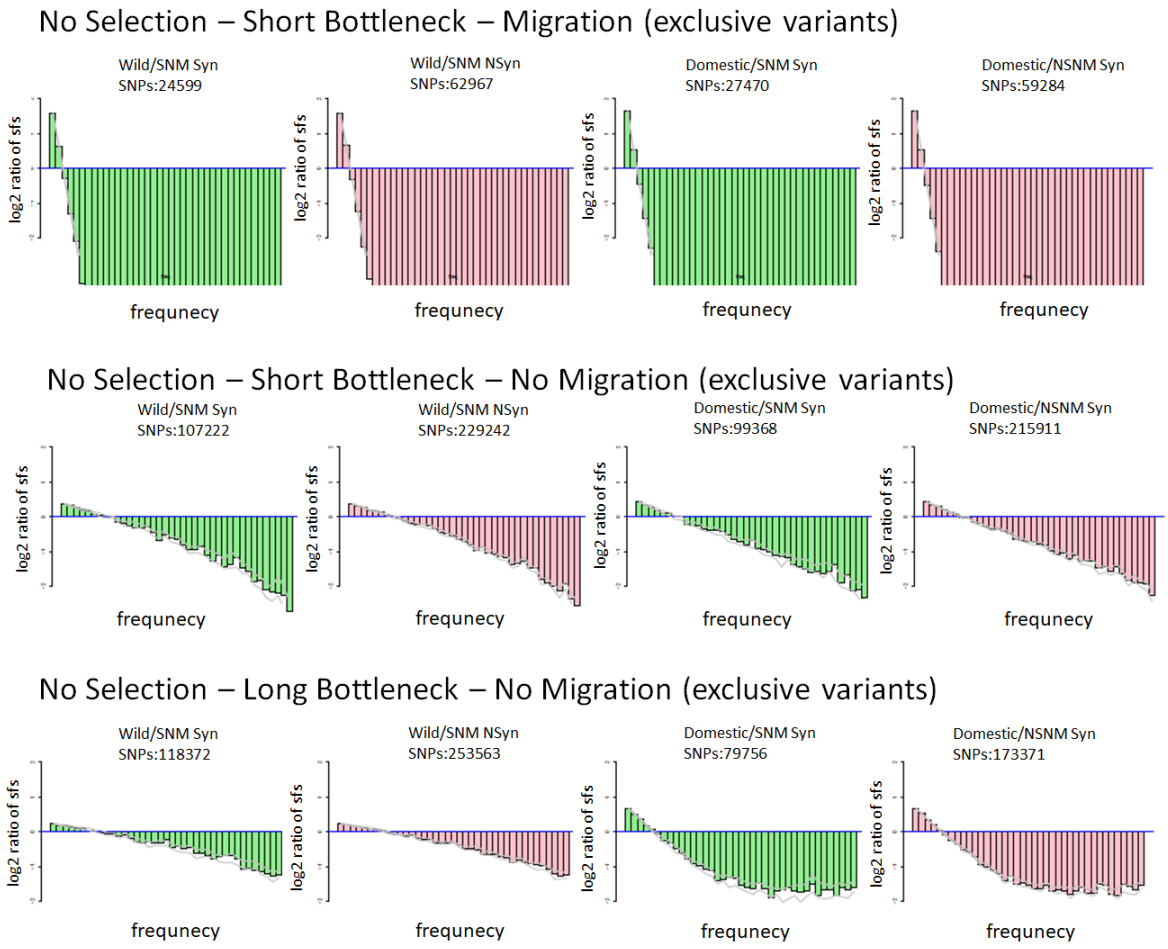

Figure A.3: log2 ratio of the expected SFS under Standard Neutral Models vs the simulated one using exclusive variants under three scenarios: Short bottleneck with migration (first row), short bottleneck without migration (second row) and long bottleneck without migration (third row).

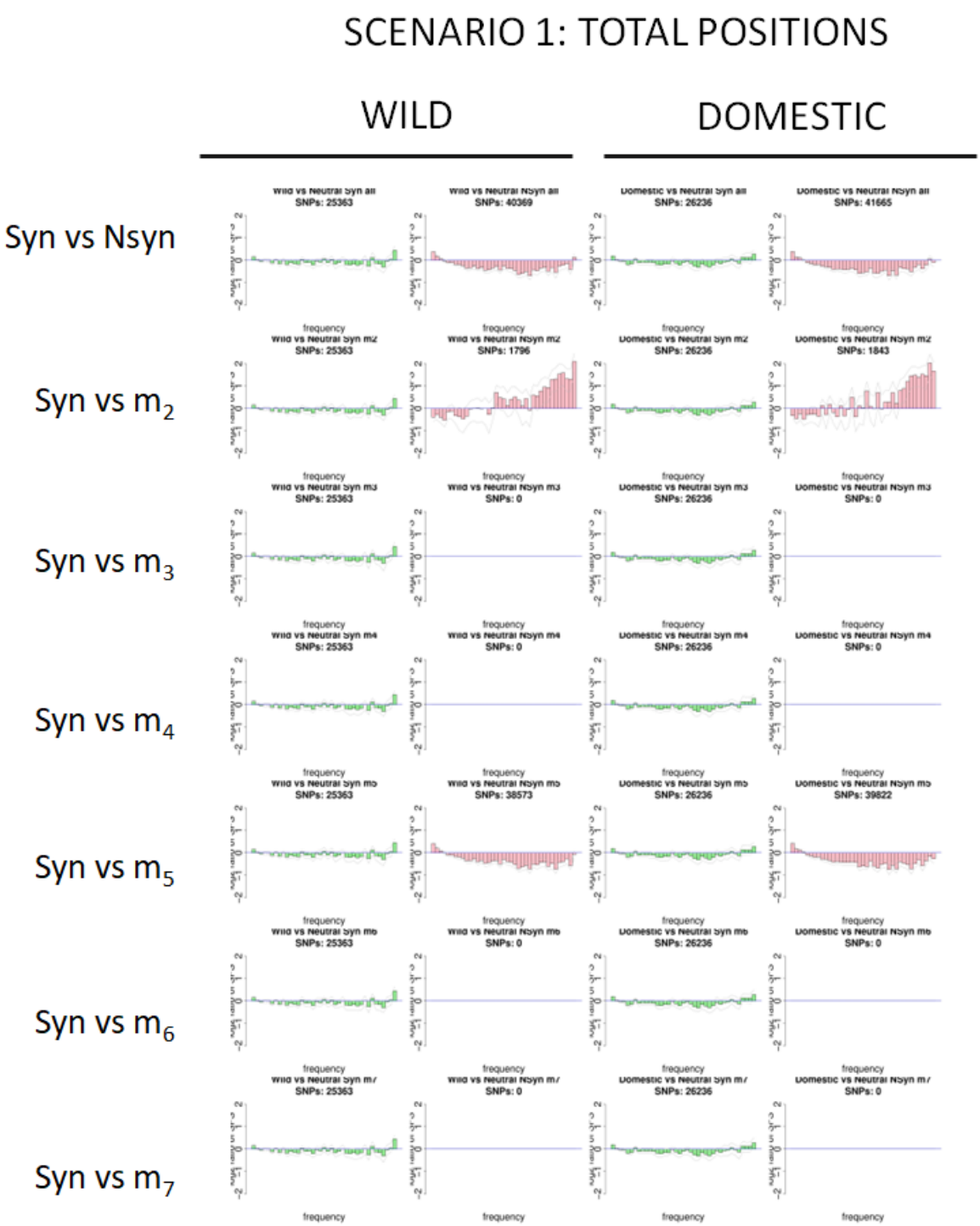

Figure A.4: log2 ratio of the simulated SFS for scenario 1 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

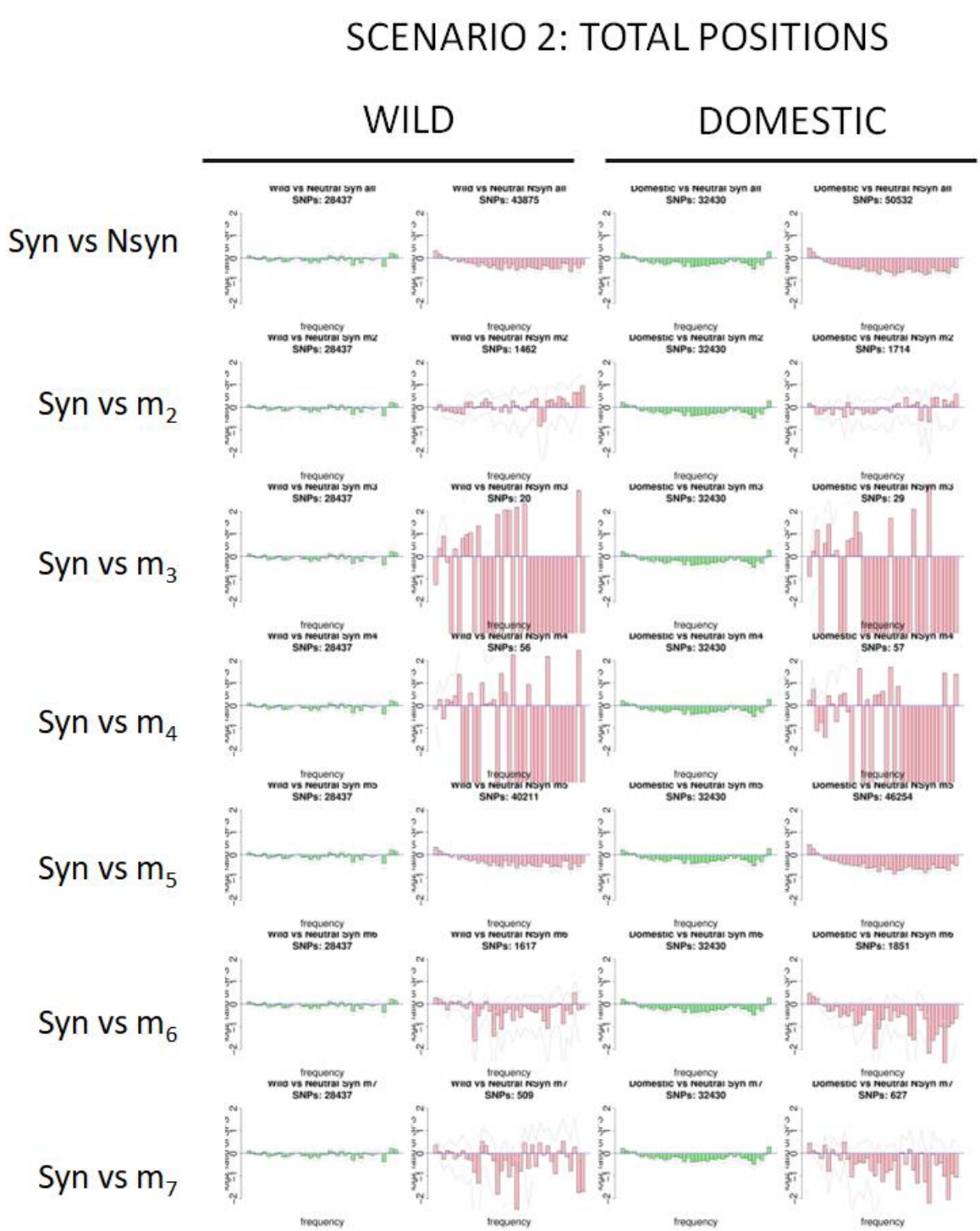

Figure A.5: log2 ratio of the simulated SFS for scenario 2 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

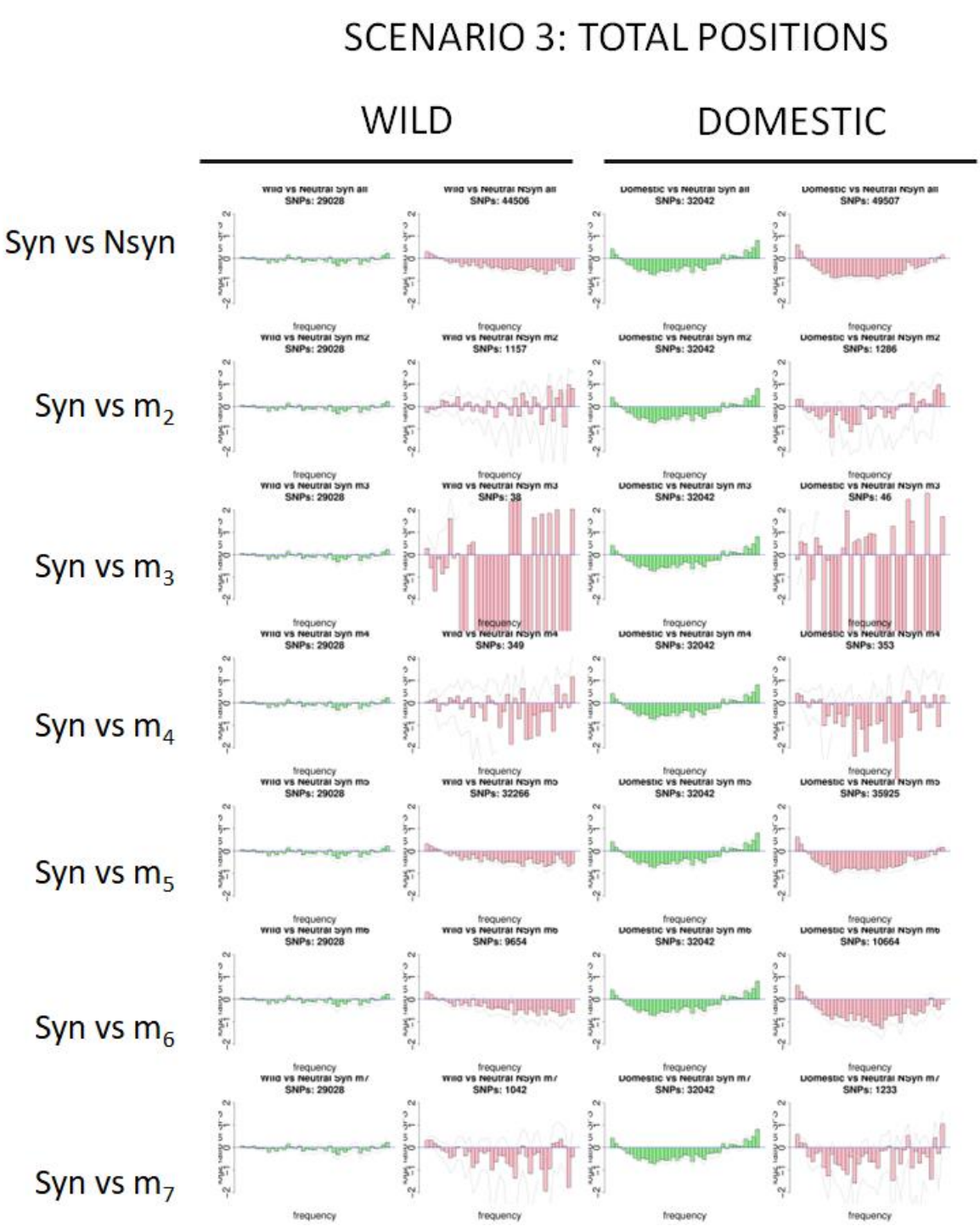

Figure A.6: log2 ratio of the simulated SFS for scenario 3 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

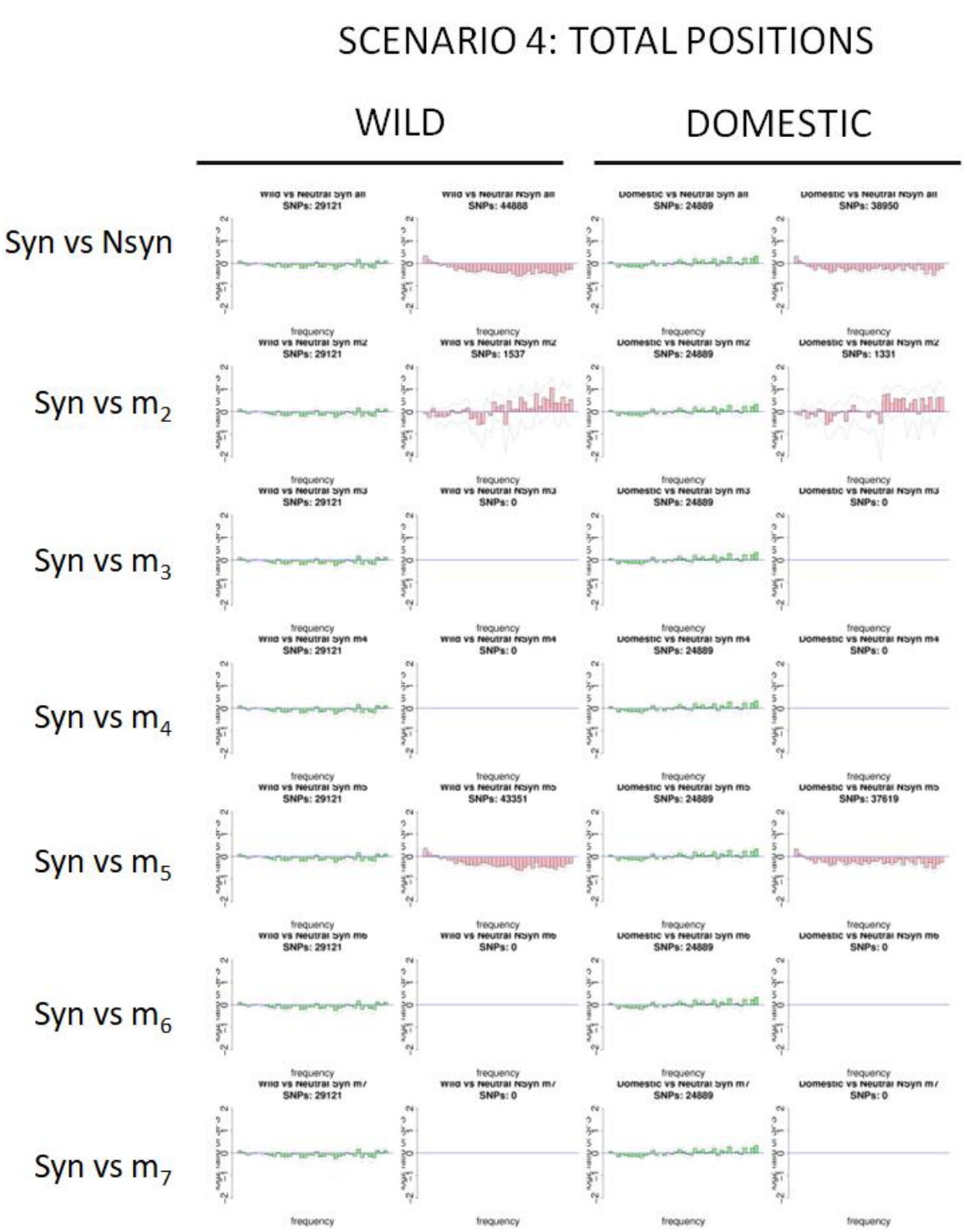

Figure A.7: log2 ratio of the simulated SFS for scenario 4 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

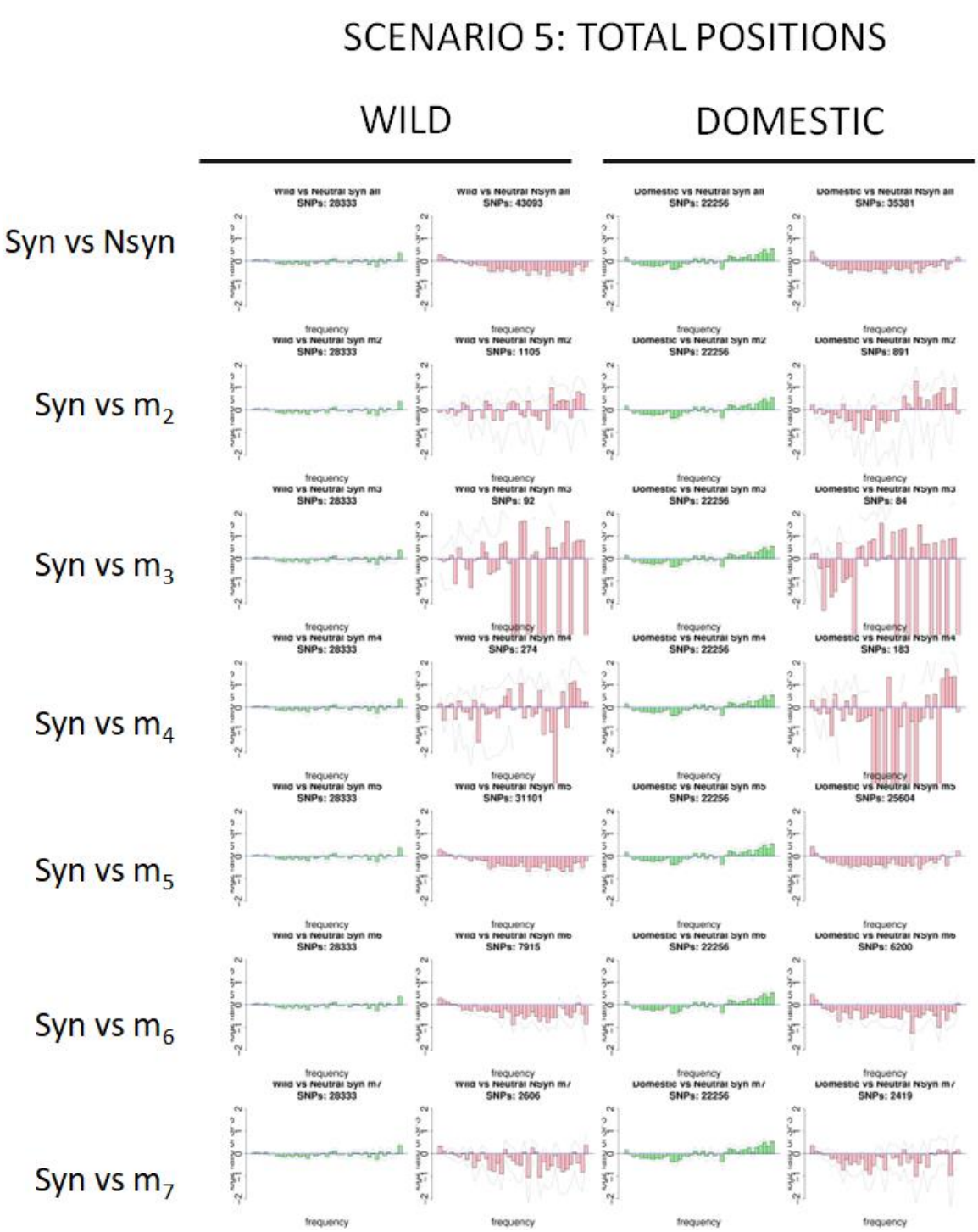

Figure A.8: log2 ratio of the simulated SFS for scenario 5 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

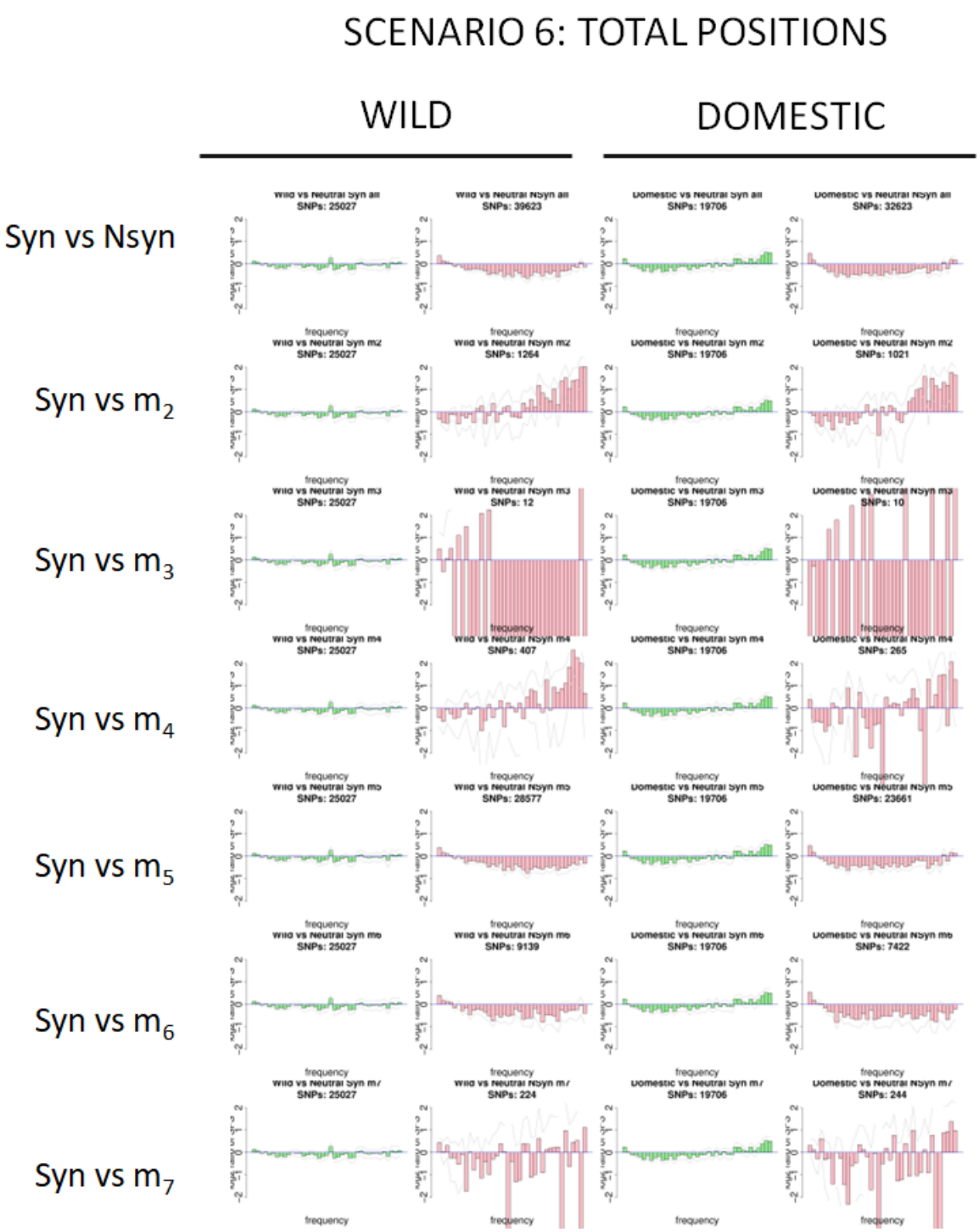

Figure A.9: log2 ratio of the simulated SFS for scenario 6 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

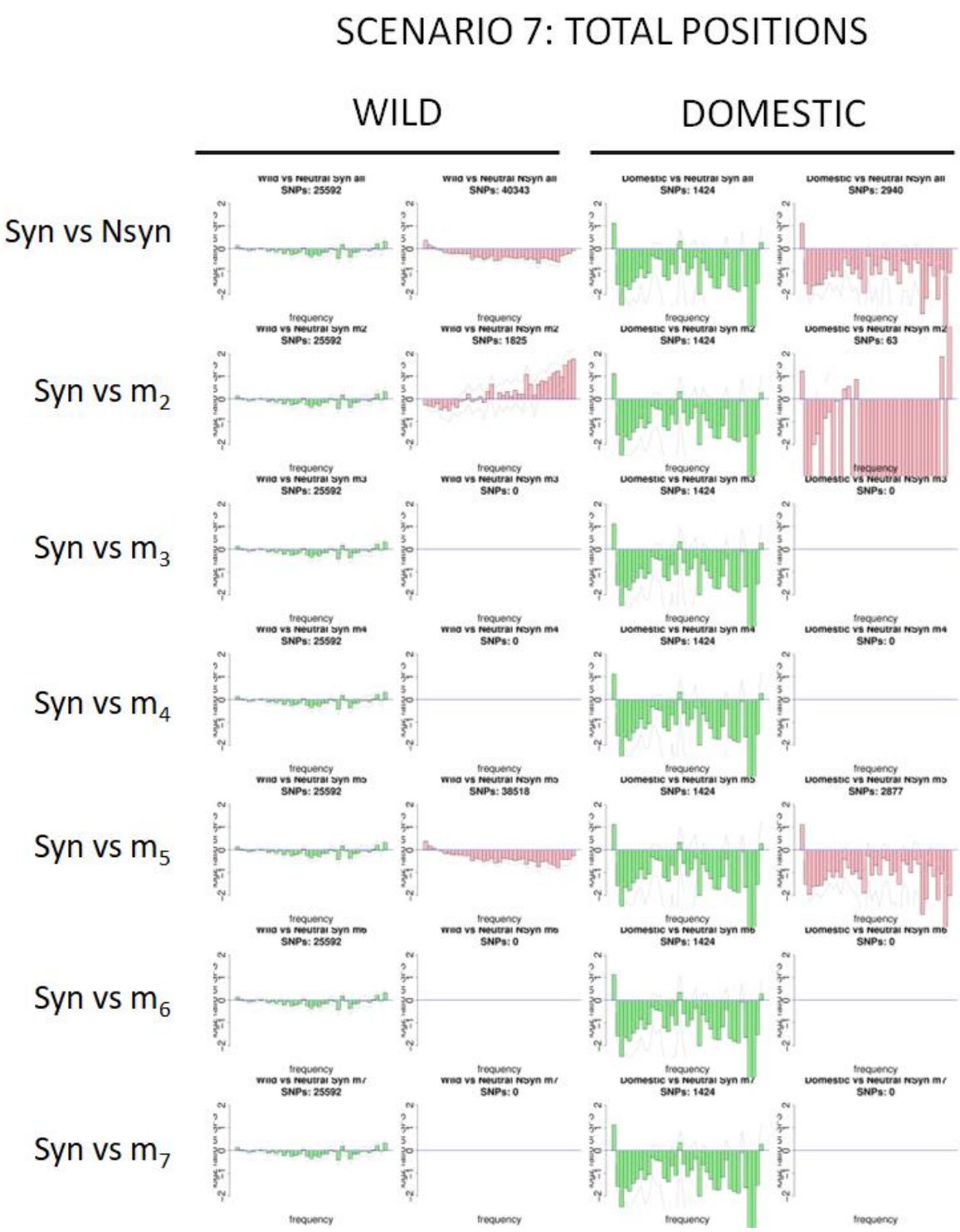

Figure A.10: log2 ratio of the simulated SFS for scenario 7 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

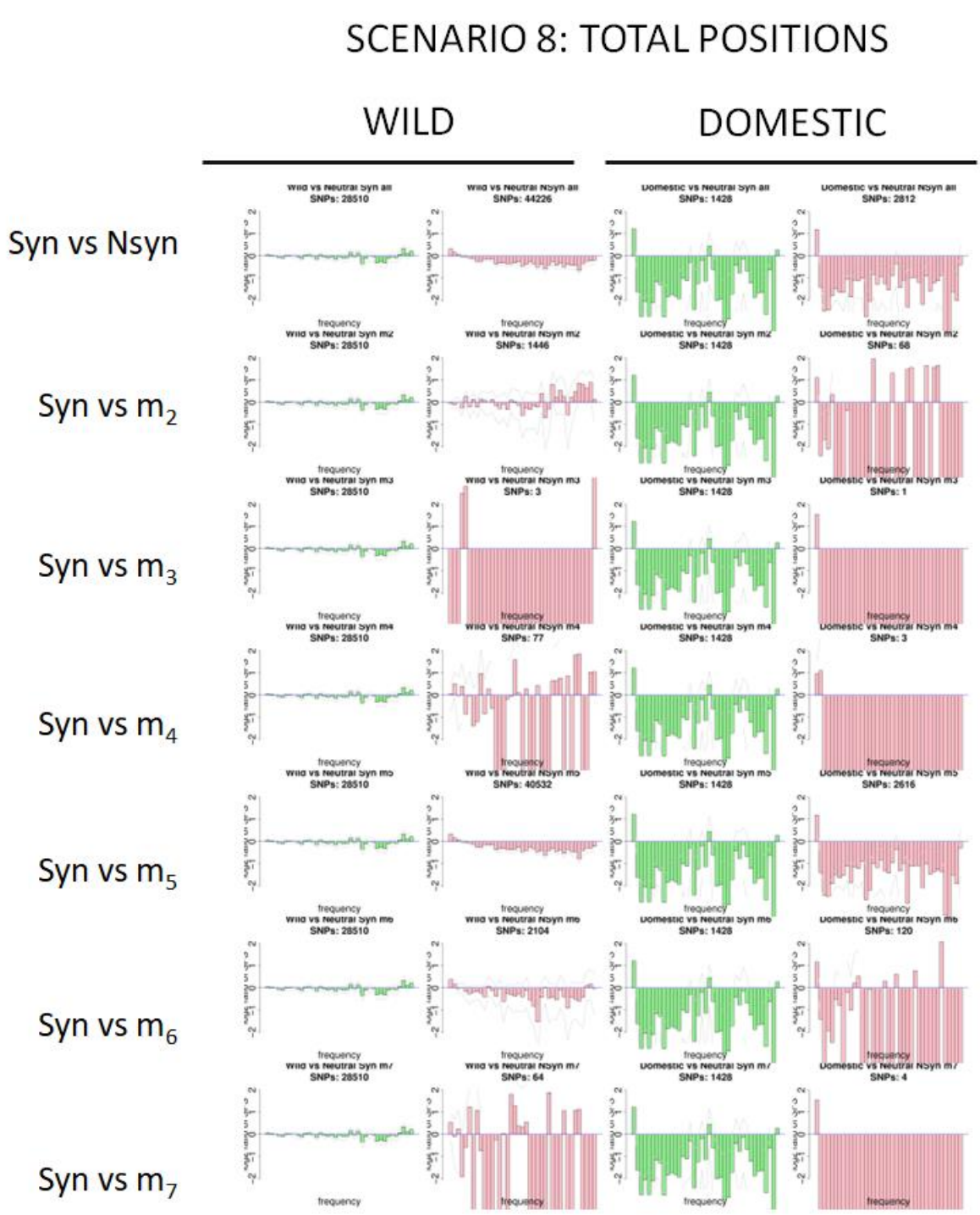

Figure A.11: log2 ratio of the simulated SFS for scenario 8 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

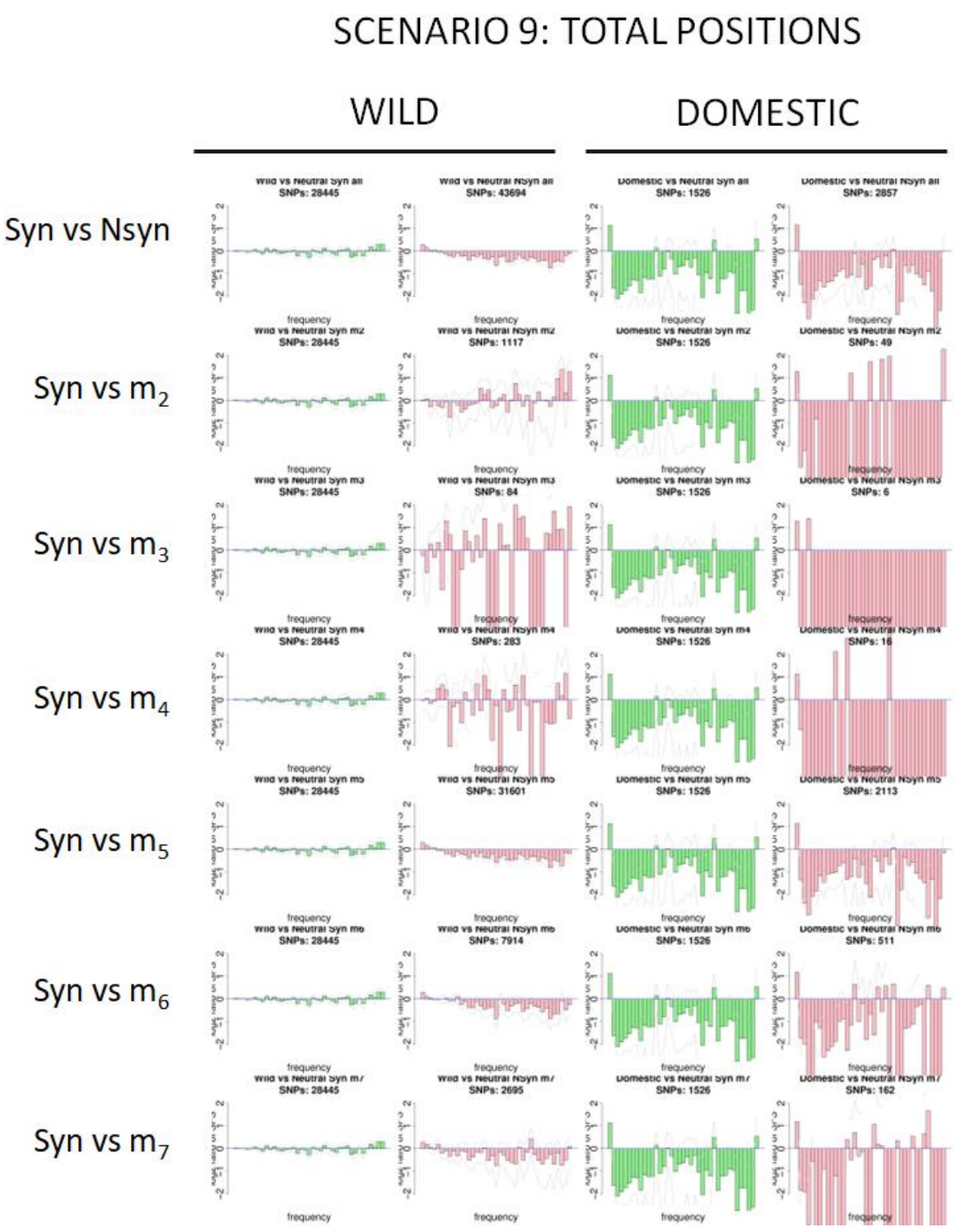

Figure A.12: log2 ratio of the simulated SFS for scenario 9 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively.

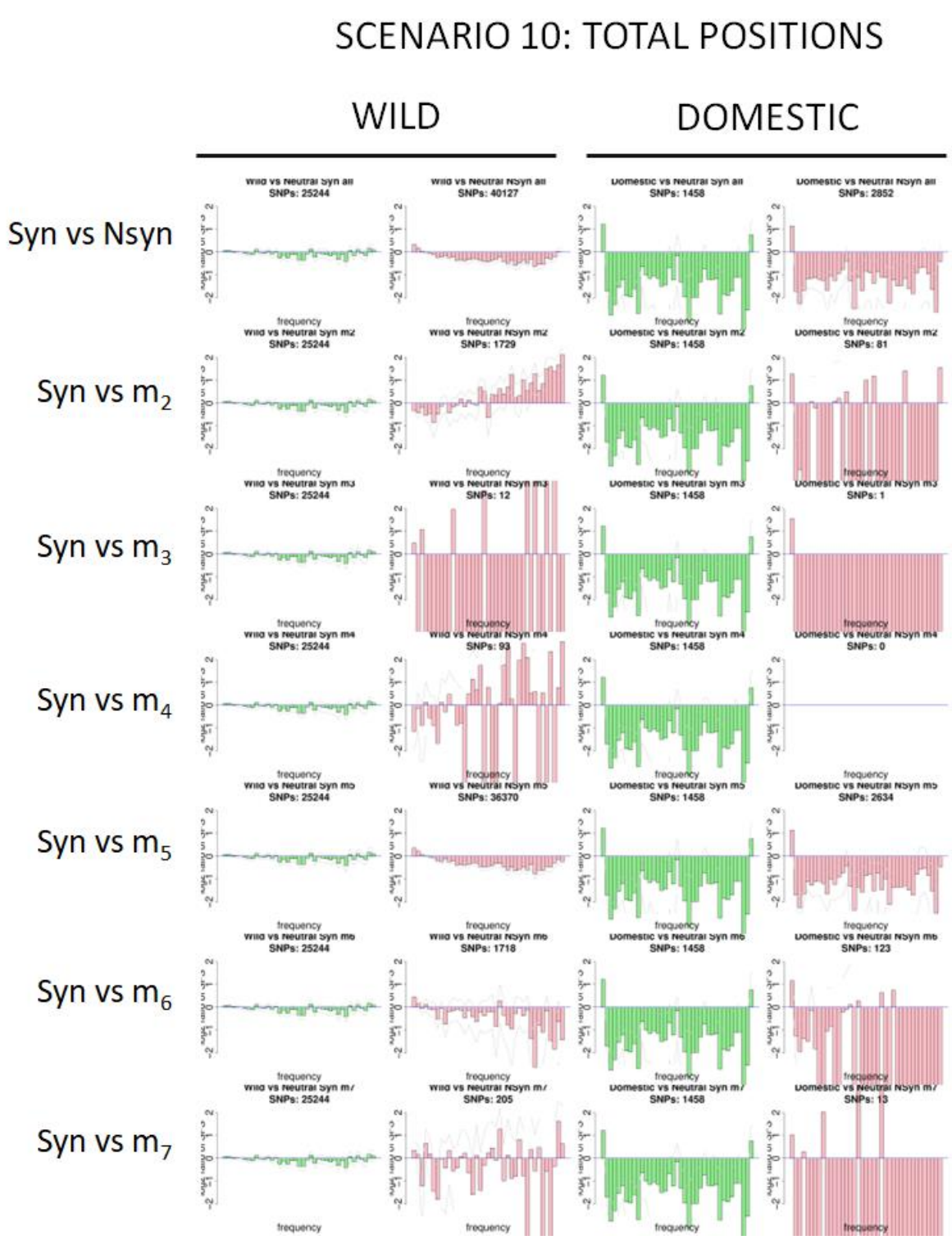

Figure A.13: log2 ratio of the simulated SFS for scenario 10 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 variants, respectively

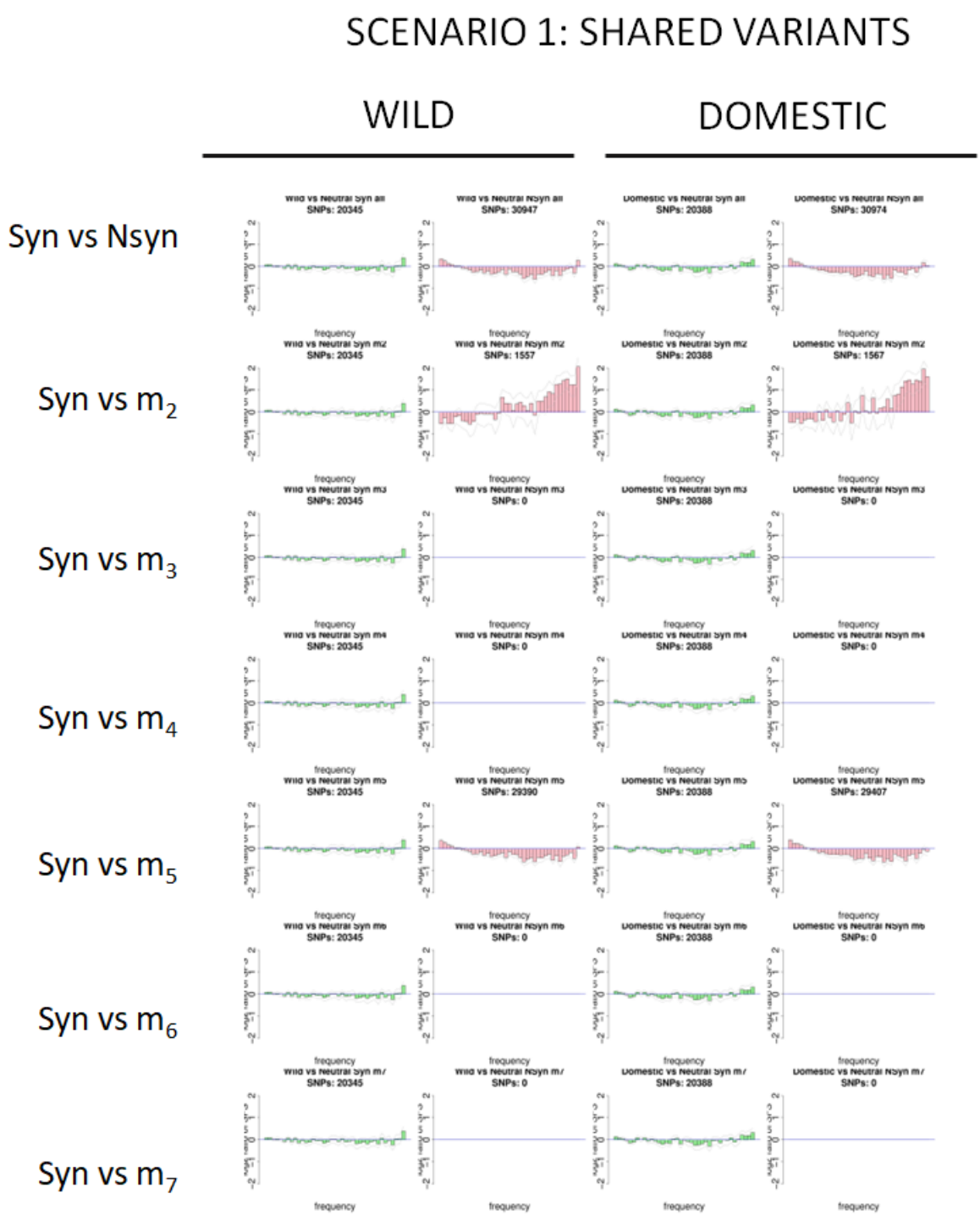

Figure A.14: log2 ratio of the simulated SFS for scenario 1 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

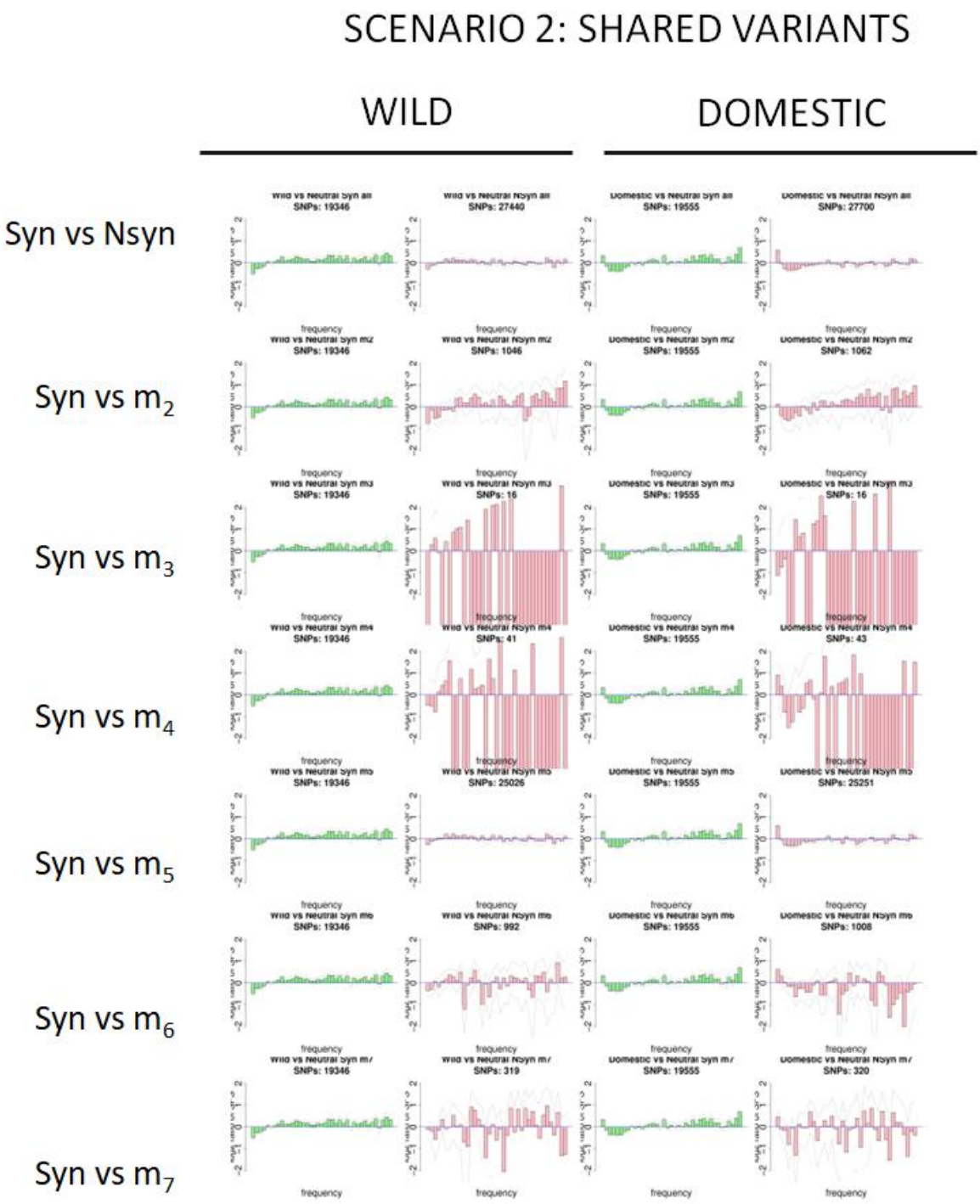

Figure A.15: log2 ratio of the simulated SFS for scenario 2 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

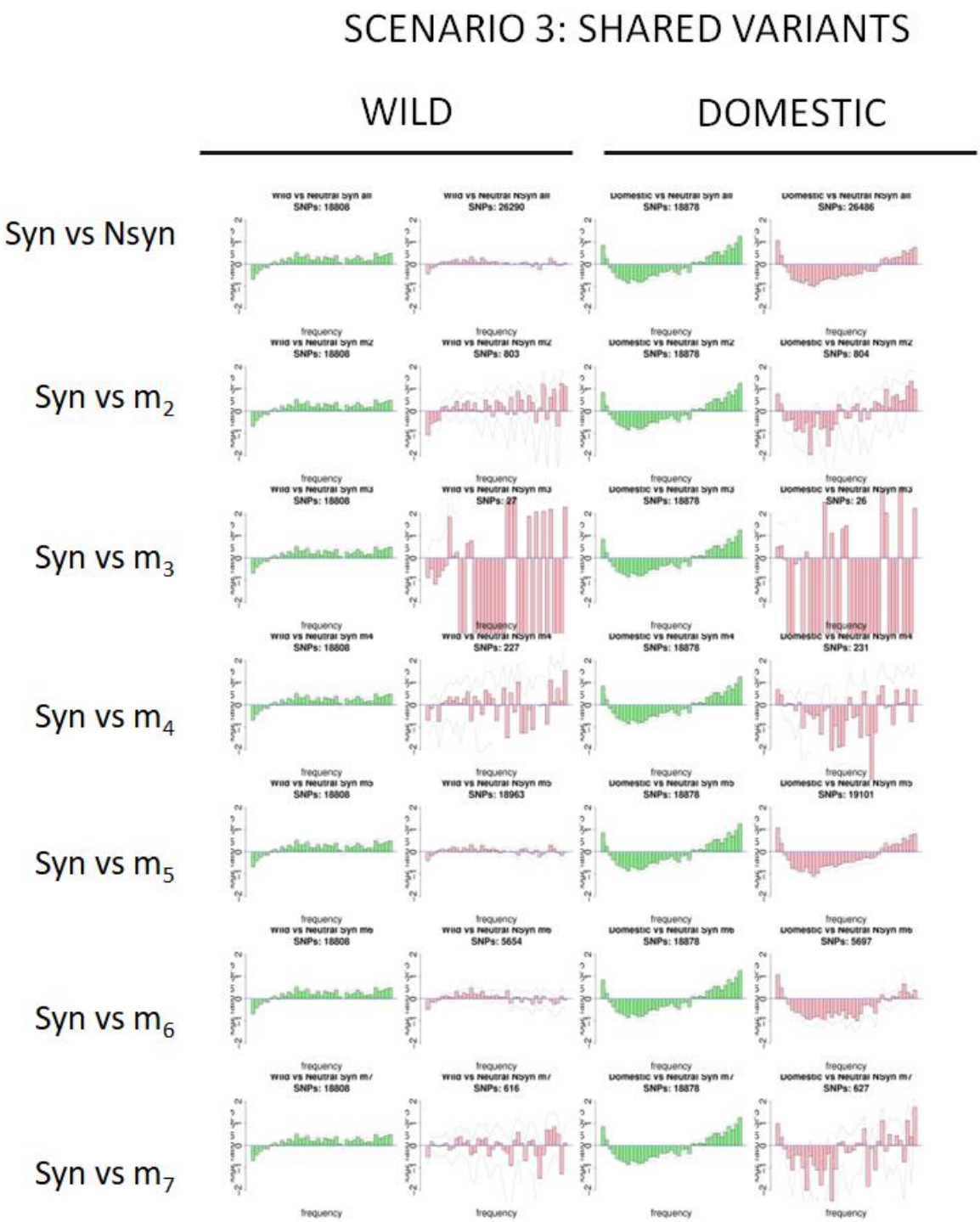

Figure A.16: log2 ratio of the simulated SFS for scenario 3 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

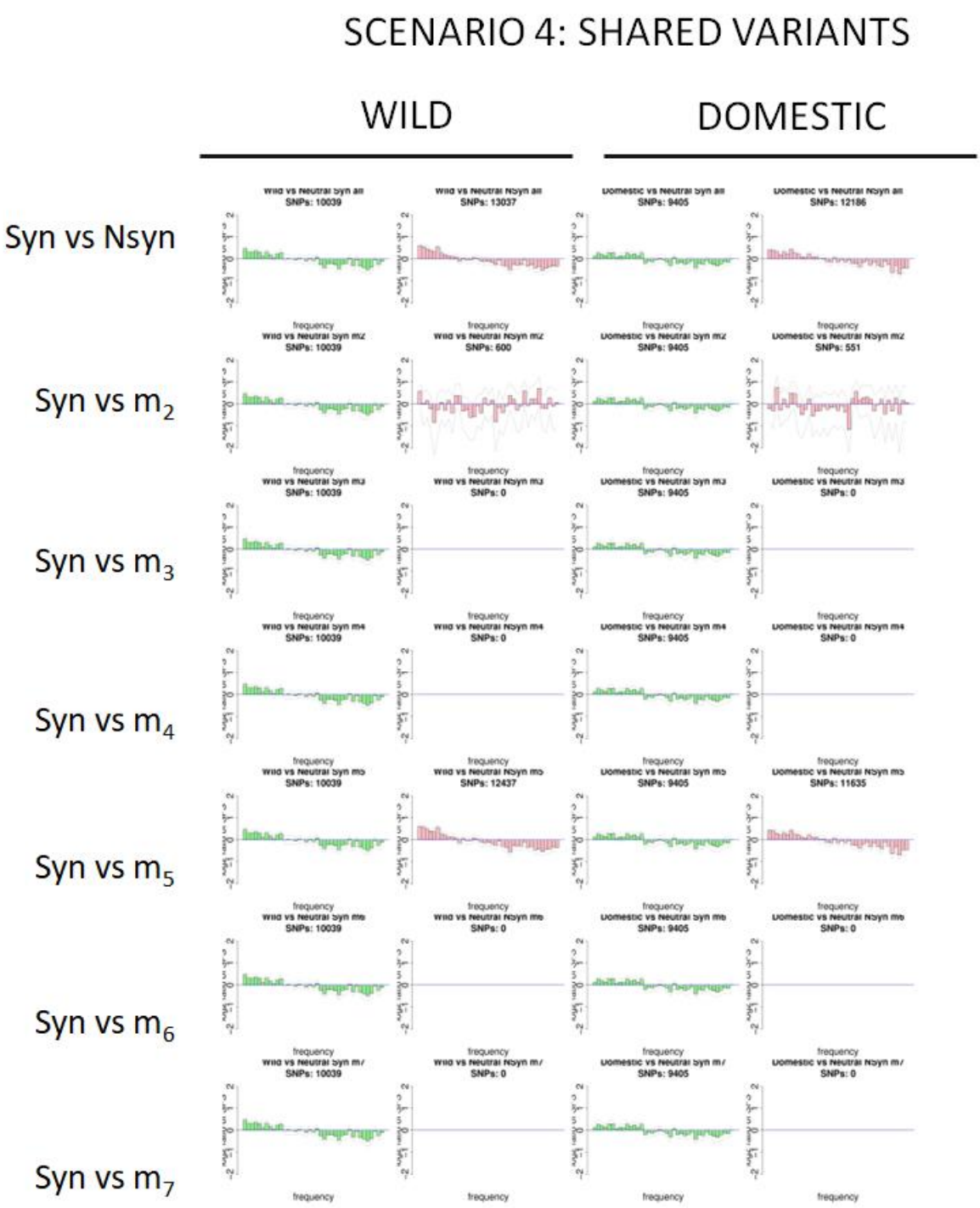

Figure A.17: log2 ratio of the simulated SFS for scenario 4 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

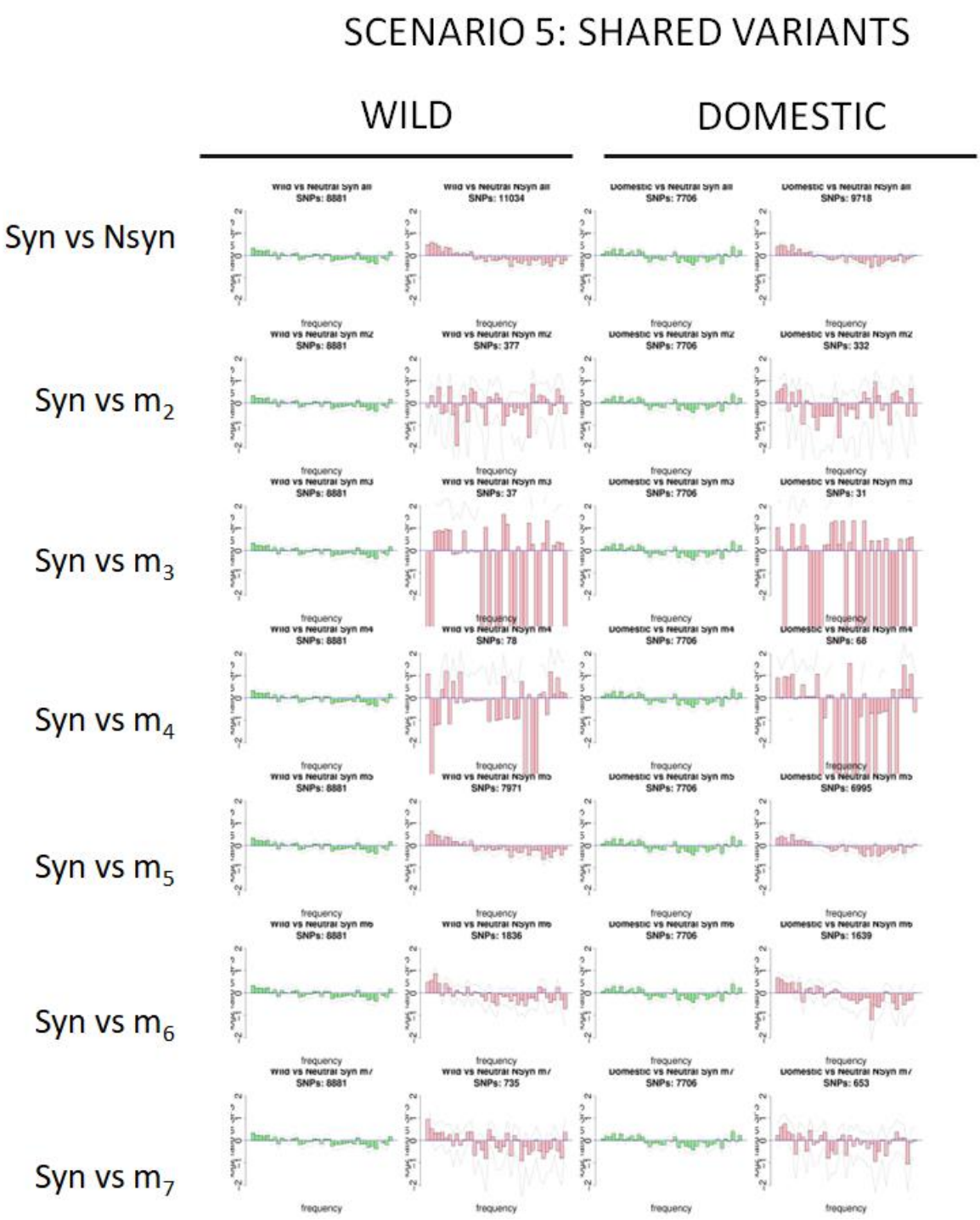

Figure A.18: log2 ratio of the simulated SFS for scenario 5 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

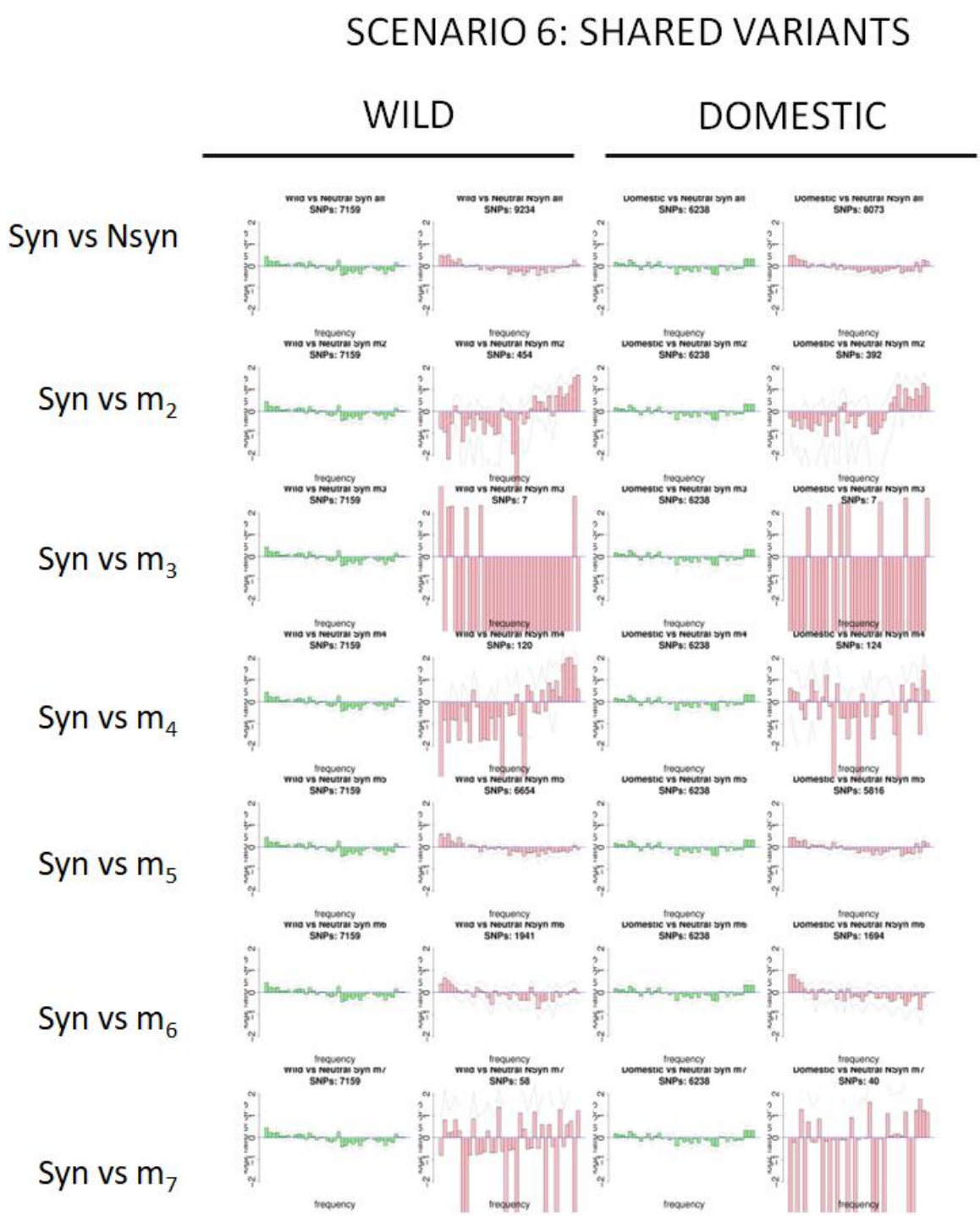

Figure A.19: log2 ratio of the simulated SFS for scenario 6 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

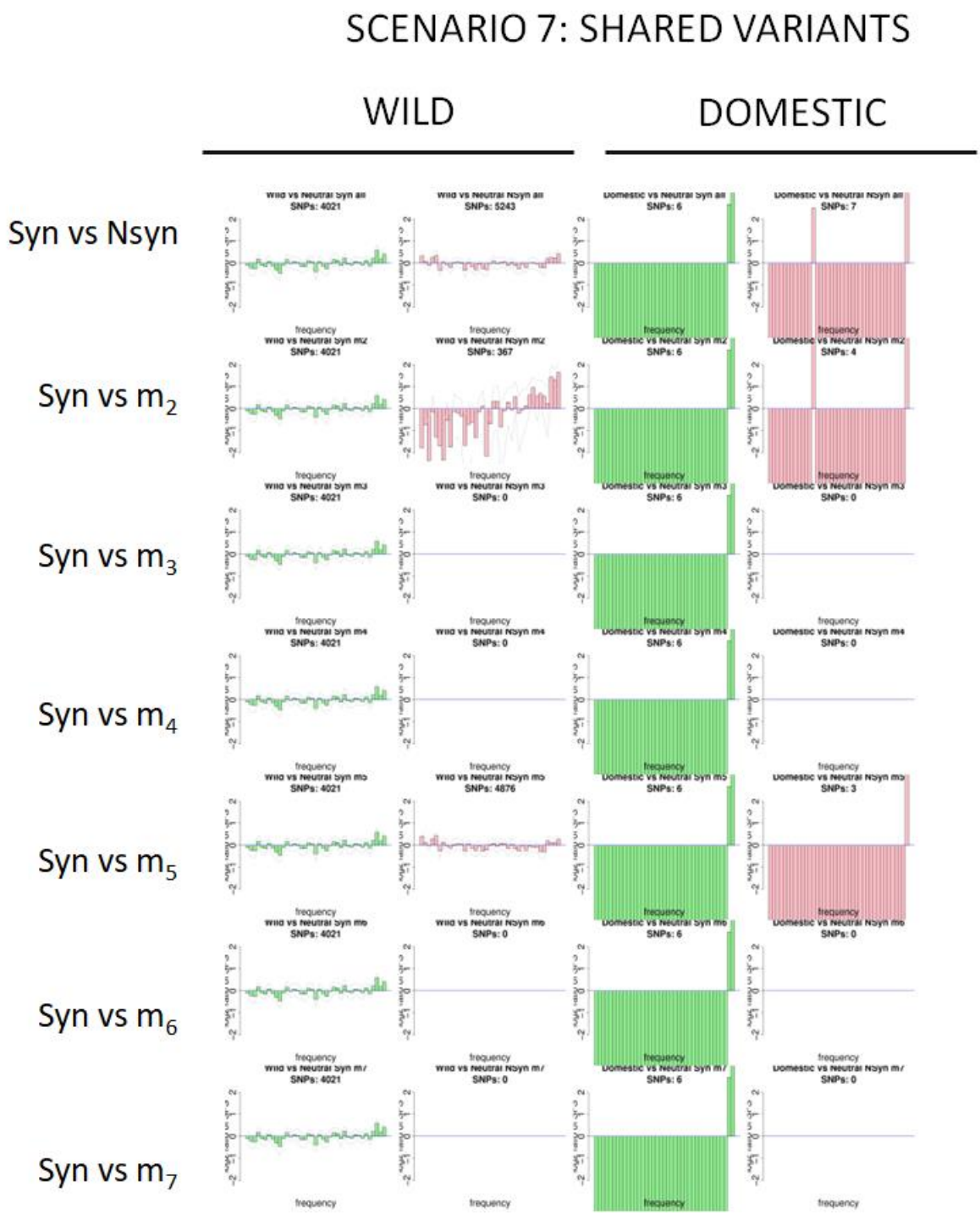

Figure A.20: log2 ratio of the simulated SFS for scenario 7 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

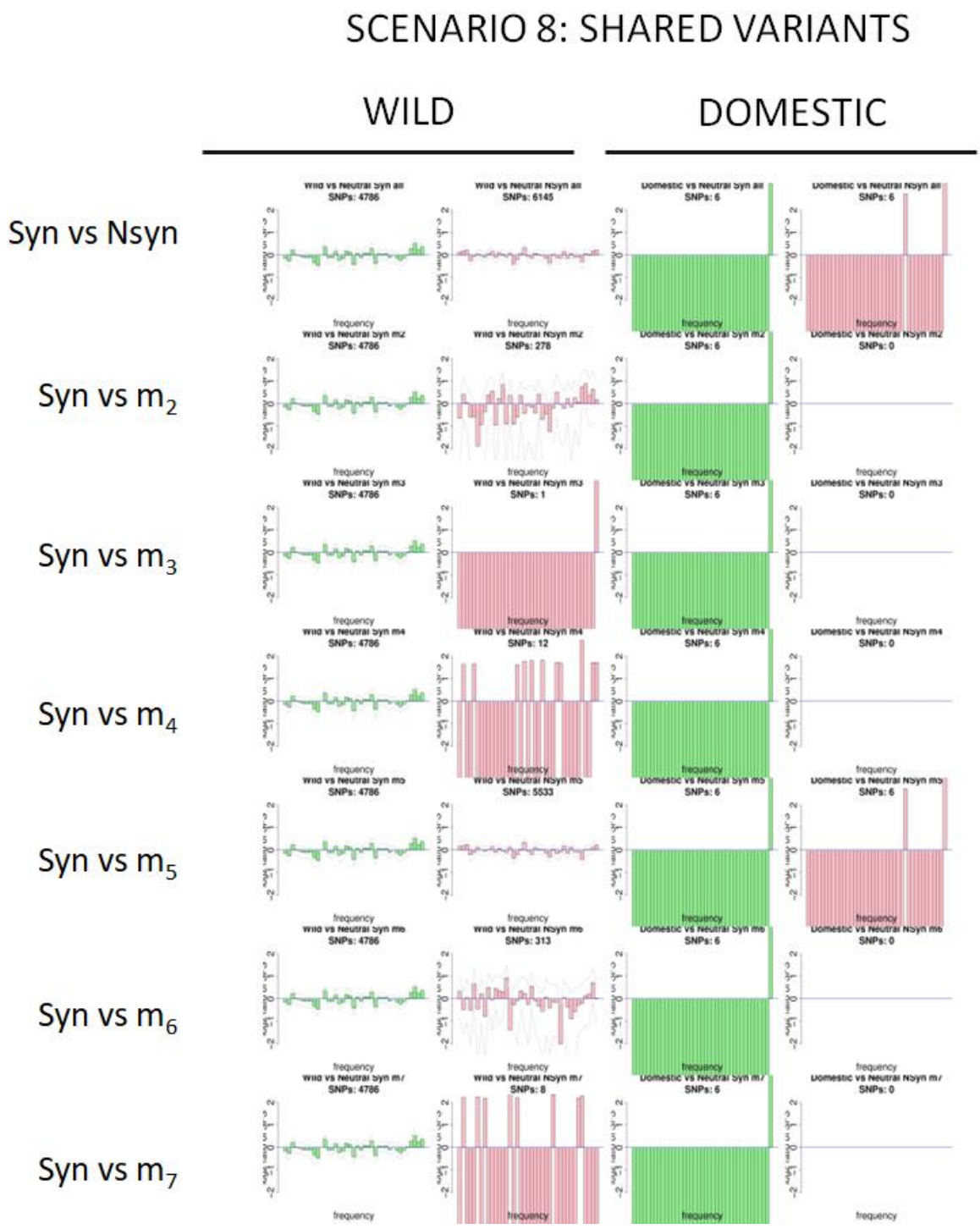

Figure A.21: log2 ratio of the simulated SFS for scenario 8 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

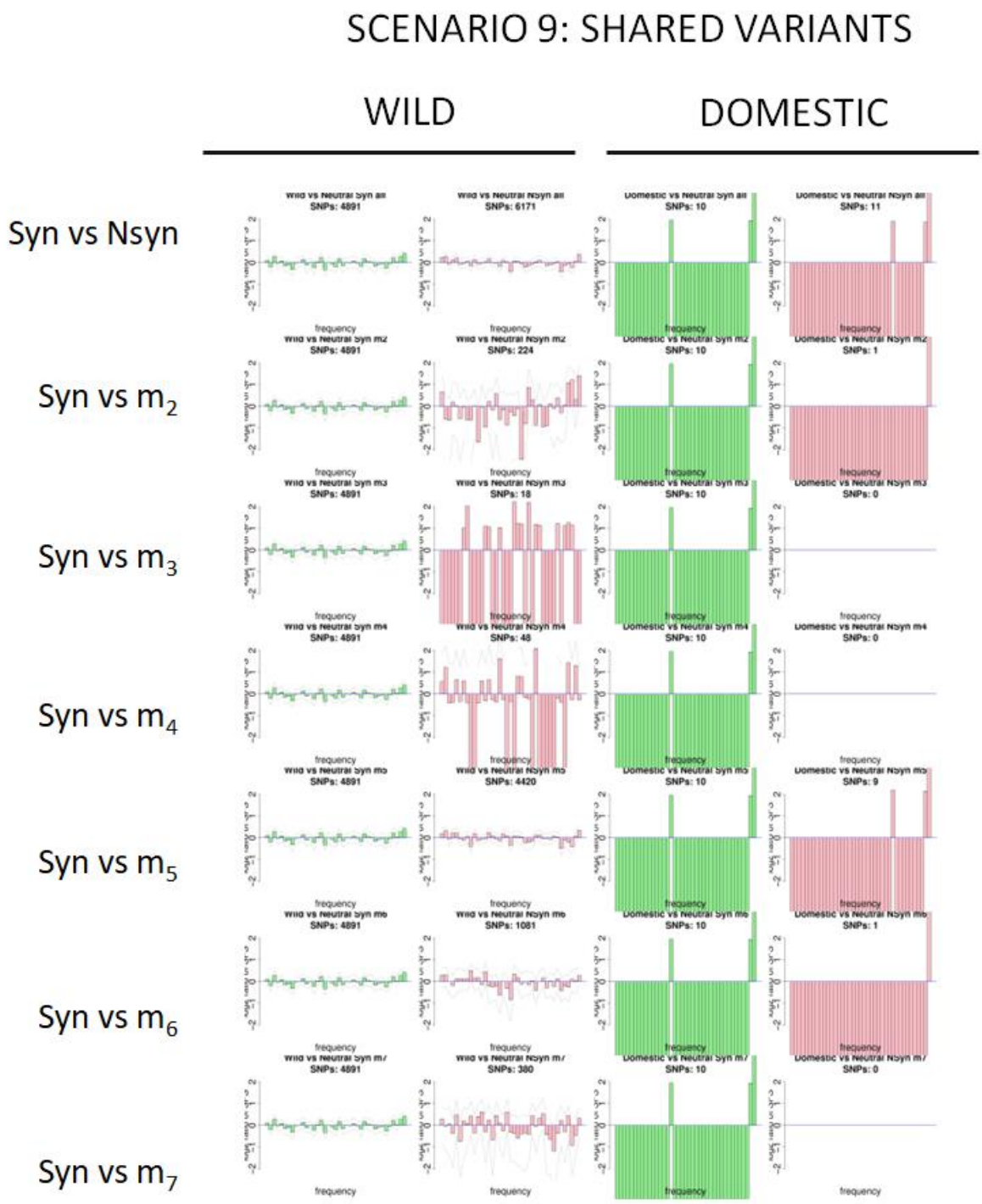

Figure A.22: log2 ratio of the simulated SFS for scenario 9 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively.

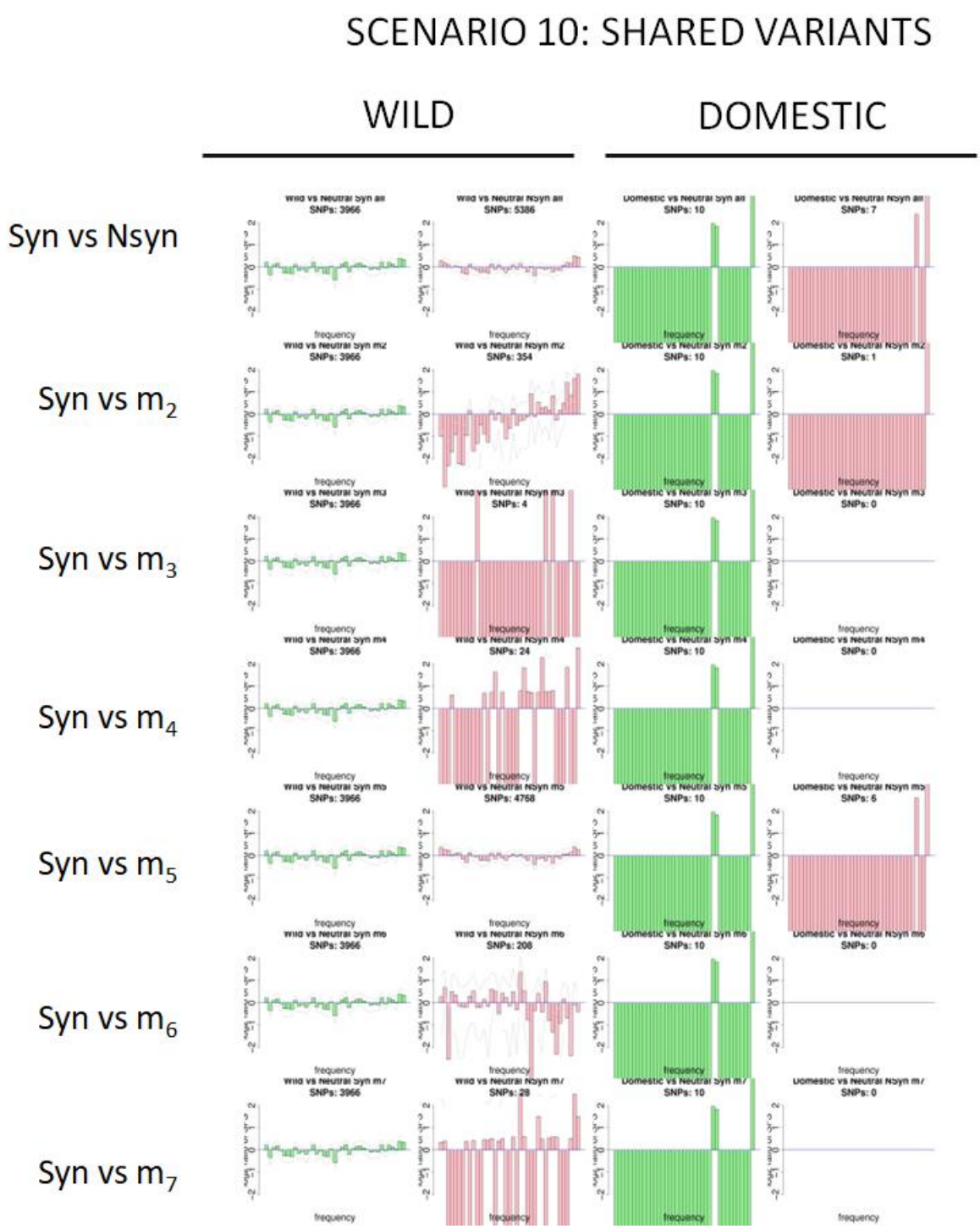

Figure A.23: log2 ratio of the simulated SFS for scenario 10 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 shared variants, respectively

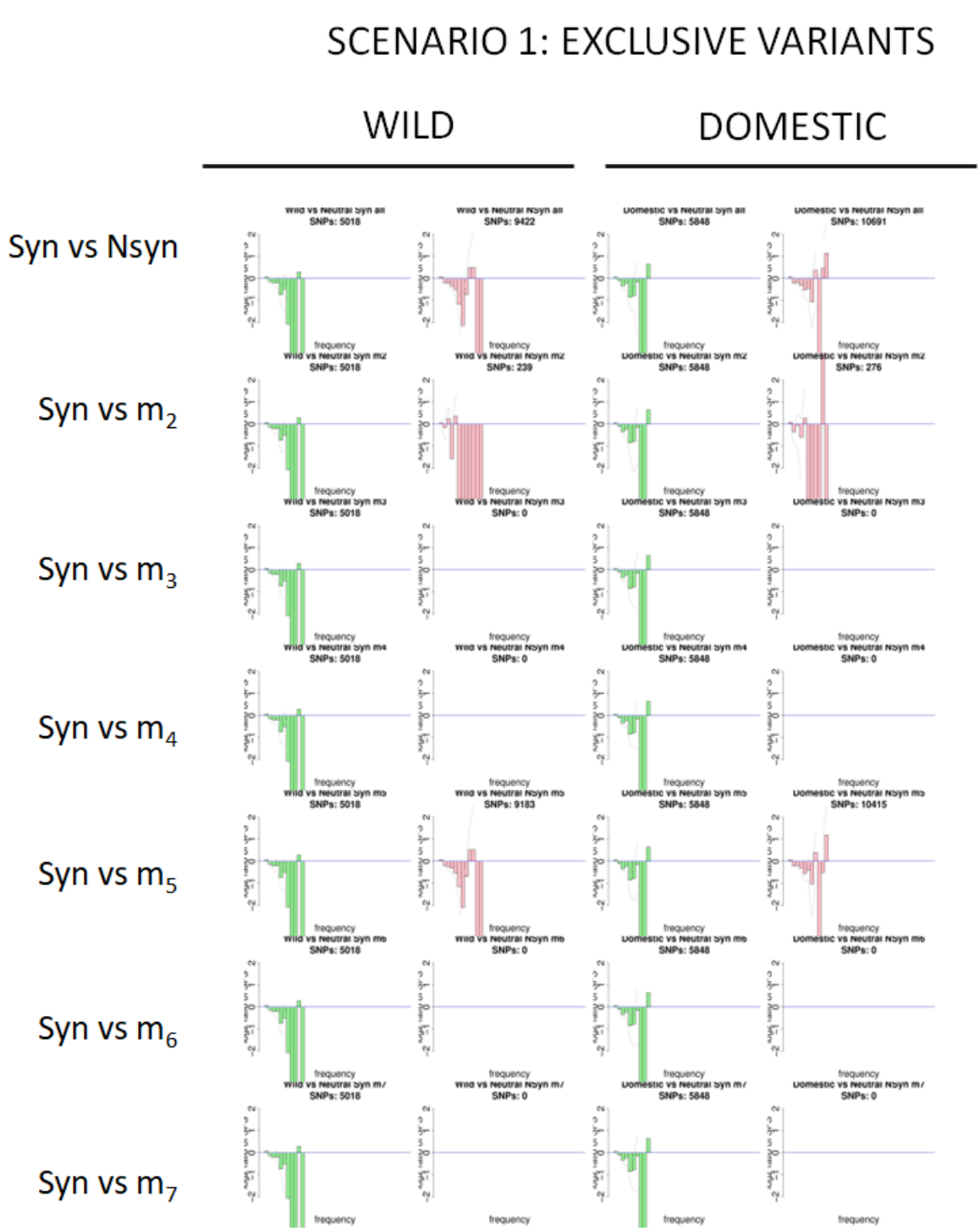

Figure A.24: log2 ratio of the simulated SFS for scenario 1 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

Figure A.25: log2 ratio of the simulated SFS for scenario 2 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

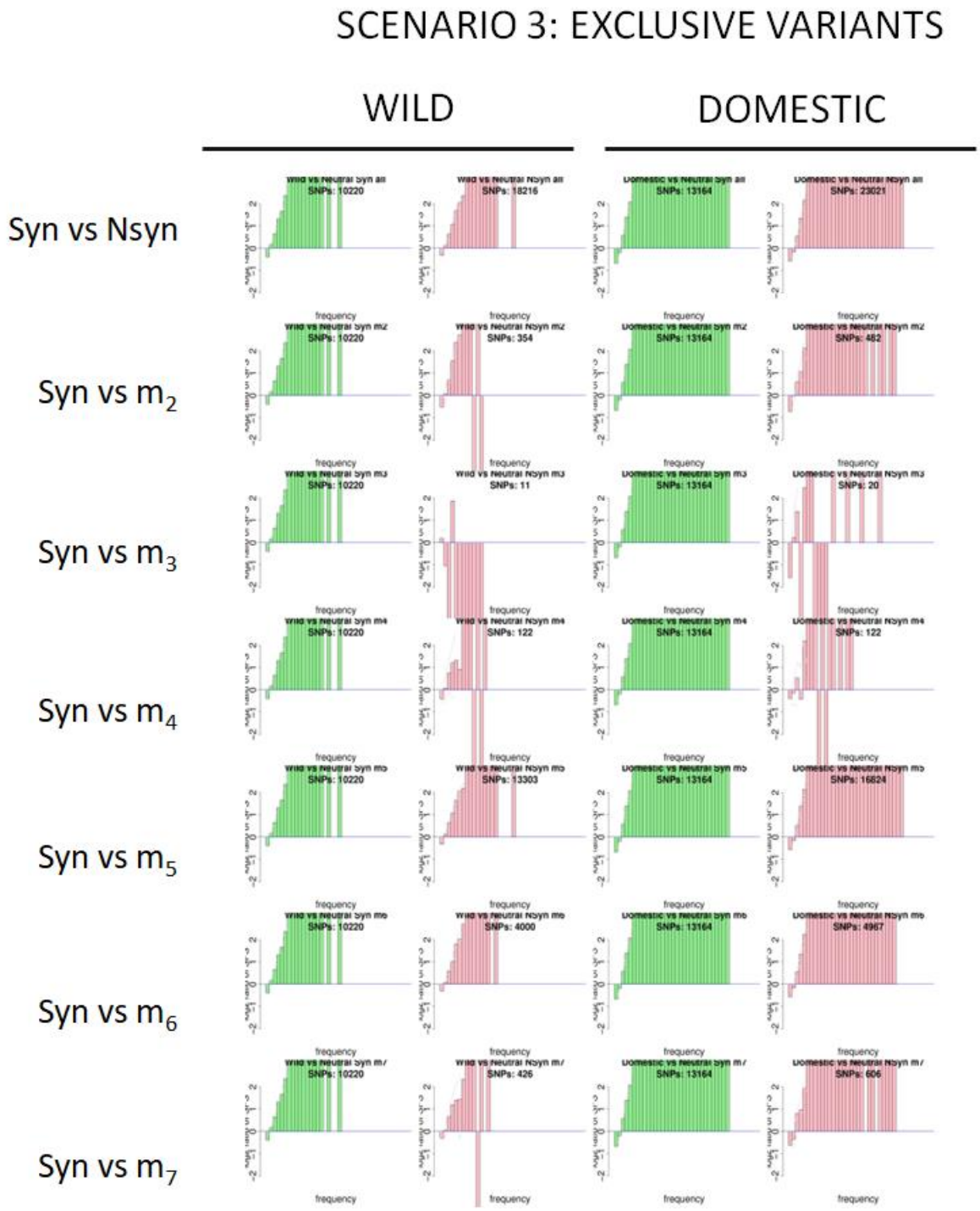

Figure A.26: log2 ratio of the simulated SFS for scenario 3 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

.

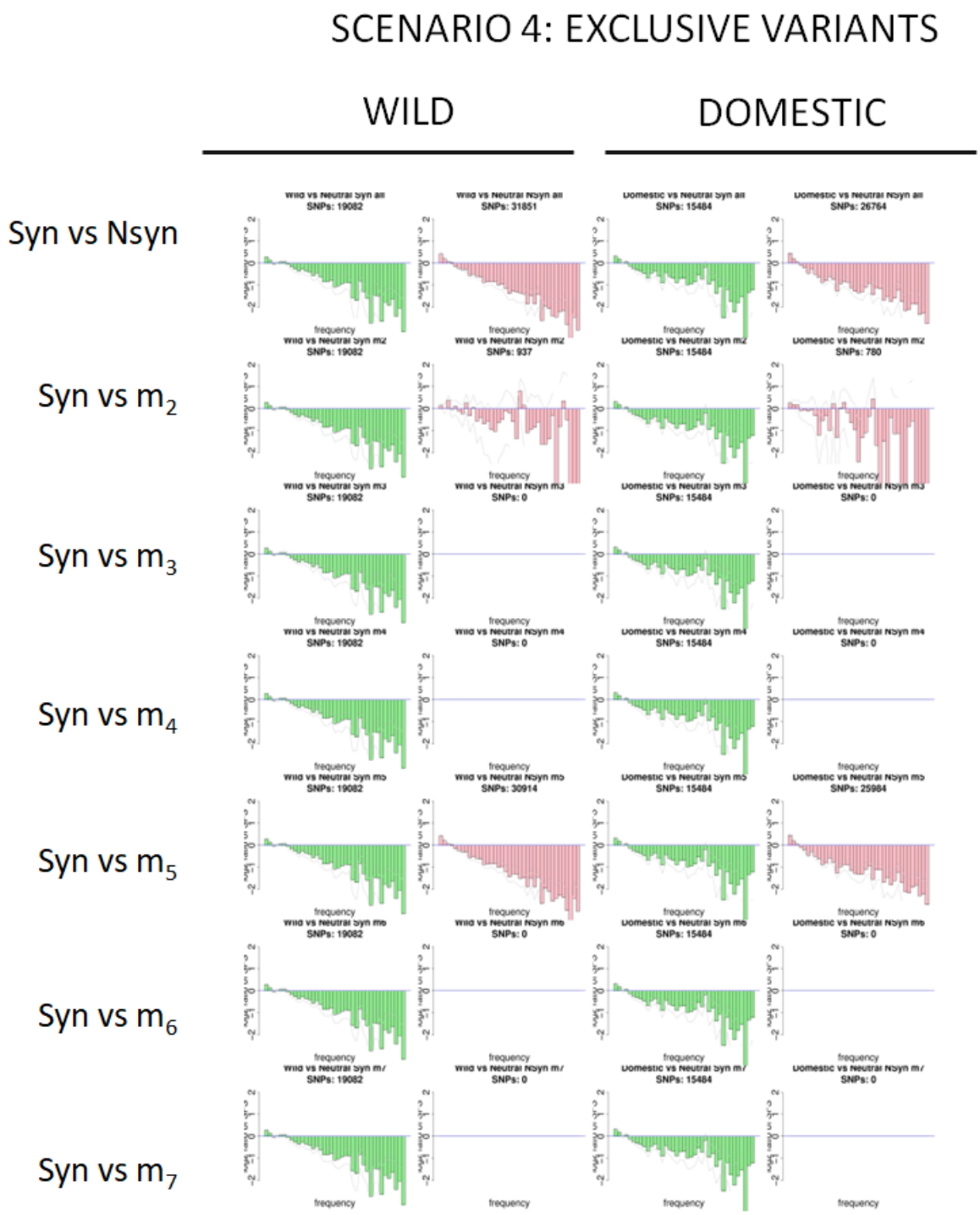

Figure A.27: log2 ratio of the simulated SFS for scenario 4 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

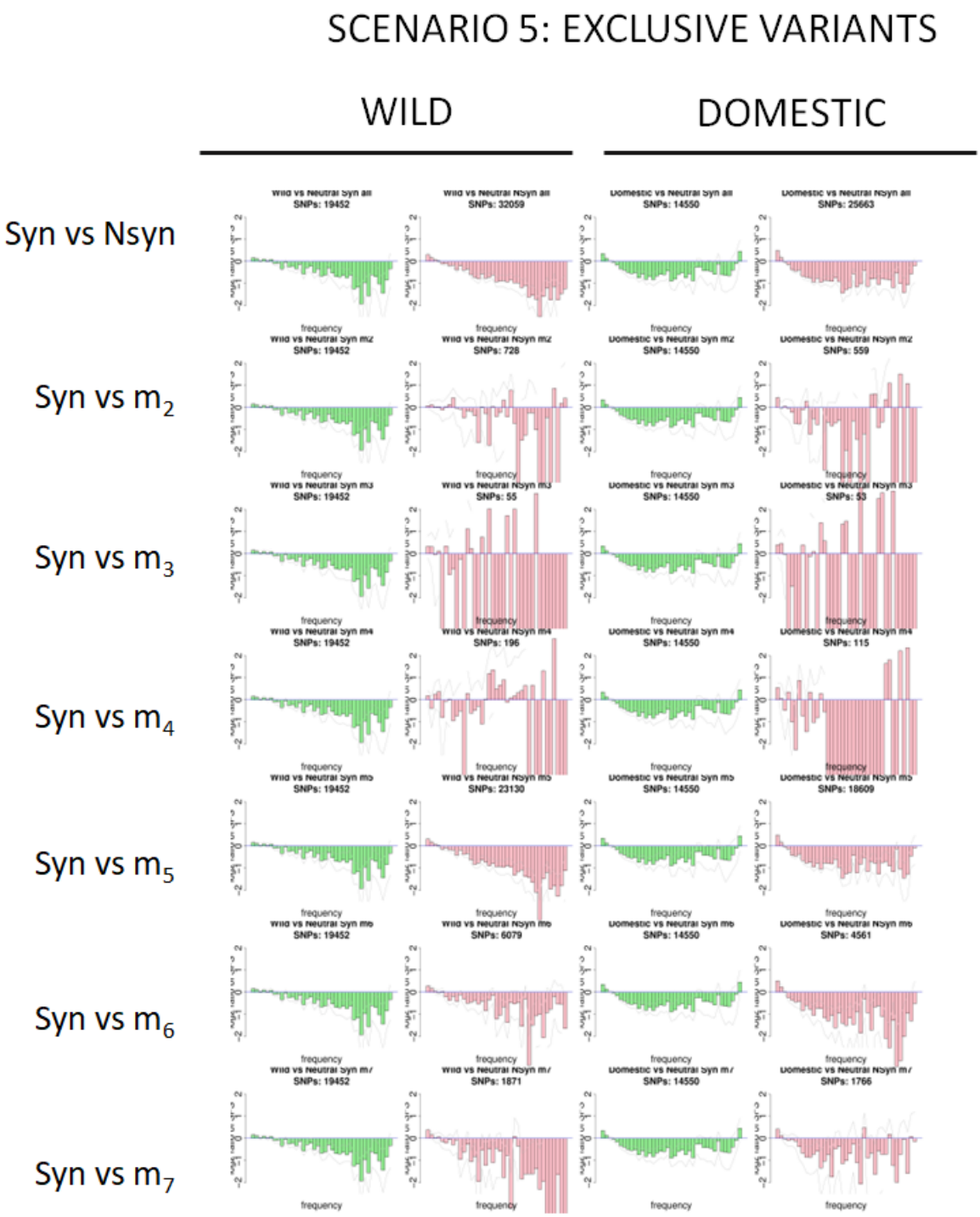

Figure A.28: log2 ratio of the simulated SFS for scenario 5 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

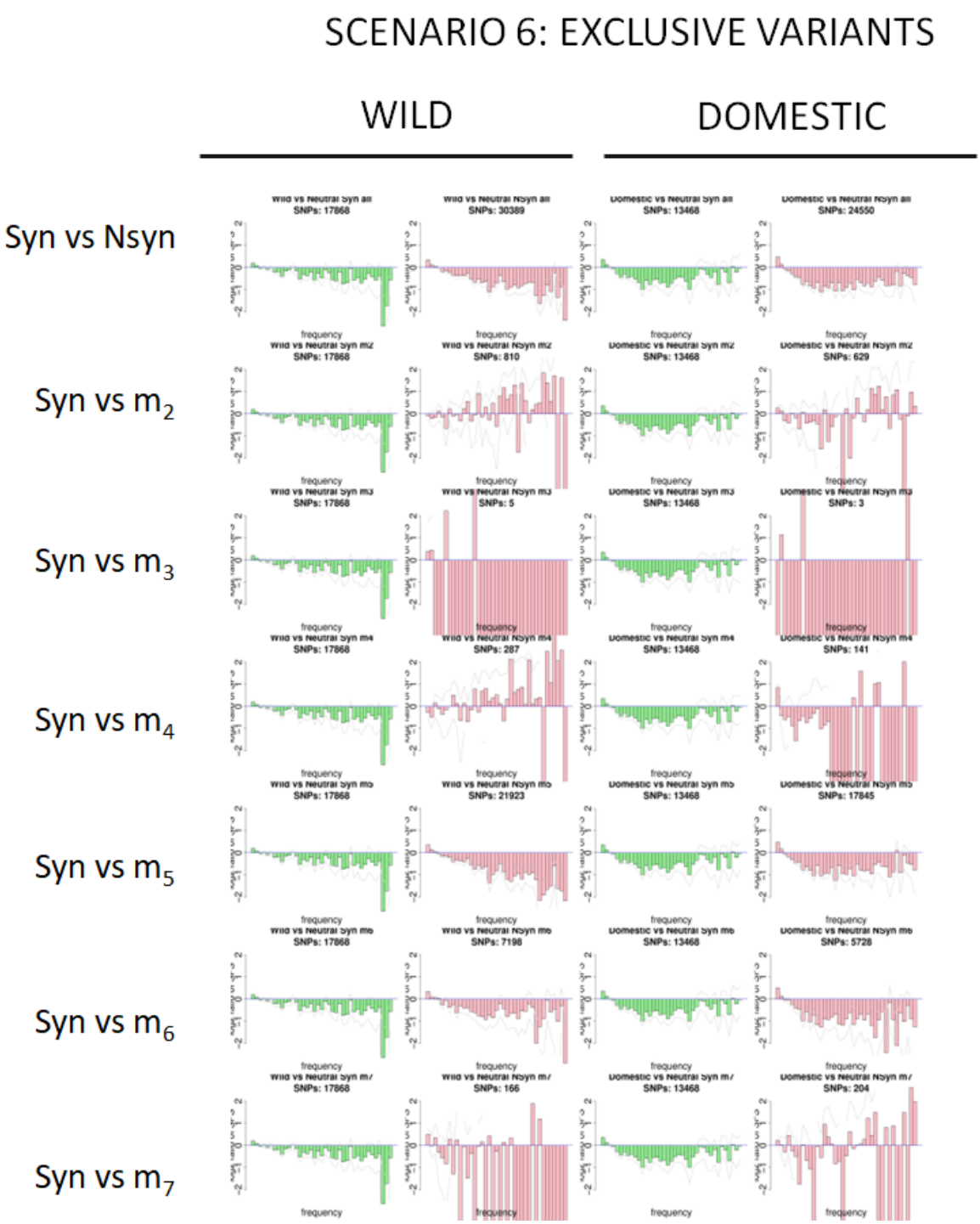

Figure A.29: log2 ratio of the simulated SFS for scenario 6 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

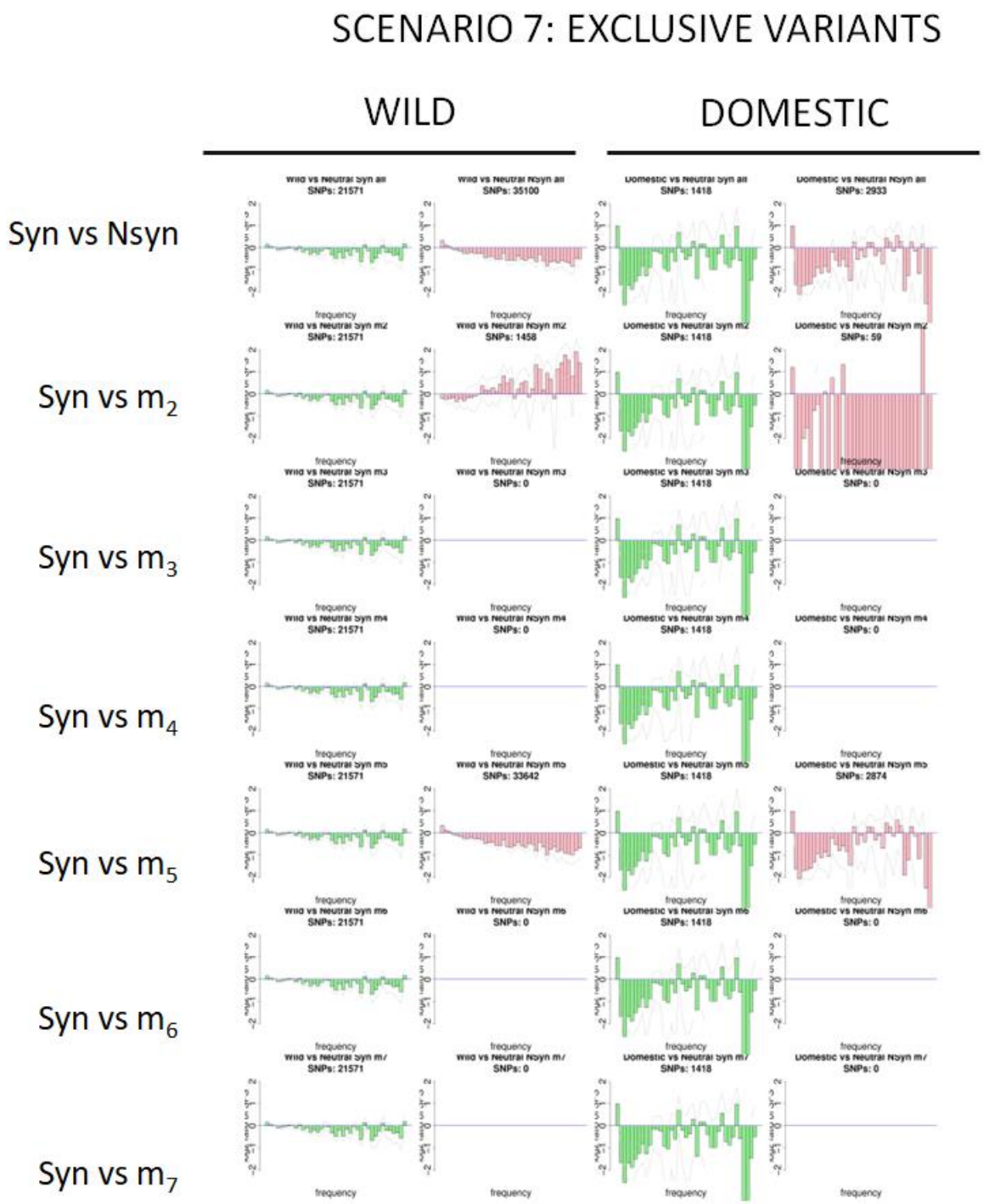

Figure A.30: log2 ratio of the simulated SFS for scenario 7 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

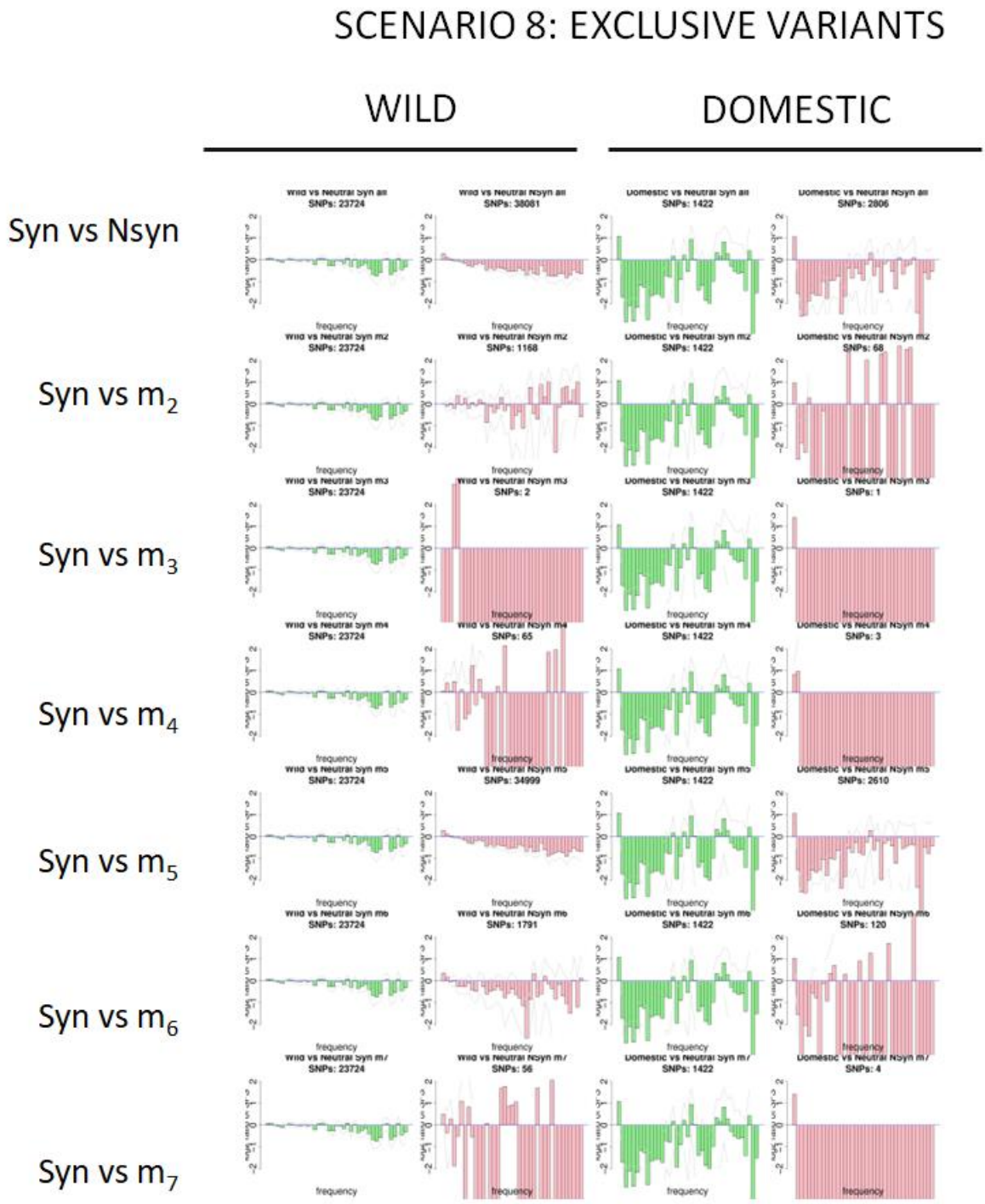

Figure A.31: log2 ratio of the simulated SFS for scenario 8 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

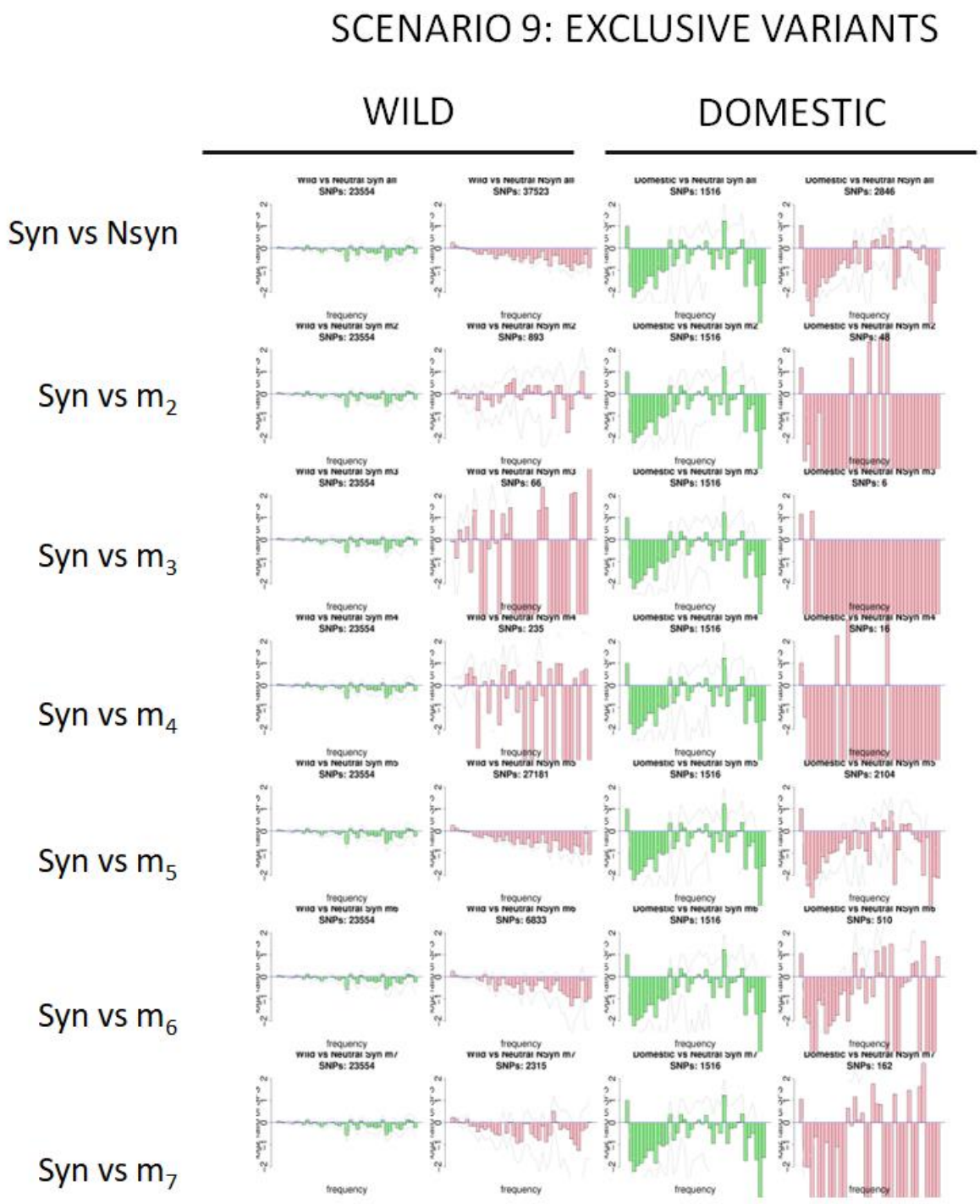

Figure A.32: log2 ratio of the simulated SFS for scenario 9 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively

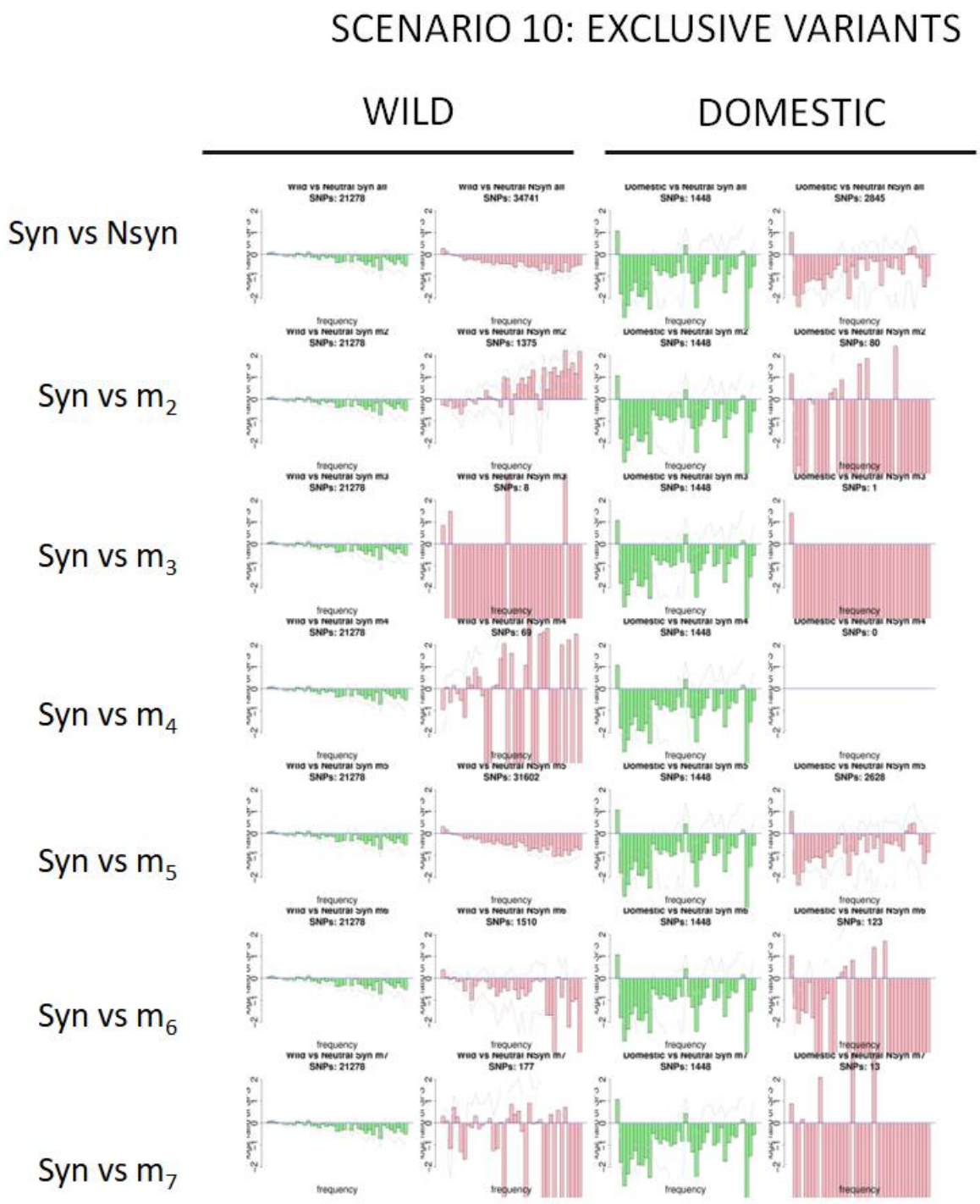

Figure A.33: log2 ratio of the simulated SFS for scenario 10 versus the expected SFS under the same demographic scenario (without selection) and per each type of variants. The seven rows correspond to nonsynonymous, m2, m3, m4, m5, m6 and m7 exclusive variants, respectively.

Appendix B

Transposable element polymorphisms improve prediction of complex agronomic traits in rice

Supplementary Information

Transposable element polymorphisms improve prediction of complex agronomic traits in rice

Table B.1: Accessions used in this study.

ACCESSION	GROUP	STATUS*	COUNTRY	ACCESSION	GROUP	STATUS*	COUNTRY
IRIS 313-10000	IND	I	South_Korea	IRIS 313-11684	IND	T	Thailand
IRIS 313-10010	IND	T	Fiji	IRIS 313-11685	IND	T	Thailand
IRIS 313-10020	AUS	T	Sri_Lanka	IRIS 313-11686	IND	T	Thailand
IRIS 313-10026	IND	NA	Madagascar	IRIS 313-11687	IND	T	Thailand
IRIS 313-10059	JAP	NA	South_Korea	IRIS 313-11689	JAP	NA	South_Korea
IRIS 313-10078	JAP	T	Japan	IRIS 313-11691	JAP	T	Bhutan
IRIS 313-10097	JAP	T	South_Korea	IRIS 313-11692	IND	T	Taiwan
IRIS 313-10109	IND	T	Nigeria	IRIS 313-11693	IND	T	Taiwan
IRIS 313-10114	IND	I	Burundi	IRIS 313-11700	IND	T	Thailand
IRIS 313-10134	IND	NA	Thailand	IRIS 313-11704	IND	T	Thailand
IRIS 313-10158	IND	NA	Ecuador	IRIS 313-11705	IND	T	Thailand
IRIS 313-10171	IND	NA	China	IRIS 313-11706	IND	T	Thailand
IRIS 313-10177	IND	NA	China	IRIS 313-11707	IND	T	Thailand
IRIS 313-10179	IND	NA	China	IRIS 313-11708	IND	T	Thailand
IRIS 313-10189	IND	NA	China	IRIS 313-11709	IND	T	Thailand
IRIS 313-10221	IND	NA	China	IRIS 313-11710	IND	T	Thailand
IRIS 313-10228	JAP	NA	China	IRIS 313-11711	IND	T	Thailand
IRIS 313-10235	IND	I	Philippines	IRIS 313-11712	AUS	NA	India
IRIS 313-10237	IND	I	Philippines	IRIS 313-11716	IND	T	Guinea
IRIS 313-10260	IND	NA	Paraguay	IRIS 313-11717	IND	T	Indonesia
IRIS 313-10301	IND	I	Brazil	IRIS 313-11719	IND	NA	Thailand
IRIS 313-10327	JAP	NA	Peru	IRIS 313-11720	IND	NA	Thailand
IRIS 313-10332	IND	I	Indonesia	IRIS 313-11721	IND	NA	Thailand
IRIS 313-10333	IND	I	Indonesia	IRIS 313-11722	IND	T	Bangladesh
IRIS 313-10337	IND	I	Indonesia	IRIS 313-11723	IND	T	Guinea
IRIS 313-10352	IND	I	Colombia	IRIS 313-11724	IND	T	Guinea
IRIS 313-10392	IND	I	Philippines	IRIS 313-11725	JAP	I	Japan
IRIS 313-10394	IND	I	Philippines	IRIS 313-11727	IND	NA	China
IRIS 313-10397	IND	I	Colombia	IRIS 313-11728	IND	NA	China
IRIS 313-10423	IND	NA	Myanmar	IRIS 313-11730	IND	NA	China
IRIS 313-10440	JAP	NA	Philippines	IRIS 313-11731	IND	NA	China
IRIS 313-10458	IND	P	China	IRIS 313-11732	IND	NA	China
IRIS 313-10469	JAP	I	Japan	IRIS 313-11733	IND	NA	China
IRIS 313-10477	IND	NA	China	IRIS 313-11734	IND	NA	China
IRIS 313-10503	IND	NA	China	IRIS 313-11736	JAP	T	Philippines
IRIS 313-10509	IND	NA	Africa	IRIS 313-11737	AUS	T	India
IRIS 313-10511	IND	T	Philippines	IRIS 313-11738	IND	T	India
IRIS 313-10515	IND	I	Taiwan	IRIS 313-11739	JAP	T	Ghana
IRIS 313-10518	IND	NA	Myanmar	IRIS 313-11740	IND	I	Ghana
IRIS 313-10519	IND	P	India	IRIS 313-11741	IND	T	Sri_Lanka
IRIS 313-10524	IND	NA	na	IRIS 313-11742	AUS	NA	India

IRIS 313-10526	IND	NA	India	IRIS 313-11744	IND	NA	China
IRIS 313-10534	AUS	P	India	IRIS 313-11745	IND	NA	China
IRIS 313-10541	JAP	P	Guinea-Bissau	IRIS 313-11746	IND	NA	China
IRIS 313-10542	IND	NA	India	IRIS 313-11747	IND	NA	China
IRIS 313-10544	IND	NA	India	IRIS 313-11748	IND	NA	China
IRIS 313-10547	IND	P	Myanmar	IRIS 313-11750	IND	NA	China
IRIS 313-10550	IND	i	Indonesia	IRIS 313-11751	IND	NA	China
IRIS 313-10560	IND	NA	China	IRIS 313-11752	IND	NA	China
IRIS 313-10561	IND	NA	China	IRIS 313-11753	IND	NA	China
IRIS 313-10563	JAP	P	na	IRIS 313-11754	ADM	NA	Madagascar
IRIS 313-10564	JAP	P	Japan	IRIS 313-11755	JAP	NA	Liberia
IRIS 313-10570	JAP	NA	Japan	IRIS 313-11756	JAP	NA	Madagascar
IRIS 313-10576	IND	NA	Sierra_Leone	IRIS 313-11757	IND	NA	Madagascar
IRIS 313-10577	JAP	T	Philippines	IRIS 313-11758	IND	NA	Ivory_Coast
IRIS 313-10578	JAP	T	Philippines	IRIS 313-11759	JAP	NA	Ivory_Coast
IRIS 313-10582	JAP	T	Philippines	IRIS 313-11760	ADM	NA	Madagascar
IRIS 313-10602	AUS	P	Bangladesh	IRIS 313-11761	ADM	NA	Ivory_Coast
IRIS 313-10603	AUS	P	Bangladesh	IRIS 313-11762	IND	NA	Madagascar
IRIS 313-10605	AUS	P	Bangladesh	IRIS 313-11763	IND	I	Cameroon
IRIS 313-10609	IND	T	Sri_Lanka	IRIS 313-11764	IND	I	Liberia
IRIS 313-10614	IND	NA	Hong_Kong	IRIS 313-11767	ADM	NA	Madagascar
IRIS 313-10617	JAP	NA	na	IRIS 313-11772	IND	T	Madagascar
IRIS 313-10623	AUS	NA	Nepal	IRIS 313-11773	IND	NA	Gambia
IRIS 313-10628	IND	T	India	IRIS 313-11784	IND	T	Sierra_Leone
IRIS 313-10642	JAP	P	Japan	IRIS 313-11786	IND	T	Gambia
IRIS 313-10652	IND	T	Laos	IRIS 313-11787	IND	T	Gambia
IRIS 313-10654	IND	P	Laos	IRIS 313-11788	JAP	T	Philippines
IRIS 313-10657	JAP	NA	Laos	IRIS 313-11789	ADM	T	Madagascar
IRIS 313-10664	IND	NA	India	IRIS 313-11790	JAP	T	Madagascar
IRIS 313-10666	IND	NA	India	IRIS 313-11791	IND	T	Madagascar
IRIS 313-10671	AUS	NA	India	IRIS 313-11792	JAP	T	Madagascar
IRIS 313-10675	AUS	T	India	IRIS 313-11794	ADM	NA	Madagascar
IRIS 313-10677	JAP	P	Japan	IRIS 313-11795	IND	NA	China
IRIS 313-10682	IND	T	Laos	IRIS 313-11796	IND	NA	China
IRIS 313-10687	IND	T	Malaysia	IRIS 313-11797	IND	NA	China
IRIS 313-10688	IND	T	Malaysia	IRIS 313-11798	IND	NA	China
IRIS 313-10693	JAP	T	Indonesia	IRIS 313-11799	IND	NA	China
IRIS 313-10697	IND	T	Malaysia	IRIS 313-11800	JAP	NA	China
IRIS 313-10703	JAP	T	Malaysia	IRIS 313-11801	IND	NA	China
IRIS 313-10706	IND	T	Malaysia	IRIS 313-11802	IND	NA	China
IRIS 313-10707	IND	T	Malaysia	IRIS 313-11804	IND	NA	China
IRIS 313-10710	JAP	T	Surinam	IRIS 313-11805	IND	NA	China
IRIS 313-10712	JAP	NA	Ivory_Coast	IRIS 313-11806	IND	NA	China
IRIS 313-10718	AUS	T	Sri_Lanka	IRIS 313-11807	IND	I	Colombia
IRIS 313-10723	IND	NA	Senegal	IRIS 313-11809	AUS	T	Kenya
IRIS 313-10725	IND	NA	Senegal	IRIS 313-11810	IND	T	Kenya
IRIS 313-10726	IND	NA	Senegal	IRIS 313-11811	ADM	T	Kenya
IRIS 313-10727	IND	NA	Senegal	IRIS 313-11812	IND	T	Kenya
IRIS 313-10728	IND	NA	Senegal	IRIS 313-11813	IND	T	Kenya
IRIS 313-10733	IND	NA	Nepal	IRIS 313-11814	IND	T	Kenya
IRIS 313-10744	JAP	T	Indonesia	IRIS 313-11815	IND	T	Kenya
IRIS 313-10748	IND	T	Vietnam	IRIS 313-11816	IND	T	Myanmar
IRIS 313-10756	IND	I	India	IRIS 313-11817	IND	T	Myanmar
IRIS 313-10762	IND	T	Indonesia	IRIS 313-11819	IND	T	Myanmar
IRIS 313-10771	ADM	T	Indonesia	IRIS 313-11820	IND	T	Myanmar

IRIS 313-10774	IND	T	Indonesia	IRIS 313-11821	IND	T	India
IRIS 313-10778	IND	T	Indonesia	IRIS 313-11822	IND	T	India
IRIS 313-10779	IND	T	Indonesia	IRIS 313-11823	IND	T	India
IRIS 313-10786	ADM	T	Indonesia	IRIS 313-11824	IND	T	India
IRIS 313-10787	ADM	T	Indonesia	IRIS 313-11825	ARO	T	India
IRIS 313-10789	JAP	T	Indonesia	IRIS 313-11829	JAP	T	Pakistan
IRIS 313-10793	JAP	T	Indonesia	IRIS 313-11832	JAP	T	Thailand
IRIS 313-10794	JAP	T	Indonesia	IRIS 313-11833	IND	T	Thailand
IRIS 313-10797	IND	T	Indonesia	IRIS 313-11835	IND	T	Thailand
IRIS 313-10798	JAP	T	Indonesia	IRIS 313-11836	IND	T	Thailand
IRIS 313-10802	JAP	T	Indonesia	IRIS 313-11840	IND	T	Thailand
IRIS 313-10805	JAP	T	Indonesia	IRIS 313-11842	IND	T	Thailand
IRIS 313-10806	IND	T	Indonesia	IRIS 313-11848	IND	T	Malaysia
IRIS 313-10810	IND	T	Indonesia	IRIS 313-11854	IND	NA	China
IRIS 313-10813	IND	T	Indonesia	IRIS 313-11866	IND	NA	China
IRIS 313-10814	IND	T	Indonesia	IRIS 313-11867	IND	NA	China
IRIS 313-10816	JAP	T	Indonesia	IRIS 313-11870	IND	NA	China
IRIS 313-10820	IND	T	Indonesia	IRIS 313-11877	IND	NA	China
IRIS 313-10822	IND	T	Indonesia	IRIS 313-11878	IND	NA	China
IRIS 313-10824	IND	T	Indonesia	IRIS 313-11882	IND	NA	China
IRIS 313-10825	IND	NA	NA	IRIS 313-11887	IND	I	Philippines
IRIS 313-10827	JAP	T	Philippines	IRIS 313-11897	JAP	T	Thailand
IRIS 313-10834	JAP	I	India	IRIS 313-11900	JAP	NA	Thailand
IRIS 313-10835	IND	I	India	IRIS 313-11902	IND	NA	Indonesia
IRIS 313-10840	JAP	T	South_Korea	IRIS 313-11909	IND	NA	China
IRIS 313-10845	AUS	NA	India	IRIS 313-11916	IND	T	Sri_Lanka
IRIS 313-10847	IND	NA	India	IRIS 313-11919	IND	T	India
IRIS 313-10850	ADM	NA	India	IRIS 313-11920	IND	NA	Thailand
IRIS 313-10851	ARO	NA	India	IRIS 313-11924	JAP	NA	Thailand
IRIS 313-10852	AUS	NA	India	IRIS 313-11927	IND	NA	Thailand
IRIS 313-10857	AUS	NA	India	IRIS 313-11929	JAP	T	Philippines
IRIS 313-10858	AUS	NA	India	IRIS 313-11930	ADM	P	Nigeria
IRIS 313-10859	AUS	NA	India	IRIS 313-11935	IND	T	Cambodia
IRIS 313-10861	AUS	NA	India	IRIS 313-11939	IND	T	Burkina_Faso
IRIS 313-10863	IND	NA	India	IRIS 313-11940	IND	T	Burkina_Faso
IRIS 313-10870	JAP	NA	India	IRIS 313-11941	IND	T	Burkina_Faso
IRIS 313-10871	AUS	NA	India	IRIS 313-11945	IND	T	Bangladesh
IRIS 313-10872	ADM	NA	India	IRIS 313-11949	IND	NA	China
IRIS 313-10883	ARO	NA	India	IRIS 313-11950	IND	NA	China
IRIS 313-10888	JAP	NA	India	IRIS 313-11953	IND	NA	China
IRIS 313-10889	JAP	NA	India	IRIS 313-11955	IND	NA	China
IRIS 313-10891	AUS	NA	India	IRIS 313-11959	IND	T	Philippines
IRIS 313-10892	AUS	NA	India	IRIS 313-11962	IND	NA	Thailand
IRIS 313-10894	AUS	NA	India	IRIS 313-11966	IND	T	China
IRIS 313-10895	JAP	NA	India	IRIS 313-11978	IND	I	Philippines
IRIS 313-10900	IND	T	Cambodia	IRIS 313-11979	IND	I	Philippines
IRIS 313-10912	IND	T	Cambodia	IRIS 313-11988	IND	T	Sierra_Leone
IRIS 313-10916	JAP	T	Cambodia	IRIS 313-11989	IND	T	Brunei_Darussalam
IRIS 313-10918	ADM	T	Philippines	IRIS 313-11994	JAP	T	Philippines
IRIS 313-10921	IND	NA	Laos	IRIS 313-11999	IND	T	Cambodia
IRIS 313-10922	JAP	NA	Laos	IRIS 313-12000	IND	T	Cambodia
IRIS 313-10923	JAP	T	Thailand	IRIS 313-12010	IND	NA	China
IRIS 313-10927	AUS	T	Nepal	IRIS 313-12024	IND	NA	na
IRIS 313-10928	IND	T	Thailand	IRIS 313-12033	IND	NA	China
IRIS 313-10930	AUS	P	Bangladesh	IRIS 313-12040	IND	T	Cambodia

IRIS 313-10936	JAP	NA	Indonesia	IRIS 313-12041	IND	T	Cambodia
IRIS 313-10937	IND	T	Indonesia	IRIS 313-12044	IND	T	Cambodia
IRIS 313-10938	IND	T	Indonesia	IRIS 313-12045	JAP	T	Indonesia
IRIS 313-10941	IND	NA	Indonesia	IRIS 313-12048	IND	T	Indonesia
IRIS 313-10942	IND	T	Indonesia	IRIS 313-12052	IND	T	India
IRIS 313-10944	IND	T	Indonesia	IRIS 313-12054	JAP	I	China
IRIS 313-10948	ADM	NA	Indonesia	IRIS 313-12058	IND	T	Cambodia
IRIS 313-10949	JAP	NA	Indonesia	IRIS 313-12060	JAP	NA	China
IRIS 313-10954	IND	T	Indonesia	IRIS 313-12061	JAP	NA	China
IRIS 313-10960	JAP	T	Indonesia	IRIS 313-12068	JAP	T	Indonesia
IRIS 313-10961	IND	T	Indonesia	IRIS 313-12071	JAP	NA	Laos
IRIS 313-10965	AUS	T	Bangladesh	IRIS 313-12094	ARO	T	Bangladesh
IRIS 313-10968	IND	NA	Brazil	IRIS 313-12097	IND	T	Cambodia
IRIS 313-10975	IND	T	Bangladesh	IRIS 313-12101	IND	T	Cambodia
IRIS 313-10980	IND	T	Bangladesh	IRIS 313-12108	JAP	NA	Malaysia
IRIS 313-10984	IND	T	Bangladesh	IRIS 313-12118	ADM	NA	Madagascar
IRIS 313-10986	IND	T	Bangladesh	IRIS 313-12121	IND	T	Laos
IRIS 313-10988	IND	P	India	IRIS 313-12127	IND	T	Laos
IRIS 313-10990	IND	T	Philippines	IRIS 313-12128	IND	T	Laos
IRIS 313-10994	JAP	T	Philippines	IRIS 313-12129	JAP	T	Laos
IRIS 313-10995	IND	T	Indonesia	IRIS 313-12131	IND	NA	Laos
IRIS 313-10997	IND	T	Indonesia	IRIS 313-12134	JAP	T	Laos
IRIS 313-10999	JAP	T	Indonesia	IRIS 313-12135	IND	NA	Malaysia
IRIS 313-11005	JAP	T	Indonesia	IRIS 313-12141	AUS	T	Bangladesh
IRIS 313-11007	JAP	T	Indonesia	IRIS 313-12146	IND	T	Cambodia
IRIS 313-11016	AUS	T	Bangladesh	IRIS 313-12164	JAP	T	Cambodia
IRIS 313-11024	AUS	T	Pakistan	IRIS 313-12183	AUS	T	Nepal
IRIS 313-11027	AUS	T	Pakistan	IRIS 313-12188	IND	NA	Laos
IRIS 313-11032	ARO	T	Pakistan	IRIS 313-12190	IND	T	Laos
IRIS 313-11033	IND	NA	Pakistan	IRIS 313-12193	IND	T	Laos
IRIS 313-11034	AUS	NA	Pakistan	IRIS 313-12228	JAP	T	Laos
IRIS 313-11035	AUS	NA	Pakistan	IRIS 313-12234	IND	NA	China
IRIS 313-11037	AUS	T	Pakistan	IRIS 313-12258	JAP	T	Laos
IRIS 313-11038	IND	P	China	IRIS 313-12259	IND	NA	Laos
IRIS 313-11039	IND	I	China	IRIS 313-12262	JAP	NA	Laos
IRIS 313-11040	IND	NA	India	IRIS 313-12268	IND	T	Myanmar
IRIS 313-11042	IND	NA	India	IRIS 313-12275	IND	NA	China
IRIS 313-11043	IND	NA	Malaysia	IRIS 313-12281	JAP	NA	Madagascar
IRIS 313-11044	JAP	NA	Malaysia	IRIS 313-12287	IND	T	Myanmar
IRIS 313-11045	JAP	NA	Malaysia	IRIS 313-12289	JAP	T	Myanmar
IRIS 313-11046	JAP	NA	Malaysia	IRIS 313-12291	IND	T	Myanmar
IRIS 313-11047	AUS	P	Bangladesh	IRIS 313-12300	IND	T	Laos
IRIS 313-11048	AUS	P	Bangladesh	IRIS 313-12303	IND	T	Laos
IRIS 313-11049	AUS	P	Bangladesh	IRIS 313-12305	IND	T	Laos
IRIS 313-11050	AUS	P	Bangladesh	IRIS 313-12307	JAP	NA	Laos
IRIS 313-11051	AUS	P	Bangladesh	IRIS 313-12312	JAP	NA	Laos
IRIS 313-11052	AUS	P	Bangladesh	IRIS 313-12321	JAP	T	Laos
IRIS 313-11053	AUS	P	Bangladesh	IRIS 313-12323	JAP	T	Laos
IRIS 313-11054	AUS	P	Bangladesh	IRIS 313-12334	IND	T	Laos
IRIS 313-11055	AUS	P	Bangladesh	IRIS 313-12349	JAP	NA	Laos
IRIS 313-11056	AUS	P	Bangladesh	IRIS 313-12350	JAP	T	Laos
IRIS 313-11057	AUS	P	Bangladesh	IRIS 313-12351	JAP	T	Laos
IRIS 313-11058	AUS	P	Bangladesh	IRIS 313-12354	IND	T	Laos
IRIS 313-11059	AUS	P	Bangladesh	IRIS 313-12355	IND	T	Laos
IRIS 313-11062	ARO	P	Bangladesh	IRIS 313-15900	IND	I	Philippines

IRIS 313-11063	AUS	P	Bangladesh	IRIS 313-15908	AUS	I	Colombia
IRIS 313-11064	AUS	P	Bangladesh	IRIS 313-15910	JAP	I	United_States
IRIS 313-11066	ARO	P	Bangladesh	IRIS 313-7638	IND	NA	Madagascar
IRIS 313-11079	IND	T	Laos	IRIS 313-7646	ADM	NA	Madagascar
IRIS 313-11081	IND	T	Laos	IRIS 313-7650	IND	NA	Madagascar
IRIS 313-11083	IND	T	Laos	IRIS 313-7685	IND	I	Philippines
IRIS 313-11085	IND	NA	Laos	IRIS 313-7688	IND	I	Philippines
IRIS 313-11089	IND	NA	Cambodia	IRIS 313-7689	IND	I	Philippines
IRIS 313-11094	JAP	T	Laos	IRIS 313-7719	IND	NA	Mali
IRIS 313-11095	IND	T	Laos	IRIS 313-7722	ADM	NA	Madagascar
IRIS 313-11097	IND	T	Philippines	IRIS 313-7725	ADM	T	Madagascar
IRIS 313-11098	IND	NA	Sierra_Leone	IRIS 313-7795	ADM	T	Madagascar
IRIS 313-11102	JAP	NA	Liberia	IRIS 313-7797	IND	I	Philippines
IRIS 313-11103	JAP	NA	Liberia	IRIS 313-7799	IND	NA	Madagascar
IRIS 313-11104	JAP	NA	Liberia	IRIS 313-7808	IND	I	Senegal
IRIS 313-11112	AUS	T	Bangladesh	IRIS 313-7816	IND	I	Senegal
IRIS 313-11113	IND	T	Bangladesh	IRIS 313-7850	JAP	NA	Madagascar
IRIS 313-11118	IND	T	Vietnam	IRIS 313-7866	ADM	I	Colombia
IRIS 313-11129	IND	NA	Myanmar	IRIS 313-7876	JAP	T	Philippines
IRIS 313-11151	IND	I	Myanmar	IRIS 313-7883	JAP	T	Indonesia
IRIS 313-11160	IND	NA	Liberia	IRIS 313-7902	JAP	I	Philippines
IRIS 313-11189	ARO	NA	Soviet_Union	IRIS 313-7909	ADM	I	Philippines
IRIS 313-11191	AUS	T	Sri_Lanka	IRIS 313-7911	IND	I	Philippines
IRIS 313-11194	IND	NA	Thailand	IRIS 313-7914	JAP	I	Ivory_Coast
IRIS 313-11202	JAP	I	China	IRIS 313-7924	ADM	I	Bolivia
IRIS 313-11205	IND	T	Bangladesh	IRIS 313-7933	ADM	T	Nepal
IRIS 313-11221	IND	P	Bangladesh	IRIS 313-7994	JAP	T	Madagascar
IRIS 313-11224	IND	P	Bangladesh	IRIS 313-8010	JAP	I	Philippines
IRIS 313-11226	IND	P	Bangladesh	IRIS 313-8024	JAP	NA	Italy
IRIS 313-11228	IND	P	Bangladesh	IRIS 313-8037	JAP	NA	Italy
IRIS 313-11229	IND	P	Bangladesh	IRIS 313-8064	JAP	I	Argentina
IRIS 313-11231	IND	P	Bangladesh	IRIS 313-8066	JAP	NA	Italy
IRIS 313-11234	IND	T	Philippines	IRIS 313-8074	JAP	I	Australia
IRIS 313-11238	JAP	NA	Brazil	IRIS 313-8085	JAP	NA	Spain
IRIS 313-11239	IND	I	Indonesia	IRIS 313-8115	JAP	NA	Portugal
IRIS 313-11240	IND	I	India	IRIS 313-8118	JAP	NA	Portugal
IRIS 313-11241	IND	I	Bangladesh	IRIS 313-8119	JAP	NA	Bulgaria
IRIS 313-11242	IND	I	India	IRIS 313-8123	JAP	NA	Portugal
IRIS 313-11244	IND	NA	India	IRIS 313-8125	JAP	NA	Bulgaria
IRIS 313-11245	IND	P	India	IRIS 313-8127	JAP	NA	Bulgaria
IRIS 313-11247	IND	I	India	IRIS 313-8129	JAP	NA	Bulgaria
IRIS 313-11249	IND	I	Philippines	IRIS 313-8140	JAP	NA	China
IRIS 313-11251	IND	I	Philippines	IRIS 313-8151	JAP	P	Portugal
IRIS 313-11252	IND	NA	India	IRIS 313-8166	JAP	NA	France
IRIS 313-11253	IND	I	Surinam	IRIS 313-8167	JAP	I	France
IRIS 313-11256	IND	NA	India	IRIS 313-8168	JAP	NA	France
IRIS 313-11257	ADM	NA	India	IRIS 313-8172	ADM	NA	Philippines
IRIS 313-11258	ARO	NA	India	IRIS 313-8173	JAP	I	United_States
IRIS 313-11260	IND	NA	India	IRIS 313-8177	JAP	NA	Italy
IRIS 313-11262	IND	NA	India	IRIS 313-8185	JAP	NA	Italy
IRIS 313-11263	ADM	NA	India	IRIS 313-8204	JAP	I	United_States
IRIS 313-11264	IND	NA	India	IRIS 313-8208	JAP	NA	Portugal
IRIS 313-11265	AUS	NA	India	IRIS 313-8293	IND	NA	Senegal
IRIS 313-11266	IND	NA	India	IRIS 313-8305	IND	T	India
IRIS 313-11267	IND	NA	India	IRIS 313-8312	IND	T	Malaysia

IRIS 313-11269	IND	NA	India	IRIS 313-8321	AUS	P	Bangladesh
IRIS 313-11270	ARO	NA	India	IRIS 313-8323	JAP	I	United_States
IRIS 313-11271	IND	NA	India	IRIS 313-8326	ARO	NA	India
IRIS 313-11273	IND	NA	India	IRIS 313-8332	IND	NA	India
IRIS 313-11274	AUS	NA	India	IRIS 313-8341	IND	NA	Vietnam
IRIS 313-11275	IND	NA	India	IRIS 313-8349	IND	T	Bangladesh
IRIS 313-11277	AUS	NA	India	IRIS 313-8356	JAP	T	Philippines
IRIS 313-11278	IND	NA	India	IRIS 313-8381	JAP	T	Malaysia
IRIS 313-11279	IND	NA	India	IRIS 313-8386	IND	NA	India
IRIS 313-11280	IND	NA	India	IRIS 313-8391	IND	NA	Burkina_Faso
IRIS 313-11281	IND	NA	India	IRIS 313-8407	IND	T	Malaysia
IRIS 313-11285	IND	NA	India	IRIS 313-8436	JAP	T	Indonesia
IRIS 313-11286	IND	NA	India	IRIS 313-8453	IND	NA	India
IRIS 313-11287	IND	NA	India	IRIS 313-8454	IND	NA	Taiwan
IRIS 313-11289	ARO	NA	India	IRIS 313-8493	IND	T	Indonesia
IRIS 313-11295	AUS	NA	India	IRIS 313-8530	IND	T	India
IRIS 313-11297	ADM	NA	India	IRIS 313-8557	IND	T	Malaysia
IRIS 313-11298	AUS	NA	India	IRIS 313-8568	IND	T	India
IRIS 313-11301	IND	NA	India	IRIS 313-8571	IND	T	Tanzania
IRIS 313-11302	IND	NA	India	IRIS 313-8585	IND	NA	India
IRIS 313-11303	IND	NA	India	IRIS 313-8586	IND	T	Thailand
IRIS 313-11316	IND	T	Indonesia	IRIS 313-8595	IND	NA	Madagascar
IRIS 313-11321	IND	T	Bangladesh	IRIS 313-8606	IND	P	na
IRIS 313-11324	AUS	T	Bangladesh	IRIS 313-8627	JAP	I	United_States
IRIS 313-11338	IND	T	Philippines	IRIS 313-8641	AUS	P	Bangladesh
IRIS 313-11345	IND	T	Philippines	IRIS 313-8658	JAP	I	United_States
IRIS 313-11350	ARO	NA	India	IRIS 313-8659	IND	NA	Myanmar
IRIS 313-11351	IND	NA	India	IRIS 313-8660	IND	T	Sri_Lanka
IRIS 313-11358	IND	NA	India	IRIS 313-8665	JAP	I	United_States
IRIS 313-11370	IND	NA	India	IRIS 313-8681	IND	T	Guinea
IRIS 313-11372	IND	NA	India	IRIS 313-8687	JAP	T	Guinea-Bissau
IRIS 313-11386	IND	T	Thailand	IRIS 313-8690	JAP	NA	Vietnam
IRIS 313-11394	IND	T	Indonesia	IRIS 313-8703	IND	P	Bangladesh
IRIS 313-11395	IND	T	Indonesia	IRIS 313-8725	IND	T	Indonesia
IRIS 313-11416	IND	NA	India	IRIS 313-8745	JAP	NA	Haiti
IRIS 313-11431	IND	I	Philippines	IRIS 313-8751	IND	NA	Myanmar
IRIS 313-11435	JAP	T	Ivory_Coast	IRIS 313-8755	JAP	I	Japan
IRIS 313-11436	JAP	T	Ivory_Coast	IRIS 313-8768	JAP	T	Ivory_Coast
IRIS 313-11443	IND	T	India	IRIS 313-8803	JAP	I	United_States
IRIS 313-11453	IND	T	India	IRIS 313-8864	AUS	T	Bangladesh
IRIS 313-11460	IND	T	India	IRIS 313-8883	JAP	T	Malaysia
IRIS 313-11461	IND	T	India	IRIS 313-8909	IND	T	Tanzania
IRIS 313-11467	IND	T	Philippines	IRIS 313-8911	ARO	T	Thailand
IRIS 313-11472	IND	T	Philippines	IRIS 313-8923	JAP	I	United_States
IRIS 313-11477	AUS	NA	India	IRIS 313-8924	IND	T	India
IRIS 313-11483	AUS	T	Bangladesh	IRIS 313-8925	IND	T	Sri_Lanka
IRIS 313-11484	AUS	T	Bangladesh	IRIS 313-8930	IND	T	Bangladesh
IRIS 313-11489	AUS	NA	India	IRIS 313-8935	IND	NA	India
IRIS 313-11493	IND	T	India	IRIS 313-8940	IND	NA	China
IRIS 313-11513	IND	NA	Ecuador	IRIS 313-8948	IND	T	Philippines
IRIS 313-11515	IND	I	na	IRIS 313-8967	IND	NA	India
IRIS 313-11516	IND	I	Philippines	IRIS 313-8982	AUS	NA	India
IRIS 313-11521	IND	T	Vietnam	IRIS 313-8985	IND	T	Thailand
IRIS 313-11522	JAP	NA	China	IRIS 313-8986	AUS	T	India
IRIS 313-11528	IND	T	Ivory_Coast	IRIS 313-8988	IND	T	India

IRIS 313-11530	IND	T	Thailand	IRIS 313-9020	IND	T	Thailand
IRIS 313-11543	IND	NA	Myanmar	IRIS 313-9023	IND	P	India
IRIS 313-11546	IND	NA	Myanmar	IRIS 313-9039	IND	T	Sri_Lanka
IRIS 313-11547	IND	NA	Myanmar	IRIS 313-9048	JAP	T	Bhutan
IRIS 313-11555	IND	T	Sierra_Leone	IRIS 313-9066	IND	P	Bangladesh
IRIS 313-11567	ARO	T	Nepal	IRIS 313-9067	IND	T	Bangladesh
IRIS 313-11575	JAP	NA	China	IRIS 313-9112	IND	NA	Thailand
IRIS 313-11582	JAP	NA	China	IRIS 313-9116	IND	NA	Thailand
IRIS 313-11591	ADM	NA	Malaysia	IRIS 313-9117	IND	T	Indonesia
IRIS 313-11596	IND	NA	India	IRIS 313-9121	IND	T	Thailand
IRIS 313-11602	AUS	NA	India	IRIS 313-9131	IND	NA	Vietnam
IRIS 313-11604	AUS	NA	India	IRIS 313-9148	IND	P	Bangladesh
IRIS 313-11607	IND	NA	India	IRIS 313-9156	IND	I	Bangladesh
IRIS 313-11615	IND	T	Guinea	IRIS 313-9160	IND	NA	Senegal
IRIS 313-11617	AUS	T	India	IRIS 313-9182	IND	NA	Myanmar
IRIS 313-11618	AUS	T	India	IRIS 313-9198	IND	NA	Laos
IRIS 313-11622	IND	NA	China	IRIS 313-9228	JAP	P	Japan
IRIS 313-11624	IND	T	Nepal	IRIS 313-9262	IND	T	Bangladesh
IRIS 313-11626	ARO	T	Nepal	IRIS 313-9294	IND	NA	Gambia
IRIS 313-11630	ARO	T	Nepal	IRIS 313-9320	IND	T	Indonesia
IRIS 313-11635	IND	T	Thailand	IRIS 313-9324	IND	T	China
IRIS 313-11638	IND	NA	India	IRIS 313-9372	IND	NA	China
IRIS 313-11642	IND	NA	India	IRIS 313-9379	JAP	T	South_Korea
IRIS 313-11643	IND	NA	India	IRIS 313-9384	IND	T	India
IRIS 313-11644	IND	NA	India	IRIS 313-9406	IND	T	Thailand
IRIS 313-11645	IND	NA	India	IRIS 313-9409	IND	T	Malaysia
IRIS 313-11646	IND	NA	India	IRIS 313-9422	AUS	T	Bangladesh
IRIS 313-11647	IND	NA	India	IRIS 313-9427	IND	NA	India
IRIS 313-11648	IND	NA	India	IRIS 313-9449	AUS	T	Pakistan
IRIS 313-11651	JAP	NA	China	IRIS 313-9464	IND	I	Surinam
IRIS 313-11652	JAP	NA	China	IRIS 313-9469	IND	T	China
IRIS 313-11654	JAP	NA	China	IRIS 313-9470	JAP	T	Indonesia
IRIS 313-11655	JAP	NA	China	IRIS 313-9472	IND	NA	Sri_Lanka
IRIS 313-11656	IND	I	Indonesia	IRIS 313-9523	JAP	I	Japan
IRIS 313-11657	IND	NA	Nigeria	IRIS 313-9570	IND	NA	China
IRIS 313-11658	JAP	T	Sierra_Leone	IRIS 313-9590	IND	T	Indonesia
IRIS 313-11659	JAP	T	Sierra_Leone	IRIS 313-9594	IND	T	Bangladesh
IRIS 313-11661	JAP	T	Bhutan	IRIS 313-9602	IND	NA	Thailand
IRIS 313-11663	IND	T	Zimbabwe	IRIS 313-9605	IND	T	India
IRIS 313-11664	IND	NA	China	IRIS 313-9626	AUS	T	Bangladesh
IRIS 313-11665	IND	NA	China	IRIS 313-9701	JAP	I	Taiwan
IRIS 313-11666	IND	NA	China	IRIS 313-9790	JAP	NA	Uruguay
IRIS 313-11667	IND	NA	China	IRIS 313-9917	IND	T	Sri_Lanka
IRIS 313-11668	IND	NA	China	IRIS 313-9922	IND	I	South_Korea
IRIS 313-11669	IND	I	China	IRIS 313-9935	IND	NA	Guyana
IRIS 313-11671	IND	T	Nepal	IRIS 313-9936	IND	NA	Sri_Lanka
IRIS 313-11673	JAP	T	Philippines	IRIS 313-9937	JAP	NA	Italy
IRIS 313-11674	IND	T	Thailand	IRIS 313-9944	IND	NA	Solomon_Islands
IRIS 313-11677	IND	T	Thailand	IRIS 313-9961	JAP	NA	Norway
IRIS 313-11678	IND	T	Thailand	IRIS 313-9963	AUS	NA	Sri_Lanka
IRIS 313-11679	IND	T	Thailand	IRIS 313-9966	IND	P	Colombia
IRIS 313-11681	IND	T	Thailand	IRIS 313-9968	IND	T	Sri_Lanka
IRIS 313-11683	IND	T	Thailand	IRIS 313-9996	JAP	I	South_Korea

* Status: I, improved; T, traditional; P, breeding and inbred lines (promising line).

Table B.2: Traits used in this study.

Phenotype	Recoding	N	Mean	SD
Culm Diameter	None	608	1.62	0.49
Culm strength	Classes {1,2,3} recoded as {1}, classes {4:9} as {2}	642	1.39	0.49
Flag leaf angle	None	639	3.89	1.68
Grain length	None	641	8.62	1.02
Grain width	None	641	3.01	0.39
Leaf length	None	606	3.17	0.68
Leaf senescence	Classes {2:9} recoded as {2}	640	1.56	0.49
Grain weight	None	641	2.47	0.49
Salt injury	Classes {1:7} recoded as {1}, class {9} as {2}	602	1.46	0.49
Time to flowering	Log transformation	642	4.59	0.23
Panicle threshability	Classes {1:5} recoded as {1}, classes {6:9} as {2}	639	1.43	0.49

Table B.3: MITE family IDs from Castanera et al. (2021).

MITE family*	MITE type**	# TIPS	# Genic TIPs	Percentage of genic TIPs
MH63fam6_341	Tourist-like	3058	942	30.8
MH63fam8_344	Tourist-like	2087	657	31.5
MH63fam13_234	MITE-adh B-like	1166	475	40.7
MH63fam14_237	MITE-adh B-like	1771	603	34.0
MH63fam29_244	unclassified	1850	585	31.6
MH63fam32_236	MITE-adh B-like	2185	624	28.6
MH63fam47_235	MITE-adh B-like	3627	1312	36.2
MH63fam50_219	MITE-adh M-like	1207	471	39.0
MH63fam51_257	unclassified	696	238	34.2
MH63fam72_365	Amy/LTP-like	971	354	36.5
MH63fam73_259	unclassified	1003	292	29.1
MH63fam106_364	Castaway-like	1067	314	29.4
N22fam5_230	MITE-adh I-like	1093	354	32.4
N22fam30_347	Tourist-like	767	223	29.1
N22fam34_480	Ditto-like	2615	742	28.4
Oryza1fam20_279	Gaijin/Gaigin-like	2170	690	31.8
SE260500111fam211_334	Tourist-like	1019	301	29.5
SE260500112fam219_340	Tourist-like	2357	696	29.5

** Family IDs from Castanera et al., (2021)*

*** Classification based on best BLAST hit to Oryza Repeat Database (http://rice.uga.edu/annotation_oryza.shtml)*

Table B.4: Percentage of bootstrap samples where prediction correlation is larger with a given marker set than with SNPs only. Within Population Scenario.

Marker set	**MITE/DTX > SNP**		**RLX/RIX > SNP**		**ALL > SNP**	
	BayesC	**RKHS**	**BayesC**	**RKHS**	**BayesC**	**RKHS**
Culm Diameter	0.90	0.92	0.63	0.83	0.99	0.99
Culm strength	0.28	0.12	0	0.08	0.22	0.12
Flag leaf angle	0.05	0.07	0	0.09	0.06	0.09
Grain length	0.82	0.20	0	0.10	0.90	0.57
Grain width	0.36	0.05	0	0.05	0.64	0.36
Leaf length	0.67	0.82	0	0.69	0.50	0.85
Leaf senescence	0.52	0.52	0	0.79	0.64	0.69
Grain weight	0.70	0.81	0	0.55	0.76	0.59
Salt injury	0.53	0.32	0	0.34	0.34	0.41
Time to flowering	0.78	0.89	0	0.48	0.59	0.59
Pan. threshability	0.24	0.18	0	0.29	0.28	0.21

ALL: All marker model

Table B.5: Percentage of bootstrap samples where prediction correlation is larger with a given marker set than with SNPs only. Across Population Scenario.

Marker	**MITE/DTX > SNP**		**RLX/RIX > SNP**		**ALL > SNP**	
	BayesC	**RKHS**	**BayesC**	**RKHS**	**BayesC**	**RKHS**
Culm Diameter	0.83	0.92	0.18	0.37	0.95	0.81
Culm strength	0.79	0.82	0.83	0.89	0.81	0.81
Flag leaf angle	0.39	0.67	0.83	0.92	0.14	0.67
Grain length	0.02	0.67	0.02	0.22	0.68	0.84
Grain width	0.93	0.99	0.04	0.99	0.27	0.99
Leaf length	0.62	0.62	0.70	0.91	0.96	0.89
Leaf senescence	0.98	0.97	0.97	0.98	0.99	0.98
Grain weight	0.74	0.58	0.30	0.37	0.95	0.83
Salt injury	0.00	0.01	0.52	0.13	0.22	0.04
Time to flowering	0.19	0.73	0.09	0.41	0.83	0.94
Pan. threshability	0.06	0.15	0.02	0.01	0.23	0.18

ALL: All marker model

Table B.6: Correlation between observed and predicted phenotypes under a linear and threshold model.

Model		**Linear**			**Threshold**	
	SNPs	**MITE /DTX**	**RLX/ RIX**	**SNPs**	**MITE/ DTX**	**RLX/ RIX**
Culm Diameter	-0.07	0.26*	0.09	0.13	0.26*	0.06
Culm Strength	0.05	0.20*	0.18	-0.04	0.11	0.15
Flag Leaf Angle	0.00	0.11	0.14	0.10	0.13	0.26*

** Best strategy*

Table B.7: Root Mean Squared Error Value (RMSE): Within Population Scenario

Marker	**SNPs**		**MITE/DTX**		**RLX/RIX**		**ALL**	
Method	**BayesC**	**RKHS**	**BayesC**	**RKHS**	**BayesC**	**RKHS**	**BayesC**	**RKHS**
Culm Diameter	1.05	1.04	1.03	1.01*	1.06	1.04	1.03	1.02
Culm strength	0.99*	1.00	1.01	1.04	1.04	1.06	1.01	1.04
Flag leaf angle	1.04*	1.05	1.14	1.14	1.14	1.13	1.06	1.08
Grain length	0.79	0.82	0.76*	0.85	0.87	0.86	0.77	0.82
Grain width	0.43*	0.51	0.49	0.63	0.60	0.63	0.43*	0.53
Leaf length	0.97	0.98	0.97	0.97	1.00	0.97	0.97	0.96*
Leaf senescence	0.89	0.90	0.89	0.90	0.89	0.86*	0.88	0.88
Grain weight	0.82	0.81	0.79	0.78*	0.81	0.79	0.81	0.80
Salt injury	0.95*	0.96	0.96	0.96	0.97	0.96	0.96	0.95*
Time to flowering	0.61*	0.61*	0.69	0.65	0.66	0.68	0.61*	0.62
Pan. threshability	0.96	0.95*	0.98	0.97	0.97	0.95*	0.97	0.96

*Asterisk * indicates the lowest value*
ALL: All marker model

Table B.8: Root Mean Squared Error Value (RMSE): Across Population Scenario

Marker	**SNPs**		**MITE/DTX**		**RLX/RIX**		**ALL**	
Method	**BayesC**	**RKHS**	**BayesC**	**RKHS**	**BayesC**	**RKHS**	**BayesC**	**RKHS**
Culm Diameter	0.96	0.97	0.94*	0.95	0.98	0.98	0.95	0.96
Culm strength	1.00	1.01	0.98*	0.99	0.98*	0.98*	0.98	1.01
Flag leaf angle	0.96	0.98	0.95	0.96	0.92*	0.93	0.96	0.97
Grain length	1.17*	1.41	1.40	1.39	1.42	1.46	1.17*	1.38
Grain width	1.24	1.29	0.99*	1.14	1.32	1.18	1.28	1.23
Leaf length	0.89	0.89	0.88	0.89	0.87	0.84*	0.86	0.86
Leaf senescence	1.02	0.99	0.88*	0.88*	0.91	0.88*	0.89	0.89
Grain weight	1.12	1.23	1.10*	1.11	1.17	1.14	1.11	1.11
Salt injury	0.96	0.97	1.00	0.99	0.95*	0.97	0.96	0.98
Time to flowering	0.79	0.78	0.76	0.77	0.79	0.76	0.78	0.75*
Pan. threshability	0.96*	0.98	0.99	0.99	0.99	1.02	0.97	0.99

*Asterisk * indicates the lowest value*
ALL: All marker model

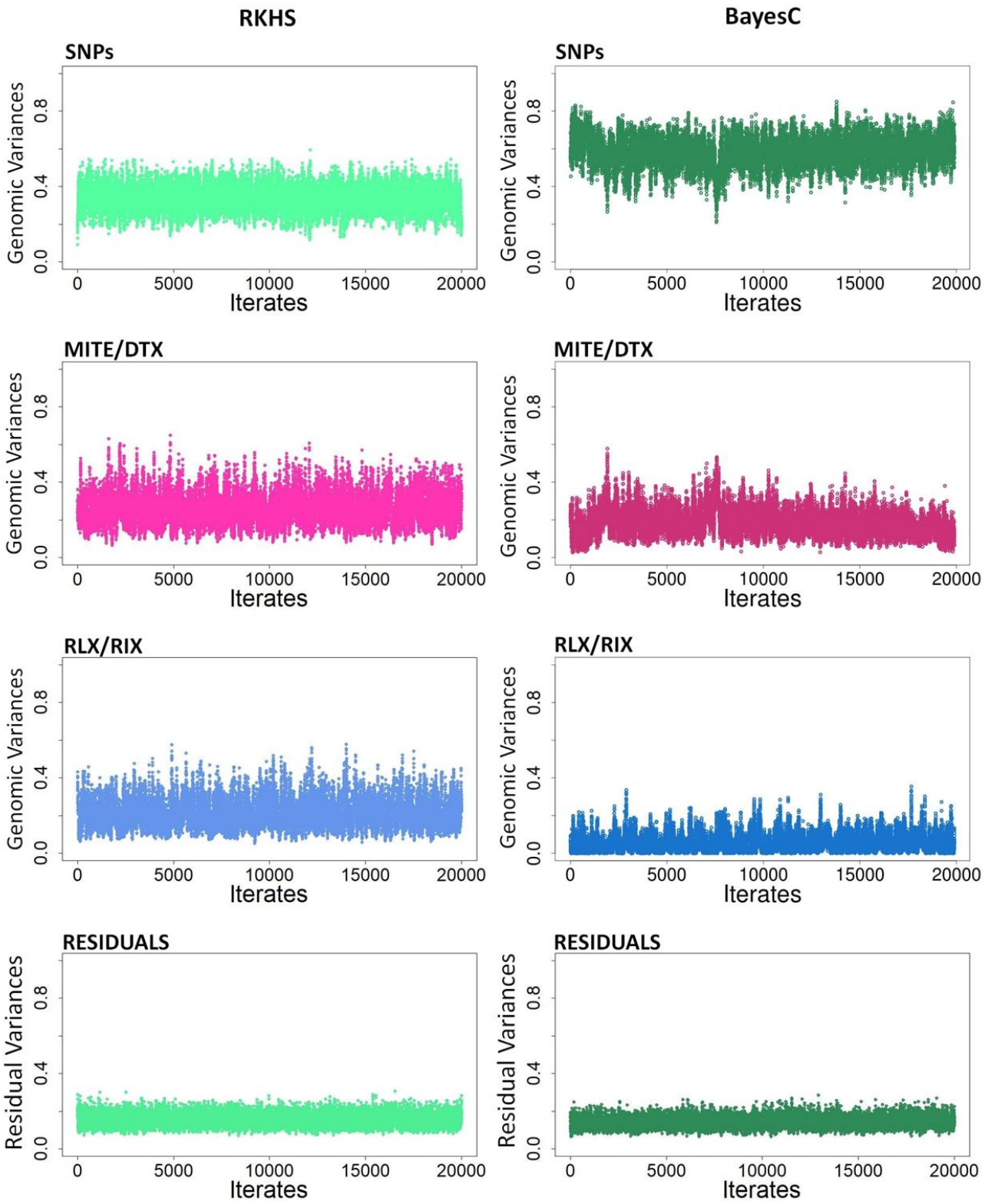

Figure B.1: Plot of variances across iterations to show convergence. Results correspond to "Grain width" under models 1a and 2a. Variances with Bayes C were computed as in https://github.com/gdlc/BGLR-R/blob/master/inst/md/heritability.md.

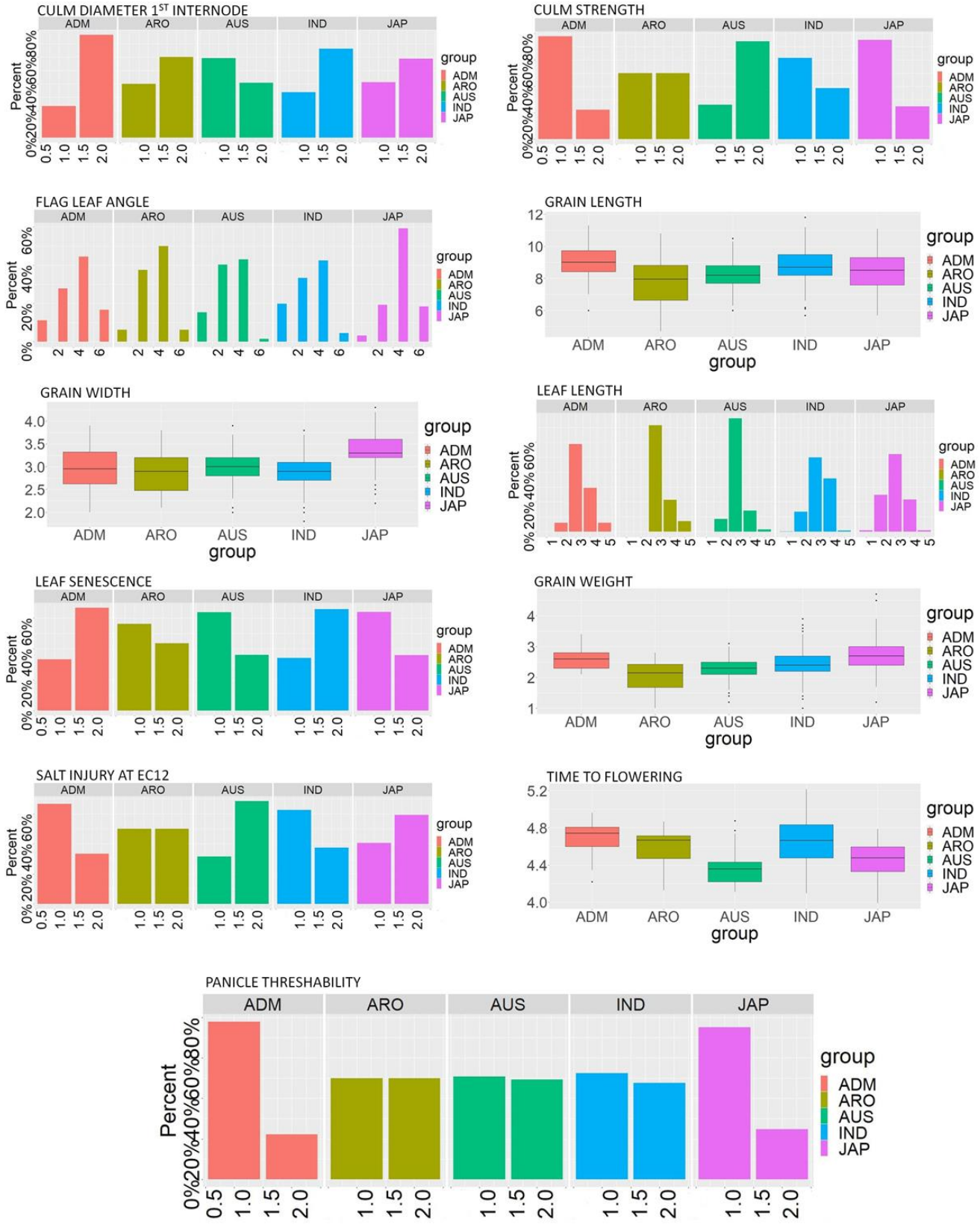

Figure B.2: Raw phenotypic distributions by populations, each shown in a different color.

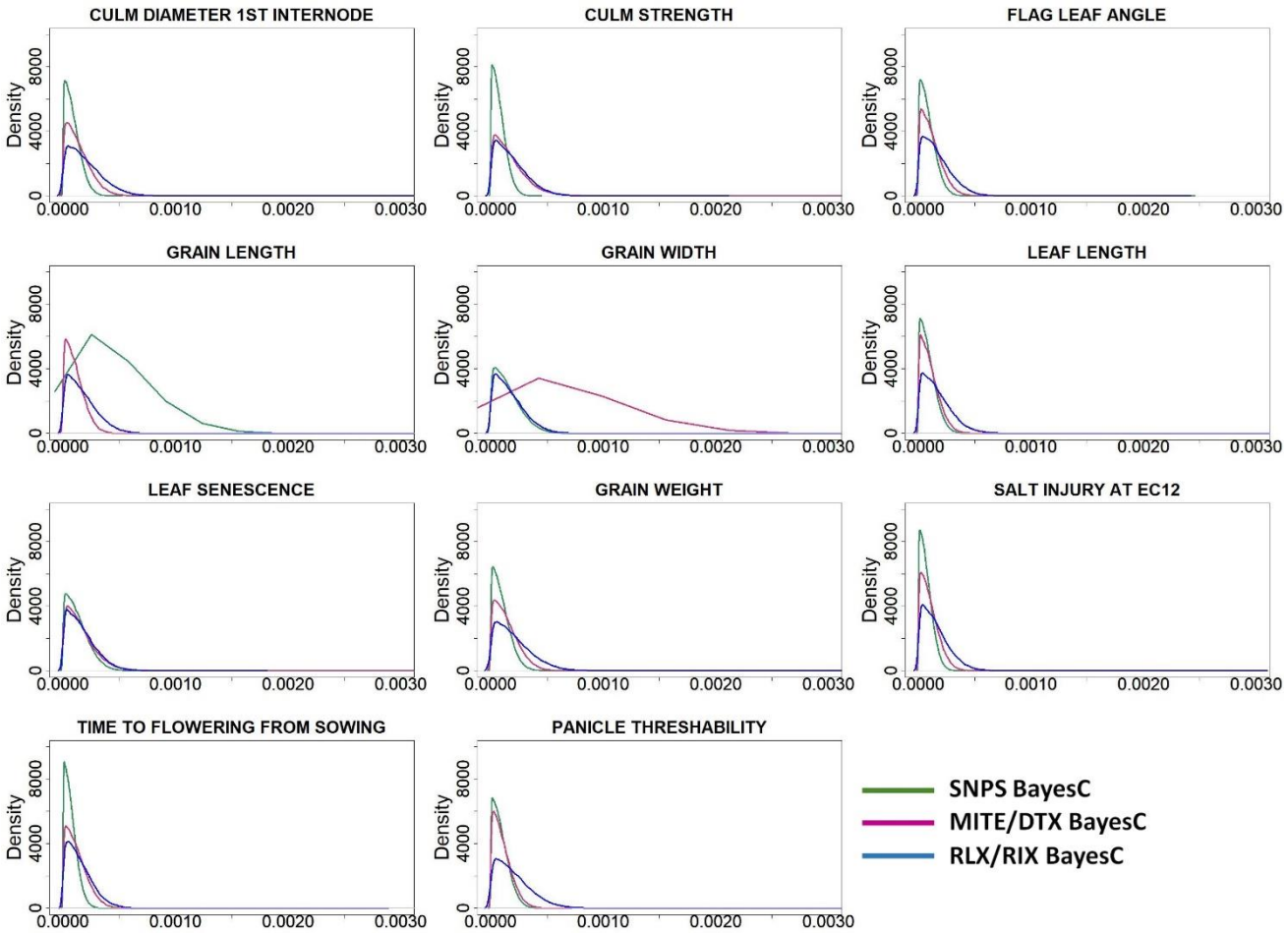

Figure B.3: Distributions of estimated marker effects from Bayes C using model 2a in the across population prediction scenario.

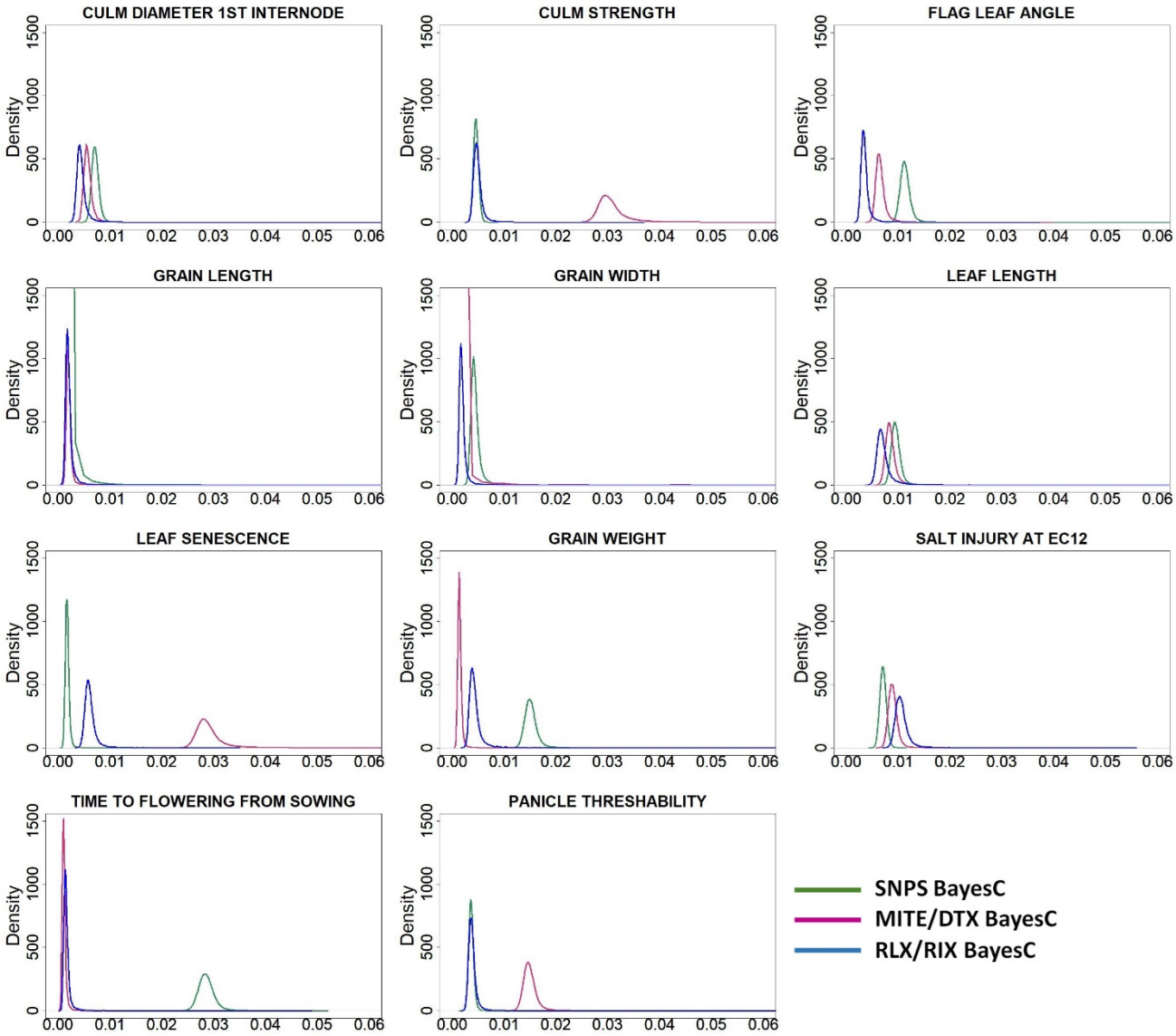

Figure B.4: Distributions of marker probabilities entering the model (d) in Bayes C using model 2a in the across population prediction scenario.

Acknowledgements

I would like to express my gratitude to those who have contributed to my professional and personal life in the last few years.

At first, I would like to thank my supervisor Dr. Sebastián E. Ramos-Onsins for giving me the opportunity to conduct my research in a such an interesting topic and trusting me even though I was coming with a totally different background. He was always willing to explain me any question, as silly it might be and to dedicate time on me.

Second, I would like to thank my supervisor Dr. Miguel Pérez-Enciso for his determining contribution to my thesis. He was always very effective, to the point giving solutions in a short period of time which for a PhD thesis is always essential.

Thank you both for all the time we spent together, for all the support and guidance.

Then, I would like to thank Prof. Josep Casacuberta and Dr. Raul Castanera for the collaboration I had with them. It has been a pleasure to work with them and I am grateful for all the help received.

Also, I would like to thank Dr. David Castellano for his guidance and help through the project we have worked together. I would like to thank Prof. Ryan Gutenkunst for his essential and always efficient contribution.

I would like to thank all my colleagues, old and news at Center for research in agricultural Genomics (CRAG) for the nice and enjoyable time I had with them. A nice working environment depends on the colleagues, and I found myself very lucky to be in this kind of environment.

I would like to thank my dear friends Charikleia, Dev, Vera and Anibal for all your friendship and support.

My dear parents and my brother for their unconditional love. I am forever grateful, and I could not have asked for better.

My dear grandfather who always encouraged knowledge and he would be very proud if he was here.

Last, but certainly not least, I would like to thank my husband Nikos for all his support, love and understanding. Thank you for all the leaps you pushed me to take while you were always there to watch my back. I could not be luckier, and I cannot wait to tackle the next chapter of our journey together.

Publications related to the thesis

The following publications have been published or are under preparation as part of this thesis:

- Ioanna-Theoni Vourlaki, Raúl Castanera, Sebastián E. Ramos-Onsins, Josep M. Casacuberta, Miguel Pérez-Enciso, "Transposable element polymorphisms improve prediction of complex agronomic traits in rice", published in Theoretical and Applied Genetics, 2022, https://doi.org/10.1007/s00122-022-04180-2.

- Ioanna-Theoni Vourlaki, David Castellano, Ryan N. Gutenkunst, Sebastian E. Ramos-Onsins, "Detection of Domestication Signals through the Analysis of the Full Distribution of Fitness Effects using Forward Simulations and Polygenic Adaptation". Under revision for journal publication, 2022, DOI: 10.1101/2022.08.24.505198.

- Ioanna-Theoni Vourlaki et al. Merging structural and nucleotide genome-wide variation for genomic prediction in rice (under preparation).

CURRICULUM VITAE

EDUCATION

2005 - 2012	Diploma of Physics, University of Patras Thesis: Chest Computational Tomography and Image Noise Analysis
2013 - 2016	MSc. School of Electronic & Computer Engineering, Technical University of Crete Thesis: Self-organized clustering of big data: Application on the extraction of cervical cancer classes from backscattering light curves Supervisor: Prof. Michalis Zervakis
2017-	PhD student Genetics Department, Universitat Autonoma de Barcelona, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB Thesis: Detecting signals of polygenic variability in domestication and in breeding. Supervisors: Dr. Sebastián E. Ramos Onsins & Prof. Miguel Pérez-Enciso

WORK EXPERIENCE

May 2012 – August 2012	Intership with Institute of Nuclear & Radiological Science & Technology, Energy and Safety "Demokritos" National Centre for Scientific Research, Athens. Title: Digital Signal Processing and Real Time Analysis of its Rate of Change during Radiation. Supervisor: Dr. George Apostolopoulos
2013 – 2016	Researcher, Digital Image and Signal Processing Laboratory, School of Electronic & Computer Engineering, Technical University of Crete
November 2017 – May 2018	Researcher, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Supervisor: Dr. Sebastián E. Ramos Onsins, Prof. Miguel Perez-Enciso

RESEARCH PROJECTS

February 2015 – September 2015: THALES (CYBERSENSORS – High Frequency Monitoring System for Integrated Water Resources Management of Rivers) project funded by the NSRF2007-13 of the Greek Ministry of Development.

TECHNICAL & COMPUTER SKILLS

Technical Skills: Mathematical Modeling on Population Genomics, Simulations, Statistical Inference, Probability Theory, Bioinformatics, Machine Learning (Unsupervised & Supervised Learning), Pattern Recognition, Data Mining, Dynamic Spectral Imaging, Cancer Diagnosis, Bootstrapping, Time series Analysis, Big Data Organization and Analysis, Distance Metrics, Deep Learning (TensorFlow and Keras), Genomic Prediction, Bayesian Linear Regression, Plant Breeding.

Computer Skills: Experienced user of R, Python, C++, Matlab

LANGUAGE SKILLS

Fluent: English (Certificate of Competence in English, IELTS)

Basic level: Spanish

Mother Tongue: Greek

COURSES/CONFERENCES ATTENDED

Course about "Manipulation of NGS data for Genomic and Population Genetic Analyses": https://www.transmittingscience.org/courses/genetics-and-genomics/manipulation-ngs-data-genomic-population-genetics-analyses/, 2018

"Short course on Basics on Coalescent Theory and Simulations using R" (by Sebastian Ramos-Onsins):https://bioinformatics.cragenomica.es/numgenomics/people/sebas/teaching/teaching.html, 2018

Short course on "Basics on Molecular Evolution and Phylogeny using R" (by Sebastian Ramos-Onsins): https://bioinformatics.cragenomica.es/numgenomics/people/sebas/teaching/teaching.html, 2018

Short Course on “Occupational Risk Prevention in Laboratories”, held in Centre for Research in Agricultural Genomics (CRAG), 20018

Participation on “Data club group”, held in Centre for Research in Agricultural Genomics (CRAG), 2018-2019

Member of “Society for the study of Evolution”, 2020

https://www.evolutionsociety.org/

Member of “Society for Molecular Biology & Evolution”, 2020

https://www.smbe.org/smbe/

Course: Phylogenomics and population Genomics: Inference and Applications (summer 2020):

https://www.ub.edu/certfem/ppgcourse/

6th International Conference on Quantitative Genetics Virtual, Australia, 3 November – 13 November 2020.

International Conference on AL applications in agriculture, 19-20 July 2022, Barcelona.

CONFERENCES-SEMINARS

13th IEEE International Conference on Bioinformatics and Bioengineering, Chania, Greece, November 10-13, 2013

IEEE Imaging Systems and Techniques (IST) Conference, Chania, Greece, 4-6 October, Presentation, 2016

Second Scientific Conference of Students of the Doctoral Program in Genetics, UAB, Barcelona, Presentation, 2018

Third Scientific Conference of Students of the Doctoral Program in Genetics, UAB, Barcelona, Presentation, 2019

Presentation of “Detection of Signals on Adaptive Polygenic Traits on Genomic Variability Patterns: The case of Domestication” at Seminar held in Centre for Research in Agricultural Genomics (CRAG), 2019.

Popgroup54 virtual on-line conference, Liverpool, 4 January – 6 January, Poster, 2021.

Society for Molecular Biology and Evolution, virtual on-line conference, Liverpool, 4 July – 8 July, Poster, 2021.

International Symposium on Rice Functional Genomics, Barcelona, 3 November - 5 November Poster, 2021.

AWARD

Best Student Paper Presentation Award, *IEEE Imaging Systems and Techniques (IST) Conference Proceedings*, Chania, Greece, 4-6 October, 2016

PUBLICATIONS

1. I. Vourlaki, G. Livanos, M. Zervakis, C. Balas, G. Giakos, "Spectral Data Self-organization Based on Bootstrapping and Clustering Approaches", *IEEE Imaging Systems and Techniques (IST) Conference Proceedings*, Macau, China, 16-18 September, 2015, DOI:10.1109/IST.2015.7294546.

2. I. Vourlaki, G. Livanos, M. Zervakis, T. Giakoumakis, G. Giakos, C. Balas, "Recursive-Mode K-means Clustering for Self-organization of Dynamic Imaging Data", *IEEE Imaging Systems and Techniques (IST) Conference Proceedings*, Chania, Greece, 4-6 October, 2016, https://doi.org/10.1109/IST.2016.7738197.

3. I. Vourlaki, G. Livanos, M. Zervakis, C. Balas, G. Giakos, "Bootstrap Clustering Approaches for Classification of Large Data: Application in Cervical Cancer Staging", *Biomedical Signal and Processing Control*, Vol. 49, March 2019, pages 2063-2073, https://doi.org/10.1016/j.bspc.2018.12.014.

4. Ioanna-Theoni Vourlaki, Raúl Castanera, Sebastián E. Ramos-Onsins, Josep M. Casacuberta, Miguel Pérez-Enciso, "Transposable element polymorphisms improve prediction of complex agronomic traits in rice", published in Theoretical and Applied Genetics, 2022, https://doi.org/10.1007/s00122-022-04180-2.

5. Ioanna-Theoni Vourlaki, David Castellano, Ryan N. Gutenkunst, Sebastian E. Ramos-Onsins, "Detection of Domestication Signals through the Analysis of the Full Distribution of Fitness Effects using Forward Simulations and Polygenic Adaptation". Under revision for journal publication, 2022.