TECHNISCHE UNIVERSITÄT MÜNCHEN
Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Professur für Populationsgenetik

Inference of the Demographic History of Domesticated Species Using Approximate Bayesian Computation and Likelihood-based Methods

Florence Mathilde Valérie Parat

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. Nat.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Donna Ankerst

Prüfende/-r der Dissertation:
1. Prof. Dr. Aurélien Tellier
2. Prof. Dr. Chris-Carolin Schön

Die Dissertation wurde am 14.08.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 12.02.2020 angenommen.

À mon grand-père et à la mémoire de ma grand-mère,

les docteurs Guy et Monique Parat.

Acknowledgments

I would like to express my deepest gratitude to Prof. Aurélien Tellier, my supervisor, for his enthusiasm, kind encouragements and useful critics of my work. Thank you for the trust that you put in me, the freedom that you gave me to pursue my own ideas, at my own pace, and your understanding even when that pace was very slow.

I am also very grateful to the members of my thesis committee, Prof. Wolfgang Stephan (LMU) and Dr. Fabian Freund (University of Hohenheim), for their patient explanations, insightful comments and for their questions, which pushed me to always deepen my understanding of population genetics. I wish to thank Prof. Chris-Carolin Schön and Prof. Donna Ankerst for agreeing to be part of my examining committee.

I would like to thank all my colleagues at the population genetics group (Hanna Märkle, Saurabh Pophaly, Daniela Scheikl, Remco Stam, Mélissa Vérin, Amaryllis Vidali, Daniel Živković) for their help, useful discussions and, sometimes much needed, moral support. Many thanks to Silke Bauer, who saved me more than once from my paperwork anxiety and simply makes life easier for everyone in the group. Each one of you taught me things, made me laugh, made me think and cheered me up at one point or another during my PhD, so thank you, I would not have made it through without you. A special thanks to Dr. Sidonie Bellot for her help with preparing courses and teaching, and shared reflection on academia life…

I would further like to thank Prof. Chris-Carolin Schön, Dr. Eva Bauer and Dr. Grit Haseneyer/Schwertfirm for sharing their data and actively collaborating with me on the

rye project (Chapter 3 and 4). I wish to acknowledge as well Prof. Rüdi (Hans-Rudolf) Fries and Dr. Hubert Pausch for kindly sharing their data (see Chapter 6).

I am grateful to the Synbreed network for founding my research and, equally importantly, for the great people I met through the intense scientific meetings and workshops we had together, and who gave me a window into the world of both academia and the industry R&D world. I am also grateful for the opportunity it gave me to travel to the University of Fribourg and start a collaboration with Prof. Daniel Wegmann and later Prof. Sándor Szilagyi, who I both thank very much for their help with the project in Chapter 5 and 6. Thank you Sándor for your openness, for teaching me so much with patience and valuing my ideas.

I want to thank my parents for their unconditional love and support, and for passing on to me their curiosity, the wish to better understand the world which put me on this path and their sense of humor for the many times that path seemed too steep. Iain Hunter, thank you with all of my heart, for your kindness, for believing in me even when I did not and for your unwavering support throughout this long journey; I did not always make it easy!

Last but not least, I thank my sister and my brother, all my family and friends (especially Manuel Glaser and Kathrin Jung, flat-mate and office neighbor but most importantly great friends though all the ups and downs of PhD life), and my new colleagues for their support during these last years.

Abstract

The genetic makeup of all populations of animals or plants have been forged by their own unique demographic history. However among these populations, domesticated species show some common particularities. A good understanding of the genetic diversity of domesticated species is needed to maximize the selection efficacy of breeding programs. In the first part of this thesis we study the genetic diversity, genetic structure and demographic history of cultivated rye (*Secale cereale* L.). We genotyped 620 individuals from 14 global rye populations with a different end use (grain or forage) at 32 genome-wide simple sequence repeat markers. We reveal the relationships among these populations, their sizes and the timing of domestication events using population genetics and model-based inference with approximate Bayesian computation. Our main results demonstrate (i) a high within-population variation and genetic diversity, (ii) an unexpected absence of reduction in diversity with an increasing improvement level and (iii) patterns suggestive of multiple domestication events. We suggest that the main drivers of diversification of winter rye are the end use of rye in two early regions of cultivation: rye forage in the Mediterranean area and grain in northeast Europe. The lower diversity and stronger differentiation of eastern European populations were most likely due to more intensive cultivation and breeding of rye in this region, in contrast to the Mediterranean region where it was considered a secondary crop or even a weed. We discuss the relevance of our results for the management of gene bank resources and the pitfalls of inference methods applied to crop domestication due to violation of model assumptions and model complexity.

In a second part of the thesis, we focus on animal domesticates for which pedigree data is collected. We show that, whereas genetic data is only informative about the population mutation rate, the product of the effective population size times the mutation rate, and not about these quantities individually, this hurdle can be overcome by combining genetic data with pedigree information. To successfully use pedigree data, however, important aspects of real populations such as the presence of two sexes, unbalanced sex ratios and overlapping generations have to be taken into account. We present here an extension of the classic Wright-Fisher model accounting for these effects and show that the coalescent process under this model reduces to the classic Kingman coalescent with specific scaling parameters. We further derive the probability of a pedigree under that model and show how pedigree data can thus be used to infer demographic parameters. We present a computationally efficient Markov chain Monte Carlo-based inference approach, combining pedigree information and genetic data summarized by the site frequency spectrum that allows for the joint inference of the mutation rate, sex-specific population sizes and the fraction of overlapping generations. Using simulations we then show that these parameters can be accurately inferred from pedigrees spanning just a few generations, as are available for many species. We finally apply the pedigree part of the model to a real cattle pedigree from the Fleckvieh breed. We finally discuss future possible extensions of the model and the inference framework necessary for applications to wild and domesticated species, among others, accounting for more complex demographies and the uncertainty in assigning pedigree individuals to specific generations.

Zusammenfassung

Die genetische Beschaffenheit von Tier- und Pflanzenpopulationen wird durch deren einzigartige demographische Entwicklung bestimmt. Um die Selektionswirksamkeit von Zuchtprogrammen zu steigern, ist ein umfangreiches Wissen der genetischen Diversität erforderlich. Diese Thesis vermittelt Einblicke bezüglich der genetischen Diversität, Struktur und demographischen Entwicklung von Roggen (*Secale cereal* L.). Dies geschieht durch die Verwendung der ABC auf 32 SSR Markern, in 14 Populationen mit verschiedenem Endverbrauch (Getreide bzw. Futtermittel). Der zweite Teil der Thesis beschäftigt sich mit domestizierten Tierarten. Wir ermittelten, dass die effektive Populationsgröße und Mutationsrate durch die Kombination aus genetischen und Stammbaum-Daten, getrennt betrachtet werden können. Zu diesem Zweck erweiterten wir das klassische Wright-Fisher-Modell, um zwei Geschlechter, unausgeglichene Geschlechterverhältnisse, sowie sich überschneidende Generationen zu berücksichtigen. Letztlich entwickelten wir daraus einen Inferenzansatz, der in Generationen, in denen Stammbaumdaten fehlen, ein neu skaliertes n-Koaleszenzmodell verwendet.

Résumé

La composition génétique des populations animales et végétales est forgée par leur histoire démographique. Comprendre la diversité génétique des espèces domestiquées est essentielle pour optimiser l'efficacité des programmes de sélection. La première partie de cette thèse présente une étude de la diversité génétique, de la structure et de l'historique démographique du seigle (*Secale cereal* L.) à l'aide de *approximate Bayesian computations* sur 32 marqueurs SSR dans 14 populations utilisée à des fins différentes (grain ou fourrage). La deuxième partie de la thèse porte sur les animaux domestiques. Nous montrons qu'il est possible d'estimer séparement la taille effective de la population et le taux de mutation en combinant des données génétiques au pedigree. À cette fin, nous étendons le modèle classique de Wright-Fisher à deux sexes, des sex-ratios deséquilibrés et des générations chevauchantes. Nous développons une approche d'inférence utilisant un n-coalescent mis à l'echelle dans les générations hors du pedigree.

Contents

List of Figures

List of Tables

Authors' contributions

Analyses and results from chapter 3 and 4 are published in Parat *et al.* (2016).

Chapter 3

G.S., E.B. and C.-C.S. designed the study. T.M. contributed resources. F.P., G.S. and U.R. performed statistical analysis. All authors contributed to writing the manuscript.

Chapter 4

F.P. and A.T. designed the study and wrote the manuscript. F.P. performed statistical inference, and analyzed the results.

Chapter 5

F.P., A.T. and D.W. designed the study. F.P. created and implemented the inference model, and wrote the manuscript.

Chapter 6

F.P. designed the study, performed the simulations and statistical inference, analyzed the results and wrote the manuscript.

Parat, F., G. Schwertfirm, U. Rudolph, T. Miedaner, V. Korzun, E. Bauer, C.-C. Schön, and A. Tellier, 2016 Geography and end use drive the diversification of worldwide winter rye populations. Molecular Ecology **25**: 500-514.

Chapter 1

General introduction

1.1 Domestication

1.1.1 Definition of domestication

Domestication is a relationship between two populations of different species, in this thesis, humans, the domesticator, and a plant or animal population of interest, the domesticate. There is little agreement on the characterization of this relationship. While some authors mostly stress the human side of the relationship, highlighting the intentionality of domestication and the control over reproduction and every other aspects of the domesticates' life cycle (e.g., Cauvin 2000), other authors put more weight on the other partner in that relationship, underlining Darwin's concept of *unconscious selection* and the evolutionary mechanisms that allows the plant or animal to take advantage of domestication to increase its fitness benefit (Rindos 1984). The third way is to see domestication as a balance between the benefits of both the domesticator and the domesticate. In this sense, Zeder (2015) defined domestication as ”a sustained multigenerational, mutualistic relationship in which one organism assumes a significant degree of influence over the reproduction and care of another organism in order to secure a more predictable supply of a resource of interest, and through which the partner organism gains advantage over individuals that remain outside this relationship, thereby benefiting and often increasing the fitness of both the domesticator and the target domesticate.”

1.1.2 The domestication syndrome

Domestication often entails morphologic (phenotypic) changes. Some of these modified traits, such as larger seeds or fruits, apical dominance, and loss of dispersal mechanisms, are found in almost every domesticated plant species. They are known under the name *domestication syndrome* (Hammer 1984).

The evolution of traits due to domestication can be the result of intentional selection by humans but often started as unconscious selection (Smith 2006). For example, in the case of grains, early farmers were sowing part of the harvested grain, thus, giving a competitive advantage to the grain that simultaneously reached maturity, did not shatter or fall from the ear, and could therefore be harvested and potentially sown (Hillman and Davies 1990). It is still debated whether these morphologic changes are crucial to domestication, and can therefore be used as archaeological markers, or appeared slowly after long periods of cultivation as a side effect of the use of a particular agricultural technique as using a sickle to harvest grain (Balter 2007; Fuller and Allaby 2009).

In mammals, some traits also occur in association with domestication independently in many species such as de-pigmentation patches, reduced facial skeleton, floppy ”lop” ears, or smaller brain size. Although these changes are suspected to be related to selection for docility and to arise as a symptom of modified neural crest cell migration or multiplication (Wilkins *et al.* 2014), the exact mechanisms relating morphological traits and domestication, and even more so, their genetic basis are still uncertain.

1.1.3 The genetic impact of domestication: demography versus selection

One of the first impact of domestication on the plant or animal partner is the modification of the selective pressures acting on the domesticate. In most cases, the domesticators are not intentionally applying a selective pressure on a particular trait but rather, by their actions, relax the selective pressures which typically applies to the free-living organisms (e.g., predation or water access) and shifts the selection to new factors arising from their relationship (e.g., docility).

By sampling a few individuals from the wild and implanting them in new environments, domestication also strengthens drift by creating bottlenecks and potential founder effects in domesticated populations (Tenaillon *et al.* 2004). These events are part of what population geneticists call the demographic history or, simply, demography of the population. Demography can be defined as the set of changes that occur over a period of time in populations mostly with reference to size, geographical distribution and migration.

The demographic events linked to domestication lead to a loss of selection efficacy and an increase of drift. Alleles get randomly fixed accelerating further the genetic, and sometimes morphologic, differentiation of the domesticates compared to the wild population (Glémin and Bataillon 2009).

Population geneticists have developed a multitude of tools and models to study the effect of genetic drift, mutation, migration and natural selection on the patterns of genetic diversity. Many of these tools have been applied to domesticated populations with success.

1.1.4 Further steps of improvement

Deliberate breeding to encourage specific traits appears often later in the domestication process and leads to the development of so called *improvement traits*. Improvement of the domesticates' characteristics that suit the needs of humans is a continuous process but technological steps can modify the speed and the genetic implication of this process. For example, in plants the term *landraces* has many definitions but it is often used to describe a diverse but distinct population that acquired its specificities through isolation, local adaptation and unconscious selection or very basic mass selection. Later on, with the generalization of international breeding companies and, even later, gene banks and marker assisted selection (MAS), so called *varieties* were defined. These varieties usually present a lower level of genetic diversity over the whole genome. In the most extreme cases, there is no diversity left and varieties are completely homozygous (e.g., lettuce) or diversity is limited to fixed differences between completely homozygous parents (e.g., maize). However varieties can also contain small chromosomal regions of high diversity due to the recent introgression of favorable traits from landraces or wild relatives.

In the case of animals, breeds were defined very early on based on the morphological appearance of the animal and close monitoring of reproduction. These breeds were first obtained by local crossings and improved using mass selection. But, similarly to the case of plants, with the increased use of techniques such as artificial insemination or MAS, less weight is put on the visible traits of the animals but rather on the performance. As a result, the reproductive population size within breeds has decreased leading to a decrease of genetic diversity whereas the census size of some commercially successful breeds have strongly increased.

1.2 Inferring demography in domesticated species

1.2.1 From history to the detection of domestication genes

This thesis consists of two main parts with one common purpose: the study of demographic history through genetics in domesticated species. Understanding these species' demography in terms of genetics offers a window into the past that is complementary to archaeological studies. The ancestors of the genetic material sampled at present time only represents a small part of the domesticated individuals. Conversely, plant or animal rests found by archaeologists are not necessarily actual ancestors of our current crops or livestock (Gross and Olsen 2010). These differences explain some of the discrepancies between genetic and archaeological results, or even paleogenomics results such as the study of ancient maize DNA (Jaenicke-Després *et al.* 2003). While archeology can help trace the historical spread of a crop, knowledge of the genetic structure of contemporary populations can shed light on other aspects of the domestication history, such as the development of different cultivation methods or crop end uses, and their impact on the current genetic diversity.

Studying the neutral demography of domestication also brings insights into the relative weight of random genetic drift and selection. Indeed, demographic events and increased genetic drift due to domestication can leave traces in the genome easily confused with the ones of selection. The inferred demographic models can serve as a null model to detect sites under selection and find candidate genes involved in domestication and improvement (Tenaillon *et al.* 2004).

Last but not least, a better qualification of available genetic material, such as population structure and effective population size are necessary for a better population management in breeding programs. This knowledge helps to maintain the genetic diversity of a species and allows a better use of this material for breeding and improvement, fully exploiting the genetic potential of populations.

1.2.2 Cultivated rye

The first part of the thesis is focused on inferring demography in a plant domesticate, cultivated rye (*Secale cereale* L.). Scientific literature describing the history of rye is scarce compared to cereals of higher economic importance. However, rye constituted in the past an important resource and still does in several regions of northern and eastern Europe.

Rye is used as a grain crop for bread making, brewing, distilling and animal feed and as a forage crop in the form of green chop, pasture, green manure or haylage (Miedaner 2010). Modern rye breeding populations, referred to as varieties in this thesis, are adapted to two main end uses, grain or forage, or to a mixture of both; however the impact of use on genetic diversity and structure remains unclear.

From an archaeological perspective, there is a general agreement that cultivated rye originated from the Fertile Crescent (Khush 1963). Wild relatives or weedy forms reached Europe probably through a northern route, and remains were present at archaeological sites dating to the late Neolithic age in Poland and Romania, to the Bronze Age in the Czech Republic, Slovakia and Ukraine and to the Iron Age in Germany, Denmark and Crimea (see references in Zohary *et al.* 2013). Studies on archaeological remains indicated that rye most likely spread as a weed among wheat and barley fields throughout Europe (Behre 1992). Therefore, the first domestication of rye most probably happened through conscious or unconscious selection by early Neolithic farmers, around 4,500 BC (Behre 1992; Khush 1963).

A number of research studies have investigated the genetic diversity and structure of rye populations. Based on different marker systems and plant material sets, previous studies identified three factors that led to a clustering of rye populations: spring or winter growth habit (Ma *et al.* 2004), population history (geographical origin, relatedness and dispersion routes) and level of improvement (Persson *et al.* 2001). Indeed, because landraces are old, locally adapted populations that underwent little or no mass selection, they usually present higher levels of diversity than the officially registered varieties that are usually more recent (later than the 1940s for rye) and are products from breeding cycles (Villa *et al.* 2005). However, the literature also contains conflicting results concerning expected levels of genetic diversity between populations. For example, contrary to a study on Portuguese rye populations by Matos *et al.* (2001), Ribeiro *et al.* (2012) suggest that the diversity of Portuguese populations was higher than the diversity of northern Europe (Persson and Von Bothmer 2002). Additionally, as expected in most crops but contrary to what has been shown in previous studies (e.g., Persson *et al.* 2001), Ribeiro *et al.* (2012) found higher levels of diversity in rye landraces than in varieties. To address these discrepancies in the literature, we analyzed rye population samples over a wide geographical range using simple sequence repeat (SSR) markers. In addition to assessing the genetic diversity and structure of several populations of grain, forage and weedy rye, we use this information to infer their history.

1.2.3 Pedigree records in animal domesticates

In the second part of the thesis, we build a new framework for the analysis of genetic data from populations with available individual-based pedigrees. Modern population genetics is primarily based on coalescent theory (Wakeley 2009), which assumes no prior information on the true (parent-offspring) relationship between genetic lineages. We postulate that the pedigree of a sample contains information about the demography of the population at least partially complementary to the information contained in the genetic data. A method exploiting the full information should therefore improve the inference of demographic parameters. Indeed, several methods have been proposed to use pedigree information to infer demographic processes using the increase of inbreeding over time under a given reproduction model (Falconer and Mackay 1996; Gutiérrez *et al.* 2008). Additionally, a method that allows the simulation of pedigrees under a given demographic and reproductive model and draws genealogies inside these pedigrees was developed (Gasbarra *et al.* 2005), and could be used, in an ABC framework, for inference. However, there is currently no general inference framework for such data.

Pedigree information is available for many populations or species, in particular for managed populations under conservation management or domesticated animals under active breeding (e.g., Clutton-Brock *et al.* 1982; Ellegren 1999; Cunningham *et al.* 2001; Mc Parland *et al.* 2007). Yet many of these species have important life history traits that are not reflected in the standard Wright-Fisher model, including overlapping generations and two sexes. Additionally, in most domesticated species, fewer males are reproducing than females but males can reproduce over a longer time period, spanning several generations. While such life history traits have an impact on the response of the population to selection and are therefore accounted for in breeding programs (Hill 1974), changes in allele frequencies due to drift remain well described by scaling the models with an appropriate effective population size (N_e; Wright 1931; Engen *et al.* 2007). As a consequence, demographic inference in domesticated species using coalescent theory does usually not model overlapping generation or sex biased population sizes.

However, simple scaling does not extend to models incorporating pedigree information. To address this, we present here a Wright-Fisher-based diploid two-sex model with overlapping generations well describing pedigrees observed from domesticated breeding programs. Other specific traits that did not prevent us from modelling the population using pedigree information, such as skewed offspring distribution or reproductive success (i.e., selection at the phenotypic, individual, level) are implicitly accounted for through their impact on the effective population size of each sex. This is a particularly interesting aspect of our model as differences in offspring number variance between males and females are common in both domesticated and wild animal populations. We then show how this model results in a simple scaling of the standard coalescent in the absence of pedigree information and derive some analytical and numerical solutions to obtain estimates of the model parameters using the information contained in the SFS and pedigree data jointly. This also allows for the estimation of important life history characteristics such as the degree of overlapping generations and sex specific population sizes from such data.

An additional advantage of our framework is its ability to infer effective population sizes (N_e) and mutation rates (μ) jointly. Under the standard coalescent framework, both parameters are simply scaling the coalescent tree and hence only their product can be estimated, usually in the form $\theta = 4N_e\mu$. Wakeley and Takahashi (2003) showed that a joint estimation becomes feasible if the sample size n exceeds N_e since the rate of coalescence in the first few generations is a function of the ratio n/N_e and hence contains information about N_e regardless of μ. This was later used to infer gene specific mutation rates in humans from deep sequencing data Nelson *et al.* (2012); Schaibley *et al.* (2013). As we show here, pedigree information also contains information about N_e independent of μ, enabling the joint inference of both demography and mutation rate even in case where $n \ll N_e$.

1.2.4 Outline

Seven chapters constitute this thesis. Following this brief introduction, Chapter 2 introduces the theoretical concepts and methods used in this thesis and that might not be familiar to the reader. After which, the first part focusing on plant domestication and improvement describes the use of mirosatellite data to investigate cultivated rye. It presents in Chapter 3 the description of genetic diversity and demography of weedy and domesticated rye populations at different levels of improvement and in Chapter 4 the inference of neutral demographic models for a subset of these rye populations. These results have been published in Molecular Ecology (Parat *et al.* 2016). The second part centered on animal domesticates that have pedigree records, with cattle as an example, presents in Chapter 5 a new method to infer demography using jointly pedigree data and genetic information and in Chapter 6, applications of this model to simulated and real data. Finally a general conclusion summarizes the results and points of discussion of this work.

Chapter 2

State of knowledge

A classic approach to infer demographic and selective forces in domesticated and wild species is to model the effect of such forces on allele frequencies at polymorphic sites. The polymorphism data consists of simple sequence repeats (SSR) or single nucleotide polymorphisms (SNP) for several individuals of a given population. The underlying parameters of the evolutionary forces are inferred by matching the observed patterns to the ones produced by simulations or to expectations from a model. The inference is based on likelihood or Bayesian methods. In the following, I introduce the neutral population models, mutational models, summary statistics for genetic data, and inference methods that are used (as such or adapted) in this thesis.

2.1 The models

2.1.1 The Wright-Fisher model and its coalescent approximation

The Wright-Fisher model (Fisher 1922, 1930b; Wright 1931) is one of the most widely used models to describe the impact of drift and demographic forces on allele frequencies. This is mainly due to the simplicity of interpreting it in both biological and mathematical terms. In its original form, this model describes a population of diploid monoecious organisms of constant size N_{cens}. In the following, we call $N = 2N_{\text{cens}}$ the number of chromosomes at any generation. Generations are discrete and the number of offspring for each parent is binomially distributed. (See Box 1 for a more rigorous description.) When N is sufficiently large, it can be convenient to approximate the binomial distribution by a Poisson distribution of mean and variance 1. In other terms, the generations are not overlapping and the N descendant chromosomes in generation $g + 1$ are randomly equiprobably distributed among, and only among, the N potential parental chromosomes of generation g (Figure 2.1).

Biologically, this means that the expected reproductive success of every individual in the population is the same, i.e., there is no selection. There is no population substructure such as geographical distance or ecological barriers and the variance of offspring numbers is low, $1 - (2N_{\text{cens}})^{-1}$. The genetic process of allele transmission is then added to this demographic ancestry process: given two possible alleles segregating in the population, the allele carried by a parent is transmitted to all its descendants.

Because the population is finite and reproduction is a random process, some individual will not have any descendant and will therefore leave no genetic evidence that can be observed at present (Figure 2.1).

Box 1: The Wright-Fisher model

In a forward in time setting, the Wright-Fisher model describes a constant diploid population of constant size N_{cens}, with $N = 2N_{\text{cens}}$ chromosomes, that reproduce with discrete generations. Each chromosome in a generation $g + 1$ is the child of one chromosome from generation g. Conversely, the number of chromosome children of the j-th chromosome from generation g is a random variable ν_j assuming a symmetric multinomial distribution: $P\{\nu_j = n_j(j = 1, 2, \ldots, N)\} = N!/n_1!n_2! \ldots n_N!N^N$. To insure that the population remains constant, ν_j is constrained such that $\sum_{j=1}^{N} \nu_j = N$. Let us assume that each chromosome carries a bi-allelic locus with alleles A_1 or A_2 and that no mutation can occur. The probability that an allele in i copies in the present generation is found in j copies in the next generation is then

$$p_{ij} = \binom{N}{j} (i/N)^j [1 - (i/N)]^{N-j}, \quad i, j = 0, 1, 2, \ldots, N. \tag{B.1}$$

Denoting $X(g)$, the number of chromosomes carrying the allele A_1 at generation g, we observe that $X(.)$ is a Markovian random variable with transition matrix $\mathbf{P} = \{p_{ij}\}$.

Some properties of this model are well known (Ewens 2004), especially:

1. The largest non-unit eigenvalue of $\mathbf{P}$: $\lambda_{\mathbf{max}} = 1 - N^{-1}$
2. The probability that two chromosomes taken at random have the same chromosome parent: $\pi_2 = N^{-1}$
3. The variance of offspring allele frequency: $\mathbb{V}[x(g+1)|x(g)] = x(g)[1 - x(g)]N^{-1}$, where $x(g) = X(g)/N$ is the fraction of individuals of generation g carrying the allele A_1.

This stochastic process and the changes in allele frequencies that it entails are known as *genetic drift*. Furthermore, allele frequencies of individuals that have descendants at present time will only be observed if these descendants are sequenced. In most cases, the n sampled chromosomes represent only a small proportion of the reproductive population ($n << N$). Note that due to the progress in sequencing technology, this hypothesis can be violated. Indeed, some populations are sampled extensively and the sample size can be of the order of magnitude or larger than the reproductive population. Such oversampling leads to different ancestry shapes and requires adjustments of the derived models that will not be detailed here (Wakeley and Takahashi 2003).

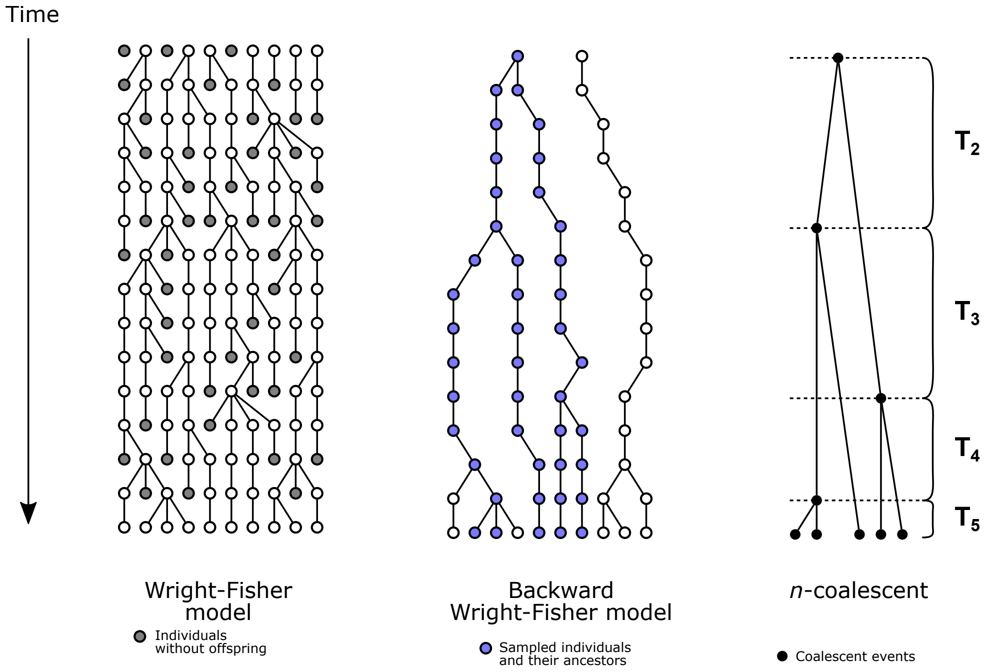

Time
T_2
T_3
T_4
T_5
Wright-Fisher model
● Individuals without offspring
Backward Wright-Fisher model
● Sampled individuals and their ancestors
n-coalescent
● Coalescent events

Figure 2.1: From the Wright-Fisher model to the n-coalescent in a haploid population.

Within the same Wright-Fisher model, but looking only at the ancestry of a sample rather than at the whole population, we can observe very different genealogical patterns (Figure 2.1). From this observation and assuming that the population is very large

($N \to \infty$), Kingman (1982a) derived a continuous, backward in time approximation: the n-coalescent.

Box 2: The n-coalescent process (Kingman 1982b)

For any natural number n, let E_n denote the finite set of equivalence relations on $\{1, 2, \ldots, n\}$. Let ξ be an equivalence relations on $\{1, 2, \ldots, n\}$ and η an equivalence relation that can be obtained from ξ by merging two of its equivalence classes. Such merger is referred to as a coalescent event. For $R \in E_n$, if the continuous-time Markov chain $\{R_t; t \geq 0\}$ is an n-coalescent, and ignoring terms of order $(\delta t)^2$, we can write:

$$\Pr(R_{t+\delta t} = \eta | R_t = \xi) = \delta t \tag{B.2}$$

thus

$$\begin{cases} \Pr(R_{t+\delta t} \neq \xi | R_t = \xi) = \sum_\eta P(R_{t+\delta t} = \eta | R_t = \xi) = d_i \delta t \\ \Pr(R_{t+\delta t} = \xi | R_t = \xi) = 1 - d_i \delta t, \end{cases} \tag{B.3}$$

where i is the number of equivalence classes of ξ, and $d_i = \binom{i}{2} = i(i-1)/2$ is the number of ways to choose two equivalence classes from ξ to be merged.

It follows that the sojourn times or waiting times T_i in any state ξ, are independent and follow an exponential distribution with parameter d_i, such that the probability density of the T_i is $d_i \exp(-d_i t)$ with $t > 0$.

Kingman's n-coalescent is the mathematical process describing the path, backward in time, of the ancestral lineages of n sampled chromosomes. This describes the sequential connections between their genetic ancestors (see Figure 2.1). The n-coalescent is a binary branching process. Starting at time 0 with n lineages from the n sampled chromosomes, the n-coalescent process merges pairs of lineages successively until only one linage is left, the most recent common ancestor (MRCA) of the sample. (For a more formal definition, see Box 2.) The times T_i between two successive coalescent events during which there are exactly i ancestral lineages are independent and exponentially distributed with parameter $\binom{i}{2}$ (Figure 2.1).

Kingman (1982a) also shows that this process describes the ancestral genetic process for a sample of finite size n in the limit as N approaches infinity in the Wright-Fisher model. The intuition behind this result is that, the number of lineages from the sample's ancestry is finite ($i \leq n$), therefore, as N grows to infinity, the probability that some of these lineages share a parent approaches 0. Moreover, the probability of more than two such lineages share the same parent, so-called multiple mergers, or that several groups of lineages share a parent at the same generation, simultaneous mergers (Figure 2.2), declines even faster to 0 and can therefore be neglected. (For a more formal derivation, see Box 3.)

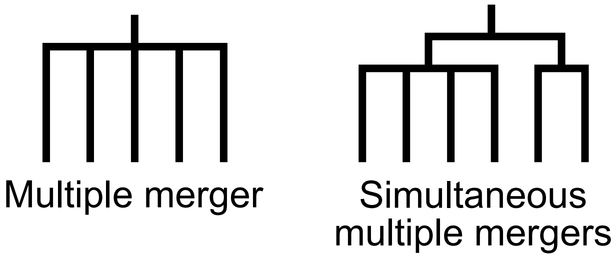


Figure 2.2: Schematic representation of multiple merger and simultaneous multiple mergers in coalescent genealogies.

Box 3: The Wright-Fisher model and the n-coalescent

As noted by Kingman (1982b), in a backward setting, the Wright-Fisher model can be formulated as: each chromosome of the generation $g + 1$ chooses its parent at random, independently and uniformly from the N chromosomes of g. Therefore, two chromosomes have a different parental chromosome at the previous generation with probability $1 - (1/N)$ and have the same parent with probability $1/N$. Three chromosomes have the same parent with probability $1/N^2$, four with probability $1/N^3$ and so on.

Using the same notations as in Box 2, we can describe this process backward in time as a discrete-time Markov chain $\{\mathcal{R}_g\}$ whose state space is E_n, the finite set of equivalence relations on $\{1, 2, \ldots, n\}$, with transition probability:

$$\begin{cases} \Pr(\{\mathcal{R}_{g+1}\} = \eta | \mathcal{R}_g = \xi) = \frac{1}{N} \\ \Pr(\{\mathcal{R}_{g+1}\} = \xi | \mathcal{R}_g = \xi) = 1 - \frac{d_i}{N} + \mathcal{O}(\frac{1}{N^2}) \\ \Pr(\{\mathcal{R}_{g+1}\} = \zeta | \mathcal{R}_g = \xi) = \mathcal{O}(\frac{1}{N^2}), \end{cases} \tag{B.4}$$

where ζ is an equivalence relations on $\{1, 2, \ldots, n\}$ that can be formed from η by merging more than two of its equivalence classes.

We recognize the similarities between this formulation of the Wright-Fisher model and the earlier formulation of the n-coalescent (Box 2) when the time is rescaled in N. Indeed, it can be proven (Kingman 1982b), that if $\mathcal{R}_{[N_t]}$ denotes the process on E_n in time rescaled in N, it converges in distribution as $N \to \infty$ to the continuous-time Markov chain R_t, that we described as the n-coalescent.

2.1.2 The concept of effective population size

Natural populations are never as simple as the models used to describe them. The effective population size (N_e) is defined as the size an ideal population should have to exhibit the same amount of genetic drift or inbreeding than our natural population. A small N_e implies strong genetic drift and fast loss of heterozygosity (i.e., fast loss of genetic diversity). N_e does not exist in nature, but it is a useful construct to understand the different evolutive forces; it serves as a conversion rate between the standard model and the reality or at least a more complex, more realistic model.

The three traditional types of N_e used in the Wright-Fisher framework are N_e^e (Ewens 2004), based on the eigenvalue of the transition matrix $\mathbf{P}$ described in Box 1, N_e^i, the inbreeding population size based on the probability that two chromosomes taken at random are descendant of the same parental chromosome, and N_e^v, based on the variance of the frequency of an allele given its frequency at the previous generation (Kimura and Crow 1963). The respective formulas can be directly derived from the three equations in Box 1.

If the natural population follows strictly a Wright-Fisher model, the census size N_{cens} equates the effective population size and all three definitions of N_e give the same result. In more realistic, more complex models including separate reproductive types (sex), population size changes or population structure, the census size will differ from N_e and N_e^e, N_e^i, and N_e^v might take different values or, in the case of N_e^v, not even exist (Ewens 1982).

From its definition, it is obvious that N_e^i relates to the probability of two lineages coalescing and therefore, in almost all cases, the coalescent effective population size N_e^c, if it exists, is equal to the inbreeding effective size as defined in Nordborg and Krone (2002). Therefore, in the following, N_e designates N_e^i when used in a discrete Wright-Fisher setting and N_e^c when used for a continuous coalescent process.

Similarly to the Wright-Fisher case, different definitions of N_e^c have been proposed to fit specific cases such as the presence of recombination and hitchhiking (Gillespie 2000) or to be very general and accept even other structures of genealogy than Kingman's

n-coalescent (Möhle 2001). Other definitions limit the existence of N_e^c to models that converge to the n-coalescent (Nordborg and Krone 2002; Sano *et al.* 2004; Wakeley and Sargsyan 2009), or, even further, to n-coalescent with a linear change in time scale (Sjödin *et al.* 2005), but yield the same value of N_e^c if it exists.

As stated in the previous section, the coalescent process arises from the Wright-Fisher model as a limit when N goes to infinity and time is rescaled by N. The coalescent process itself is independent of the population size, meaning that the waiting times between coalescent events relative to each other are not affected by N but the absolute waiting times in generations grow linearly with N. Time scaling is an important concept in coalescent theory as many other models converge to a n-coalescent when time is scaled "appropriately". Therefore, in the literature, the scaling factor is sometimes included in N_e for convenience but, due to the many different definitions of N_e and potentially non-linear scaling factors, this might obscure the meaning of N_e and bring further confusion into its biological interpretation (Nordborg 2001).

2.1.3 Model extensions and robustness of the coalescent

Under its original form, the Wright-Fisher model and its coalescent approximation have few but very strong assumptions that are violated in most, if not all, natural populations: diploid monoecious individuals, a constant population size and a binomial offspring distribution. Several extensions have been proposed to encompass a wider range of possible demographic scenarios and life history traits.

Dioecious populations

The first hypothesis that needs to be relaxed is the monoecious reproduction system. Monoecy describes a sexual reproduction system without mating types or where all individuals carry both mating types (hermaphrodites). The relaxation of this constraint is already discussed in the seminal paper from Wright (1931), where it is shown that patterns of allele frequency fluctuations are approximately equivalent in a group of monoecious individuals with random fertilization or in a population equally divided between females and males. It is also shown that, in a population comprising N_f reproducing females and N_m reproducing males, replacing the population size by

$$\frac{4N_mN_f}{N_m + N_f} \tag{2.1}$$

results in the same rate of loss of heterozygosity as in a monoecious population where selfing is prevented. If N_f and N_m are different, the effective population size is thus mainly under the influence of the smallest of the two population sizes. A comparable result has been derived for the coalescent of a two-sex population by Möhle (1998a). Although Möhle's model differs in that it considers a strictly monogamous population where couples are formed randomly but siblings always share both their mother and their father (full-siblings), the coalescent limit holds in both models with simply a different scaling factor. Genetic drift is doubled in the monogamous model compared to the *random union of gametes* model.

Both dioecy with unbalanced sex ratios and monogamy are common traits in domesticated populations. It is especially true for animal domesticates where males can have a much larger progeny than females and therefore, a smaller number of males is usually used for reproduction. A certain degree of monogamy can occur in species that produce litters (i.e., multiple births) as well as in plant breeding, where all seeds obtained from a plant by plant crossing are full siblings.

Changing population sizes

The second hypothesis that can be relaxed is the constant population size. It is straightforward to include population size changes in a Wright-Fisher model by changing N in function of time in the transition probability of the Markov chain (Box 1). Interestingly, as in the case of dioecious populations, the effective size is influenced largely by the occurrence of small sizes. In other terms, population size fluctuations tend to increase drift and inbreeding and therefore decrease diversity compared to the same average but constant population size. In the case of random or cyclic size changes and if the population stays relatively big, N_e can be approximated by the harmonic mean of the population sizes (Wright 1938b; Nei *et al.* 1975).

Looking backward in time, the intuition for this result is that the probability of two lineages not coalescing before a time T or, expressed differently, the probability that the waiting time for the coalescent event is equal or greater that T when observing two lineages can be written:

$$\Pr(t \geq T) = \prod_{t=0}^{T-1} 1 - \frac{1}{2N_t}, \tag{2.2}$$

where N_t is the population size at a time t. If N_t is large enough for every t in that interval, we can use the approximation $\log(1-y) \approx -y$ and write:

$$\Pr(t \geq T) \approx \exp\left(\frac{1}{2N_{HT}}\right), \tag{2.3}$$

where N_{HT} is the harmonic mean of N over T generations (Charlesworth and Charlesworth 2010). However, such an approach is only relevant for rapid fluctuations of the population size (faster than the coalescent process) without very severe bottlenecks (Charlesworth and Charlesworth 2010).

To study population size changes at a larger time scale, one can build explicit models of deterministic population size changes. In these models, the times T_i between coalescent events are not independent anymore. Two main types of such models have been studied in population genetics: the geometric (or exponential) growth model and the stepwise population size change model.

The exponential growth model can be defined as follows: a population growing exponentially at rate r, forward in time, up to a present size N_0 has size $N(t) = N_0e^{-rt}$, t generations in the past. It is often used due to two important properties: (1) it has some

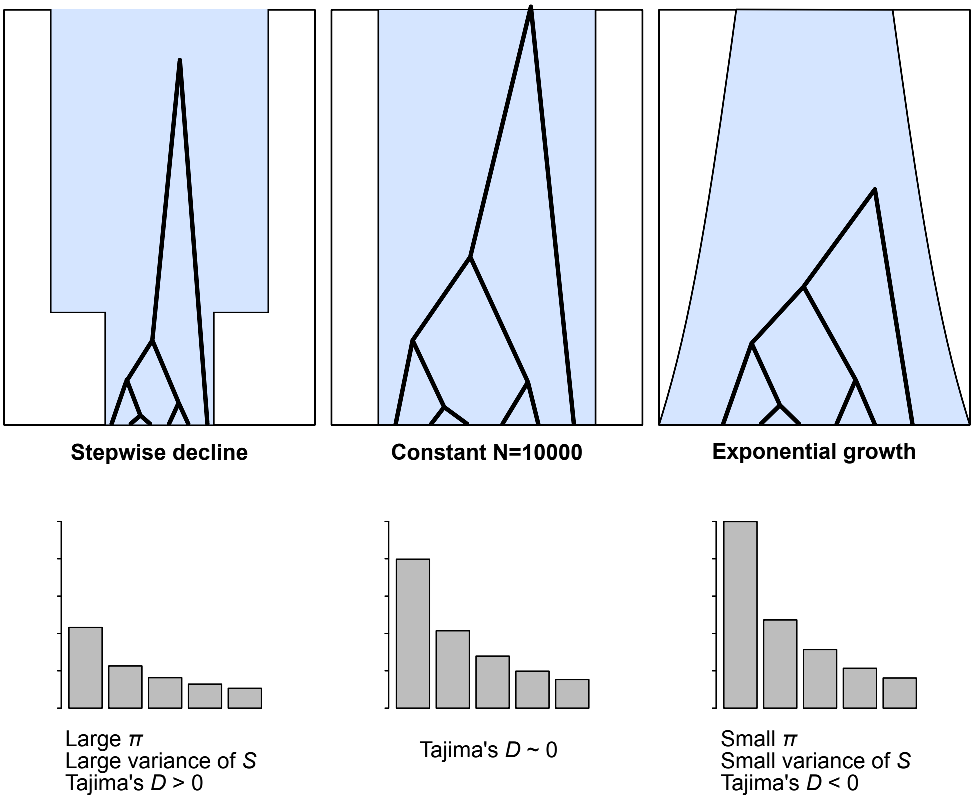


Figure 2.3: The impact of different demographic models on (A) the coalescent tree and (B) the site frequency spectrum.

biological relevance, as it is the continuous version of the geometric model that represent the first phase of a logistic growth, and describes well the growth of a population in an unlimited environment (i.e., infinite resources, no carrying capacity, density-independence); (2) the probability that the first coalescence among i lineages occurs in generation t, derived from $\Pr(t \geq T)$ (Slatkin and Hudson 1991), as well as the joint distribution of times T_i between coalescent events can be analytically derived (Griffiths and Tavare 1994). These probabilities provide a foundation to use this model in inference methods by facilitating the simulation of genealogies or allowing the calculation of maximum likelihood estimates of population demographic parameters, namely r and the starting time of the expansion.

When the population is exponentially growing, the probability that two lineages share a common ancestor at the next generation (i.e., coalesce) decreases, external branches tend to be longer and genealogies are more star-shaped (Figure 2.3). Conversely, exponential population decay increases the length of internal branches.

Coalescent simulations based on the method from Slatkin and Hudson (1991) can also be adapted to other deterministic models of population size changes such as the piecewise-constant (also called stepwise, Figure 2.3) and piecewise-exponential model (Donnelly and Tavare 1995). This method is described in Box 4. Generalization of the joint probabilities calculated analytically in the exponential case is more complex and leads to numerically instable integrals that can only be calculated for small sample sizes (Griffiths and Tavaré 1998; Polanski *et al.* 2003) or requires additional approximations (Chen and Chen 2013).

By approximating size changes with piecewise constant population sizes or by combining these result into a piecewise exponential model, one can model fairly complex demographic scenarios with good precision. However, increasing model complexity implies an increase in the number of parameters defining the demographic scenario, and therefore, one needs to find efficient compromises in model parametrization to avoid over-parametrization (Lapierre *et al.* 2017).

Population size changes play an important role in the genetic makeup of domesticated species. Domestication is traditionally thought to be associated to a rapid decrease in population size, the ”domestication bottleneck”, often modeled as a stepwise or exponential population size decrease, followed by some recovery. However, such short strong bottlenecks were not found in several perennial crops and recent studies of ancient DNA have also challenged this hypothesis in several annual crops such as maize, sorghum and barley (Allaby *et al.* 2019). The complex history of these species and the ongoing debate on the existence of a domestication bottleneck versus more progressive effects of domestication on population size reduction and diversity loss of crops, show the need for accurate and versatile models of population size variation.

Box 4: Simulating genealogical trees in varying populations

Let us assume that the offspring distribution variance is that of a Wright-Fisher model and that time is measured in $N = M(0)$ generations where $M(g)$ is the population size g generations in the past. If there exist an increasing continuous function $\Lambda(t)$ with density $\lambda(.)$, for which

$$\lim_{N\to\infty} \sum_{i=1}^{\lfloor Nt \rfloor} 1/M(i) = \Lambda(t), \tag{B.5}$$

where $0 < \Lambda(t) < \infty$ for $t > 0$, then,

$$\Pr(T_i < t | T_n + \ldots + T_{i+1} = \tau_i) = \exp\left[-\binom{i}{2} \int_{\tau_i}^{t_i+\tau_i} \lambda(s) ds \right]. \tag{B.6}$$

There, the density $\lambda(t)$ represents the ratio of the population size now to the population size t generations ago, that is e^{rt} for the exponential growth model and for the stepwise model, $\lambda(t)$ is 1 before the step and N_0/N_1 after the step. One can simulate coalescent times following the algorithm

```
Draw $n-1$ independent random variable $U_n, \ldots, U_2$ uniformly distributed between 0 and 1 ;
Set $\tau_n = 0$ and $i = n$ ;
while $j > 1$ do
    Simulate $T_i$ by solving the equation
      $U_i = \Pr(T_i < t | T_n + \ldots + T_{i+1} = \tau_i)$ ;
    Set $\tau_{i-1} = \tau_i + T_i$ ;
    $j = j - 1$ ;
end
```

Population structure

The standard Wright-Fisher model hypothesizes random mating, also called *panmixia*, implying that individuals are spread homogeneously in the population and that all pairs of individuals have the same probability to mate. Panmixia, is a central concept in population genetics, it is one of the strongest hypotheses of the Hardy Weinberg equilibrium (HWE, Box 5, Hardy 1908; Weinberg 1908). In nature, however, physical distances or geographic barriers (e.g., rivers, deserts), as well as behavioral (e.g., social herd or pack life) and genetic constraints such as mating types can lead to some pairs of individuals having a lower (or higher) probability than others to mate together. This violation of the random reproduction between individuals is often called *population structure*. This leads to a deficit of heterozygotes compared to the expected value following the HWE in the population taken as a whole, although each subpopulation is at equilibrium (Hartl and Clark 2007). This reduction of heterozygosity can be quantified using Wright's coefficient of inbreeding calculated from the pedigree (Wright 1922) or by Wright's F-statistics (Wright 1949). The latter is defined as the fraction of the decrease of heterozygosity within a level of structure to the heterozygosity among levels of structure.

The index of fixation (F_{ST}), is one of these statistics and measures all effects of population structure combined:

$$F_{\mathrm{ST}} = \frac{H_T - H_S}{H_T}, \tag{2.4}$$

where H_S is the average expected heterozygosity under HWE within random mating subpopulations (S) and H_T is the expected heterozygosity under HWE in the total population (T), taken as a whole. F_{ST} can theoretically vary between 0 and 1 but rarely reaches 1. Following Wright's guidelines (Wright 1978), values lower than 0.05 indicate no to little genetic differentiation whereas values above 0.25 already indicate substantial genetic differentiation. By analogy and considering a diploid individual as a level of structure, two more indices are commonly defined, F_{IS}, the inbreeding coefficient of an individual (I) relative to the subpopulation, thus comparing expected and observed heterozygosity, non-zero values revealing non-panmictic subpopulations, and F_{IT}, the inbreeding coefficient of an individual relative to the total population, that results from the combination of all processes within and among subpopulations. These indices are ultimately related through the formula: $1 - F_{\mathrm{IT}} = (1 - F_{\mathrm{IS}})(1 - F_{\mathrm{ST}})$.

Box 5: Hardy-Weinberg principle

In an ideal population assuming:

- Diploid organism with sexual reproduction
- Identical allele frequencies in males and females
- Random mating (panmixia)
- Non-overlapping generations
- Infinite population size (i.e., no genetic drift)
- No migration
- No mutation
- No natural selection,

the genotype frequencies at a bi-allelic site with alleles A_1 and A_2 follow the relation: $f(A_1A_1) = p^2$; $f(A_1A_2) = 2p(1-p)$; $f(A_2A_2) = (1-p)^2$, where p is the allele frequency of the allele A_1 in the population.

The mechanism that results in this relation can be represented by a simple table:

		Male gametes	
		A_1	A_2
Female gametes	A_1	p^2	$p(1-p)$
	A_2	$p(1-p)$	$(1-p)^2$

Other possible genetic effects of population structure entail the apparent increase of genetic diversity when sampling across subpopulations compared to a sample in a panmictic population of the same size, and the potential decrease of diversity when sampling within a single subpopulation (see Section 2.2 for more details).

Several models have been proposed to describe population structure and migration between subpopulations explicitly.

In the most extreme case, subpopulations are separated at a given time t and stop exchanging genetic material. The subpopulations become completely independent from each other and each behaves as a population. In a coalescent framework, this means that the lineages can only coalesce within their own subpopulation until they reach time t after which the remaining lineages in all subpopulations can freely coalesce independently of their subpopulation of origin.

In less extreme cases, the separation between the subpopulations is not total, subpopulations can still exchange genetic material (individuals, eggs, seeds, pollen,. . .). This is called *migration*. The simplest model of population structure with migration is the *island model* (Wright 1931). In this model, all subpopulations, or *demes*, are identical in size and have the same probability of exchanging migrants among themselves. Because all d demes are exchangeable, only two coalescent probability have to be considered: the probability for two lineages to coalesce if they are in the same deme (T_{ii}) and the probability if they are in different demes (T_{ij}). We recall the known results:

$$T_{ii} \approx 2dN_e \tag{2.5}$$

$$T_{ij} \approx 2dN_e + \frac{d-1}{2m}, \tag{2.6}$$

where m is the migration rate and N_e the effective population size.

This model is particularly interesting because of its simplicity but does not account for one of the most common factors of population structure, geographic distance. Indeed, in most natural populations, individuals that are geographically close have a higher chance to reproduce with one another than with distant individuals. To account for this, Wright (1943) described spatial structure in continuous populations using the concept of *isolation by distance*. The individuals are spread along a line (or across a plane) with a given density and can reproduce with distant individuals following a given migration distribution centered on their place of birth.

A discrete deme model was develop to describe the same phenomenon of isolation by distance, the *stepping stone model* (Kimura 1953; Kimura and Weiss 1964), in which populations can only exchange migrants with their nearest neighbors, thus accounting for both the geographical distance of populations and the discrete characteristics of the environment (e.g., mountains, rivers) or life history traits (gregarious behavior) of many organism. In that case, the probability of coalescence for two lineages depends on the distance between their demes.

None of these models are very realistic but they lead to more general results such as, for example, the relation between F_{ST} and $N_e m$ (e.g., Wright 1943), that can be used to estimate population structure and migration using genetic data. Furthermore, in a coalescent framework, the island model can be extended and genealogies can be simulated for variable deme sizes and different, possibly changing, migration probabilities between demes. These complex models can account for all sources of population structure including geographical distance. The simulated coalescent trees can then be used for inference. Inference results can however be biased by a misspecification of the number of existing demes, now and in the past, or the partial sampling of demes, therefore, estimated migration rates should be treated with caution (Slatkin 2005).

Control over the domesticate's reproduction is one of the pillars of the domesticator-domesticate relationship and is contrary to the principal of panmixia. By isolating individuals from their wild ancestors, adapting populations to local conditions and specific purposes or creating distinct races and varieties, the domesticator creates new reproduction barriers, and therefore, population structure. However, the domesticator might also force migration among geographically and genetically distant populations, creating complex structure patterns that would not be observed in the wild.

Overlapping generations

The Wright-Fisher model considers generations as discrete or non-overlapping. This occurs in some natural populations, for example in insect populations that reproduce simultaneously and quickly thereafter, die, such that no adult from the previous generation can mate with an offspring from a latter generation. In a less drastic manner this hypothesis is usually considered to be true for humans in the sense that, usually, the reproduction period of parents and of their children do not overlap in time. However, in many populations this hypothesis is violated. Overlapping generations can occur through different processes, for example, a long and early reproductive period as for many vertebrates and trees, or the existence of a conservation form (seeds, spores, eggs or pupae that are able to survive over several generations) as for many plants, fungi and insects. Conservation of genetic material over several generations can also be artificial, in seed libraries or sperm banks. However, one should note that conservation forms, natural or artificial, do not necessarily lead to overlapping generations. Generations only overlap when individuals reproduce with individuals of distinct generations. For example, if a parent mates with its own progeny, a common practice in breeding.

There are two main ways of dealing with populations that present overlapping generations: (1) use a so-called *continuous model* or a discrete birth-death process that does not need a strict concept of generation unlike the Wright-Fisher model, or (2) modify the Wright Fisher model to include explicitly overlapping generations.

The most common birth-death process allowing overlapping generations is defined as the Moran model (Moran 1958). The Moran model is a birth and death model for haploid populations: at each time step, $t = 1, 2, 3, \ldots$, an individual is chosen to reproduce and produces one offspring, then, at the same time step, an individual is chosen at random to die. The individual that dies can be the individual that has just reproduced but not its new offspring. This implies that the population size remains constant, that the death event and the birth event are independent, and that the lifetime of individuals in the population is geometrically distributed. From this definition, it is also clear that no more than two individuals can share the same parent at the same time step and two or more coalescent events cannot happen simultaneously. This explains why exact results can be derived for the Moran model using the n-coalescent. The direct relation between the Moran model and the n-coalescent is convenient but the Moran model presents strong constraints on the age distribution that do not necessarily fit well natural populations.

More flexible models of overlapping generations, based on the Wright-Fisher model, have been developed to describe natural populations with particular history traits like seed banking (Kaj *et al.* 2001) or perenniality (Abu Awad *et al.* 2016). In most cases, these models can be approximated by a time-rescaled coalescent assuming that N is large, but this depends on the age distribution (Blath *et al.* 2013). In simple terms, if the time spent in a dormant state, (i.e., when the lineage cannot coalesce), is much shorter than the waiting time for a coalescent event, it does not influence the structure of the genealogy but only its rate.

Separation of time scales and robustness of the coalescent

In the previous sections describing violations of the coalescent it is repeatedly noted that as long as the disturbance of the model or the movement between model compartments (two sex, population size change, migration, overlapping generations) happens much faster than coalescent events, the coalescent process is unchanged in the limit and only the time scale is affected. This property is sometimes referred to as the *separation of time scales* and has been formally investigated in the general setting of Markov processes by Möhle (1998b). The overall *robustness* of the n-coalescent (Möhle 1998c, 1999; Möhle and Sagitov 2001), that is the fact that many population models using very different assumptions converge to the n-coalescent when N_e goes to infinity, is one of the reason for the popularity of this model. It implies that, in most cases, as long as N_e is reasonably large and the genetic sample smaller, the central property of simple binary merger of the n-coalescent remained true and one can use the same process with appropriate scaling to model the population of interest.

Offspring distributions

The Wright-Fisher multinomial offspring distribution implies a very low variance of the number of offspring per parent. In many natural populations however, some parents might, by chance, have much more offspring than others. This phenomenon is distinct from selection in the sense that a high or low number of offspring is not related to a particular trait or allele and will not be passed down to the next generation. As a simple example of high offspring variance one can consider two frogs laying eggs in similar ponds, one of the pond is dried out to build a road causing the death of all offspring of one of the frogs whereas most offspring of the other frog might survive and reproduce.

Following Crow and Kimura (1970), let us consider the number of offspring ν_i of chromosome i in $1, 2, \ldots, N$ as a random variable from a distribution with mean ξ and variance σ^2.

The probability π_2 that two randomly picked chromosomes come from the same parent is then

$$\pi_2 = \frac{\sum \binom{\nu_i}{2}}{\binom{N\xi}{2}} = \frac{\sigma^2/\xi + \xi - 1}{N-1}. \tag{2.7}$$

By definition, inbreeding N_e is then the inverse of this probability. Assuming the population is constant and that all individuals are exchangeable and thus have on average same number of offspring, $\xi = 1$ and $N_e = (N-1)/\sigma^2$.

High offspring variance increases therefore genetic drift or, in a coalescent framework, shortens the waiting time between coalescent events by a factor $1/\sigma^2$.

However, if the offspring variance is very large, the probability that more than two lineages join in the same coalescent event might not be negligible anymore and the ancestral process might not converge to Kingman's n-coalescent. Two different problems are highlighted here and should be distinguished. First, the approximation of the genealogy of a sample of n chromosome by the n-coalescent might be of poor quality. Indeed, one of the underlying assumption of the n-coalescent approximation is the infinite population size. In practice, for a small sample size, a N_e of several hundreds to few thousand individuals is sufficient to use this approximation but, if the variance of offspring numbers is large (i.e., not negligible compared to N_e), N_e will be strongly reduced, leading to a poor performance of the approximation. Namely, the model produces fewer external branches than occurs in reality.

Second, Kingman (1982b) showed that, in a population of N diploid individuals, if the variance ${\sigma^2}_N$ of the (random) number of offspring genes from one parental gene ν_N converges to a positive limit as N goes to infinity and if the supremum of all moments of ν_N are bounded, then, as N goes to infinity, the ancestral properties of a sample of size n in an *exchangeable Cannings model* converge to those of the n-coalescent, where Cannings models (Cannings 1974, 1975) are a wide family of models including the Wright-Fisher and the Moran model (Box 6). Other limit processes such as the Λ-coalescent (Pitman 1999; Sagitov 1999) and the Ξ-coalescent (Schweinsberg 2000), allowing respectively *multiple* and *simultanious mutliple mergers* (Figure 2.2), arise for other exchangeable Canning models as described in Möhle and Sagitov (2001).

Biologically, such skewed offspring distribution is known to occur in marine organisms but could also apply to plants where the production of seeds exceeds largely the carrying capacity of a population, or insects, viruses and microbial pathogens that experience rapid population boom and bust (Tellier and Lemaire 2014). It might also occur to a lesser extent in domesticated animal populations where a handful of animals, usually males, are mated to a large part of the population over one or two generations to bring or increase the frequency of traits of interest into that population.

Box 6: The Cannings family (Cannings 1974, 1975)

For a population with constant size N, no mutation and two alleles (A_1 and A_2), the Cannings exchangeable model is defined as a Markov chain based on a set of exchangeable 2-dimensional random variables $\{D_i, R_i\}$, where D_i is 1 if the i-th individual survives and 0 otherwise, and R_i is the number of descendants of the i-th individual, such that

$$X_{t+1} = \sum_{1}^{X_t} D_i + \sum_{1}^{X_t} R_i,$$

where X_t is the number of individuals carrying allele A_1. In order to have a constant population size, the process is constraint by

$$\sum_{1}^{N} D_i + \sum_{1}^{N} R_i = N.$$

The Cannings models allow a more general definition of discrete time population models depending on the distribution of $\{D_i, R_i\}$ or of their sum Q_i.

For example, the Wright-Fisher model can be defined as a Cannings model with D_i is 0 for all i and R has a multinomial distribution. In the Moran model, no distinction is made between surviving at the next generation and having an offspring, and the distribution of Q is such that $Q = \mathbf{1}$ with probability $1/N$, Q is a permutation of $\{0, 2, 1, 1, \ldots, 1\}$ with probability $(N-1)/N$ and cannot take any other value.

2.1.4 Mutations

In the previous sections, we have focused on the relation between individuals (or chromosomes) which defines the shape of the genealogies and the rate of coalescent. However, when working with genetic data, we only can observe the combined result of genealogy and mutation, for example under the form of allele frequencies in the sequence data of sampled individuals. Several models have been developed to describe mutational processes for different types of genetic data. The first type of data collected was phenotypes such as the color of a flower, the texture of peas or the patches of cows. These phenotypes were in general considered as bi-allelic markers, following Mendelian inheritance, with a given mutation rate u conferring the derived phenotype and v reverting to the wild phenotype (Wright 1931). In most cases, both mutation rates are sufficiently small and the time scale at which one observes phenotypes short enough that models without mutation describe well how alleles, present at a given frequency, are segregating in the population in following generations (see the Hardy-Weinberg equilibrium in Box 5 and the original Wright-Fisher models in Box 1). The first molecular markers were proteins, isozymes and allozymes (Hunter and Markert 1957), and later appeared DNA markers detected with the help of restriction enzymes (e.g., restriction fragment length polymorphism, RFLP; Grodzicker *et al.* 1974) or PCR reactions, hybridization techniques and sequencing, allowing the detecting of simple sequence repeat (SSR, microsatellites and minisatellites), single nucleotide polymorphisms (SNPs) or copy number variation (CNV). Using DNA markers, we can now observe the interaction of mutation and genealogy.

2.2 Data and summary statistics

2.2.1 Single nucleotide polymorphisms (SNPs) and their summary statistics

Mutational models

Nowadays, single nucleotide polymorphisms (SNPs) are widely available and have become, in the last years, one of the most common form of genetic data. The two main models that have been developed to describe the behavior of such mutations are the infinitely many allele model (IAM Kimura and Crow 1964) and the infinitely many sites model (Kimura 1969). The infinite allele model considers that mutations appear randomly and that a given site can be affected several times but that the allele arising from these repeated mutations will always be distinct. Mutations are irreversible, that is, no *back mutation* is allowed. The infinite sites model is even more conservative, and states that mutations are random but because of the low rate of mutation μ per base pair, mutations do not modify the same site several times. These approximations of infinitely many sites or alleles might not hold for some fast mutating organisms like RNA lytic viruses with $\mu \sim [10^{-4}\text{–}10^{-3}]$ (Drake *et al.* 1998), or for certain regions of the genome (e.g., hot spots or hypermutable sites, Hodgkinson *et al.* 2009) but are overall reasonable in many multicellular plants and animals, especially for small sample sizes.

Typically, μ is of the order of 10^{-6}–10^{-10} per base per replication in unicellular organisms (Drake *et al.* 1998), of the order of 10^{-8} per base per generation in the nuclear genome of multicellular animals such as *Drosophila melanogaster*, humans or mice (Baer *et al.* 2007). For plants, estimations are scarce. Recent work from Ossowski *et al.* (2010) give a mutation rate of 7×10^{-9} base substitutions per site per generation for *Arabidopsis thaliana*'s nuclear genome, but the mutation rate per generation might be very variable between annual plants and long lived trees (Klekowski and Godfrey 1989). Animal mitochondrial DNA tend to have a higher mutation rate than nuclear DNA (nDNA) whereas plant mitochondrial and chloroplast DNA would tend to have a 1/6, respectively 1/2, lower mutation rate than nDNA (Wolfe *et al.* 1987). See Lynch *et al.* (2016) for a more complete review of mutation rates.

Interestingly, without recombination, both infinitely many sites and infinitely many alleles models are similar. Indeed, one can either consider each nucleotide as a site or consider a non-recombining fragment as a site and the different haplotypes as the alleles. If one sees haplotypes as different alleles, then recombination could, like mutation but with a different process, create new alleles whereas this in not permitted in the IAM (Ewens 2004).

Because a mutation is a rare event that occurs randomly during cell multiplication, the accumulation of mutation on a chromosome can be assimilated to a Poisson process with parameter μ, the mutation rate. When applying this same process along a genealogy, it means that the number of mutations that occurs on a lineage between two coalescent events is, in expectation, proportional to both the mutation rate and the length of the branch. Consequently, factors affecting the shape of the genealogy described in previous sections similarly affect the mutations observed in the genome. Conversely, in an inference perspective, it implies that the number and frequency of observed SNPs contain information about the demography, reproductive system and life history traits of the individuals. This forms the basis for the population genetics field. The proportionality of mutation and branch length has, however, one adverse effect: mutation rate and effective population size cannot be distinguished and are therefore combined in a scaling factor, the *population mutation rate* (θ); for haploids, $\theta = 2N_e\mu$ and $\theta = 4N_e\mu$ for diploids. Indeed, as shown in Section 2.1.1, time and therefore the length of the genealogy is rescaled in N_e for haploids or $2N_e$ for diploids. This rescaling factor is multiplied by 2 for historical reasons related to the concept of heterozygosity or pairwise differences (eq. 2.11).

The site frequency spectrum (SFS)

As shown in previous sections, demography influences the shape of the genealogy, in particular the relative length of internal and external branches (Figure 2.3). When looking only at a sample of genetic data from a single time point, there is no way of knowing on which branch of genealogy the observed polymorphism appeared. Nevertheless, the number of times a mutation is found in a genetic sample gives information about the type of branch on which it appeared, internal or external. Therefore, a useful and convenient way of summarizing this information is to class and sum mutations following the number of individuals in the sample carrying them. Historically, this was first proposed for the infinite allele model under the form of the Ewens' sampling formula (Ewens 1972) based on the *diffusion approximation*. This work, predates the formulation of the n-coalescent, and reaches the same results using diffusion equations instead of Markov chains. In terms of n-coalescent genealogies, this means summing the length of all branches ordered by growing number of leaves (i.e., sampled chromosomes) they lead to (Figure 2.3).

The resulting histogram is called the site frequency spectrum (SFS) and has dimension $n-1$, where n is the size of the sample. The expected number of mutations appearing at a given frequency in the whole population was derived under different mutational models (Fisher 1930a; Wright 1938a) and different offspring distributions (Haldane 1939) for discrete forward models. Later with the development of the diffusion approximation and of the n-coalescent, the case of sample SFS was treated by Kimura (1964) and Tajima (1983). All different method result, to some constant coefficient, in the expectation

$$\mathbb{E}[\xi_i] = \theta/i \quad 1 \leq i \leq n-1, \tag{2.8}$$

where ξ_i is the number of segregating sites for which the derived allele is found exactly i times in the sample.

In practice it is not always possible to infer the ancestral state of a mutated nucleic acid and, often, a mutation present once in a sample could as well be the reverse mutation present $n-1$ times in this sample. In such cases where no sufficiently good out-groups are available to distinguish the derived from the ancestral allele, the SFS is said to be *un-rooted* or *folded*, as the classes of the SFS are summed as if it was folded in its center such that $\eta_i = \frac{\xi_i+\xi_{n-i}}{1+\delta_{i,n-i}}$, where $\delta_{i,j}$, Kronnecker's δ, is 1 when $i=j$ and 0 otherwise.

Several one-dimension summary statistics have been developed to summarize the information contained in sampled genetic data. Many of them can be derived from the folded SFS, for example, the number of segregating sites (S), the nucleotide diversity (π) and Tajima's D, a statistic to detect departure from the neutral model.

The number of segregating sites

Long before the n-coalescent was formalized, the relation between θ and the expected number of observed segregating sites $\mathbb{E}(S_n)$ in a sample of size n, in the case of the Wright-Fisher model, for independent sites (Ewens 1974) or for a chromosome without recombination (Watterson 1975), infinitely many sites, and Poisson distributed mutations, was already known and, for small n, had the form:

$$\mathbb{E}(S_n) \approx \theta \sum_{i=1}^{n-1} \frac{1}{i}. \tag{2.9}$$

Rearranging this equation yields one of the unbiased estimators of θ, noted $\hat{\theta}_S$ in the following.

In his paper, Watterson (1975) gave the probability generating function for S_n without recombination, in the case of the Wright-Fisher and the Moran models, yielding the variance

$$\mathbb{V}(S_n) \approx E(S_n) + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}. \tag{2.10}$$

More interestingly, he noted, precursor to coalescent theory, that S_n is the sum of the number of new mutations occurring during generations when there were exactly $i+$

1 distinct ancestors in the genealogy Y_i, and that for small samples, the Y_i are $i-1$ independent, approximately geometrically distributed random variables in the Wright-Fisher case and exactly so in the Moran model.

Pairwise differences

Another commonly used measure of genetic variation is the average number of pairwise differences among sequences of a sample, π (Nei and Li 1979; Tajima 1983), that is

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} \tag{2.11}$$

in which d_{ij} is the number of differences between the i-th and j-th sequences. In the standard n-coalescent, $E[\pi] = \theta$ and $Var[\pi] = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2$, giving θ_π, another unbiased estimate of θ but with a higher variance than θ_S for large samples (Tajima 1983).

The number of pairwise differences can also be derived from the unfolded SFS as

$$\pi = \frac{1}{\binom{n}{2}} \sum_{n-1}^{i=1} i(n-i)\xi_i, \tag{2.12}$$

or the folded SFS since π puts symmetrical weights on the unfolded SFS classes.

Theta estimators and tests for neutrality

A common way of testing neutrality is to compare two unbiased estimator of θ. The most often used test based on this principle was proposed by Tajima (1989) and depends on the difference between $\hat{\theta}_\pi$ and $\hat{\theta}_S$.

$$D_T = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{\hat{\mathbb{V}}[\hat{\theta}_\pi - \hat{\theta}_S]}}, \tag{2.13}$$

where $\hat{\mathbb{V}}$ is the unbiased estimate of the variance of $\hat{\theta}_\pi - \hat{\theta}_S$ that can be calculated from $\mathbb{V}[\hat{\theta}_\pi - \hat{\theta}_S] = \mathbb{V}[\pi] - 2\mathbb{C}\mathrm{ov}[\pi, S]/a_1 + \mathbb{V}[S]/a_1^2$, yielding

$$\hat{\mathbb{V}} = \frac{c_1 S}{a_1} + \frac{c_2 S(S-1)}{{a_1}^2 + a_2} \tag{2.14}$$

with $c_1 = \frac{n+1}{3(n-1)} - \frac{1}{a_1}$, $c_2 = \frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{na_1} + \frac{a_2}{a_1^2}$, $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$, and $a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$.

Other estimators of deviation from neutrality based on theta estimators, include Fu and Li's D (Fu and Li 1993) and Fay and Wu's H (Fay and Wu 2000) respectively taking the form

$$D_{FL} = \frac{\hat{\theta}_S - \hat{\theta}_{\eta 1}}{\sqrt{\hat{\mathbb{V}}[\hat{\theta}_S - \hat{\theta}_{\eta 1}]}} \quad \text{and} \quad H_{FW} = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sqrt{\hat{\mathbb{V}}[\hat{\theta}_\pi - \hat{\theta}_H]}},$$

where $\hat{\theta}_{\eta 1}$ and $\hat{\theta}_H$ are unbiased estimators of θ. These are respectively, the singleton estimator, depending on the singleton class of the site frequency spectrum η_1, and Fay and Wu's estimator that is similar to θ_π but that particularly weights high frequencies in a non-symmetrical way, requiring the unfolded SFS.

2.2.2 Simple sequence repeats and their summary statistics

Simple sequence repeats (SSRs) have a different mutational dynamic than SNPs. In regions of low complexity, the amount of errors during replication is increased and repeated motifs tend to be gained or lost due to slippage (misalignment of DNA strands during replication), recombination errors or point mutations. Due to the high rate of mutation causing both gain and loss of repeat units there is a high potential for homoplasy (similar motif acquired by independent mutations) and the infinite allele model (IAM) is not adapted anymore. Conversely the concept of repeat units being added or removed implies that allelic states are not interchangeable and that there is a positive correlation between the number of mutations needed, and thus time, and the difference in repeat number. This observation led to the application of models developed for the study of charge state of proteins development such as the stepwise mutation model (SMM) (Ohta and Kimura 1973; Wehrhahn 1975). However, mutation essays showed that the number of repeat units being gained or lost is not always constant (Weber and Wong 1993), and larger, non-unit, changes in repeat numbers might occur that are not compatible with the SMM. In such cases, the two-phase mutation model or generalized stepwise mutation model (GSM), that allows for such large changes in repeat numbers is more suited (Rienzo *et al.* 1994). See Box 7 for the model definition.

Box 7: The two-phase mutation model from Rienzo *et al.* (1994)

The two-phase mutation model or generalized stepwise mutation model allows multistep mutations. When a mutation occurs, it can belong to the one-step phase with probability p or to the multistep phase with probability $1 - p$. In the one-step phase, the new allele is either one repeat unit larger or smaller than its ancestor. In the original model both probabilities are equal but this can easily be relaxed. In the multistep phase, the change in the number of repeat units is drawn from a specified distribution, g_j, that allows for various sizes of changes in repeat number. In the original model, g_j was assumed to be a symmetric geometric distribution with a specified variance, σ_j^2, such that, $g_j = C\alpha^i$ for $j \leq 1$, $g_j = g_{-j}$. The normalization constant C was chosen to satisfy $\sum_{i=1}^{\infty} g_j = 1/2$ and α is then determined by the σ_j^2. Under this original form, the two-phase mutation model requires, thus, three parameters: the mutation rate (μ), the fraction of one-step mutations (p), and, if $p < 1$, the variance of the distribution of larger step mutations (σ_j^2).

The number of sites sequenced for SSR is fixed and many statistics based on the segregating sites or the SFS are meaningless for these markers. But one can take advantage of the quantity of possible alleles, as well as the correlation between time and the length of the repeat motifs to build informative summary statistics.

The literature on the measures of differentiations for SSR markers is very rich and, as it is not the central object of this thesis, only a few examples will be detailed.

Several modifications of the traditional measure of population differentiation, the fixation index F_{ST} (Malécot 1948; Wright 1949), were proposed to build statistics adapted to the specificities of SSR markers. First Nei (1973) proposed the G_{ST} statistic, an equivalent of Wright's F_{ST} for the multiple allele case. However, the maximum value that G_{ST} can take depends on the number of populations and the total number of alleles, complicating greatly the interpretation or comparison of this statistic among studies. Therefore, Hedrick (2005) proposed a normalized version of the statistic called G'_{ST} that can take all values between 0 and 1 independently of the total number of alleles present. Jost (2008) introduced a differentiation statistic (D) in the case of multiple alleles that has similar properties than G'_{ST}.

In parallel to these measures of differentiations, several statistics were developed to take advantage of the stepwise mutation process of SSR markers. Two related measures of differentiation that account for the difference between microsatellite allelic sizes, R_{ST} (Slatkin 1995) and Φ_{ST} (Michalakis and Excoffier 1996) have been defined as analogous of G_{ST} and F_{ST} respectively. The distance $\delta\mu^2$ proposed by Goldstein *et al.* (1995) also accounts for allelic size but has been developed under a strict stepwise mutation model.

Allele size-based measures of differentiation, that assume a stepwise mutation process, might not reflect the behavior of all SSR markers and, even if they do, might be less efficient than allele identity-based statistics when the contribution to population differentiation of mutation is negligible compared to that of drift. Therefore, Hardy *et al.* (2003) proposed a test to determine whether stepwise-like mutations contributed to genetic differentiation. Furthermore, different differentiation statistics capture different aspects of population structure and can present different limitations. Using a combination of these statistics provides therefore further insights into population structure and demography.

The information contained in allele sizes cannot only be used to gain knowledge on population differentiation but also to measure population size changes, and more specifically, recent bottlenecks. For this purpose, Garza and Williamson (2001) developed a statistic, the M-ratio, based on the ratio of the number of alleles, K, by the range of allele size R. The principle underlying this statistic is that, during a bottleneck, K reduces strongly due to the loss of alleles by drift whereas only the loss of the biggest or smallest alleles reduces the range of allele sizes. Therefore, if the biggest and smallest alleles are not rarer than the average length ones, that is if the allele length frequency distribution is not bell-shaped, R will be less affected by the bottleneck than K. This statistic is informative for recent bottlenecks but its interpretation can be difficult. In particular, the molecular signature of a bottleneck can be obscured by substructure and migration (Busch *et al.* 2007) and the value of the M-ratio is sensitive to the sample size and to the violation of the stepwise mutations model. Indeed, very frequent multistep mutations can increase the number of gaps in the allele size distribution. Even when the markers follow a generalized stepwise mutation model with rare multistep mutations ($p < 0.10$), occasional mutations may greatly affect allele size, either by rare but large steps in the normal mutation process or through other processes, such as modification of the flanking sequences.

2.2.3 Pedigree data

In this thesis, we define the pedigree as the list, or tree, representing ties between individual parents and their offspring as recorded for many domesticated animals (e.g., cattle or horses). Concretely, the standard format for a pedigree is a list of individuals and for each individual the identity of its mother and its father, if known. Models based on pedigrees are very similar to those used to study genealogies but both structures are distinct and many different genealogies can be built within a given pedigree. For many populations, especially among domesticated species, pedigree data has been recorded long before any genetic data was available. Applying the same Wright-Fisher model described for genealogies to this type of data enabled the calculation of inbreeding coefficients (Wright 1922) from which population size or structure could be inferred.

2.3 Bayesian statistics and inference

Bayesian statistics use data under the form of the likelihood to update a prior belief into a posterior probability distribution. Using Bayes' rule on conditional probabilities for a continuous variable, we write

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}, \tag{2.15}$$

where $f(\theta|x)$ is the conditional density of θ given x, $f(\theta)$ is the so-called prior density of θ and $f(x)$ is the marginal probability of x, $f(x,\theta)$ is the joint density of x and θ.

In most cases, the integral in the denominator cannot be calculated; so this equation has to be solved using properties of conjugacy or numerical calculation such as Monte Carlo Markov Chains (MCMC). In more complex cases, even the likelihood cannot be calculated and one has to use likelihood free methods such as approximate Bayesian computations (ABC).

2.3.1 Monte Carlo Markov Chain (MCMC)

Monte Carlo Markov Chains is a family of algorithms used to sample from a posterior density by constructing a Markov chain that accepts this same posterior density as its stationary distribution. The most common MCMC algorithms are the so-called random walk algorithms in which the chains are moving randomly in the parameter space. In this family, the Metropolis (Metropolis *et al.* 1953) and the more general Metropolis-Hasting (Hastings 1970) algorithms are often used in cases where the conditional distribution of the target distribution cannot be exactly sampled.

Box 8: The Metropolis-Hasting algorithm

```
Given a $n$-dimensional vector $\mathbf{x}$ of parameters
Initialization: Choose arbitrary initial values for each parameter of the
  vector, $\mathbf{x_0}$
while steps do
    foreach $x_t$ in $\mathbf{x_t}$ do
        Draw a random value $x'$ from a proposal distribution $\Pr(x_t \to x')$;
        Calculate the ratio: $r = \frac{\Pr(x')}{\Pr(x_t)} \frac{\Pr(x_t \to x')}{\Pr(x' \to x_t)}$
        if $r > 1$ then
            accept: $x_{t+1} = x'$
        else
            Draw a random value $u \sim \text{Uniform(0,1)}$;
            if $u < r$ then  accept: $x_{t+1} = x'$
            else  reject: $x_{t+1} = x_t$
        end
    end
end
```

If the proposal distribution is symmetrical around 0, the ratio becomes $r = \frac{\Pr(x')}{\Pr(x_t)}$. This is the Metropolis algorithm.

The principle of the Metropolis-Hasting algorithm is to propose a new position in the parameter space based on a kernel distribution, evaluate the probability of the proposed new state, and accept or reject the move of the chain accounting for both the Likelihood ratio of the new position compared to the previous one as well as the probability of proposing the move and reversing it. A more detailed description of the algorithm can be found in Box 8.

If the jump distribution allows to reach every point of the parameter space, this algorithm guarantees that the target posterior distribution is the stationary distribution of the Markov chain. Nevertheless, the parametrization of the algorithm, in priority the jump kernels of the chain, has a very strong impact on the time (number of steps) needed to reach this equilibrium distribution and to obtain a sufficient sampling of the parameter space. If the jumps are too short, the chain might stay too long around local optima and not explore efficiently the parameter space in a realistic run time. Chains starting in different regions of the parameter space might not converge (might give different results). If the jumps are too long, the space is sampled almost randomly, proposing very unlikely regions of the parameter space. The acceptance rate is then very low and the movements of the chain look jagged with long plateaus separated by long distance jumps.

2.3.2 Approximate Bayesian Computations (ABC)

When the likelihood function is not available but data can be simulated using a model, approximate Bayesian computations (ABC, Tavaré *et al.* 1997; Pritchard *et al.* 1999; Beaumont *et al.* 2002) can be used to obtain the posterior distribution. The basis of ABC is the so-called rejection algorithm: (i) sample random parameter values in their respective marginal prior distributions, (ii) simulate a data set for each vector of parameters using the model, and finally (iii), comparing the simulated data sets to the observed one, reject the datasets that are further away from the observed data than a defined threshold ϵ (Figure 2.4).

Parameter estimation

The posterior density of model parameters given the observed data is estimated by the density of the parameters used to generate the datasets the closest to the observed data (i.e., not rejected). In most cases, the tolerance threshold ϵ needs to be greater than 0 in order to accept any simulation. However choosing a too great ϵ might lead to a strongly biased estimation of the posterior distribution. For small ϵ values, the relation between the parameters and the accepted simulated data can be considered linear. Therefore, Beaumont *et al.* (2002) proposed to apply a local weighted linear regression on the accepted data points in order to weight simulations depending on their distance to the observed data, and project the parameter values on the vertical of the observed data (Figure 2.4A). Because linearity cannot always be assumed, Blum and François (2010) proposed to apply a non-linear regression. These regression methods aspire to give an approximate answer to the following question: "knowing that we observe some value y', close to the data value y, when simulating with parameter x, which parameter should we use to observe exactly y." Such regressions can mitigate the effect of a strictly positive ϵ, but if ϵ is too large or the assumptions underlying the regression are not met, regressions can also have adverse effects on the results (Beaumont *et al.* 2002). Therefore, both the estimations of the posteriors before and after the regression should be studied (Figure 2.4B).

Model choice

Using ABC to estimate parameters relies on the use of a specific simulation model. Determining the appropriate model for a given dataset is not an easy task and can usually not be automated. However, ABC can be used to compare several models using the Bayes' factor. In that case, the same rejection procedure is applied until the last step where only the generating model is considered and not the parameter values. A first estimate of the Bayes Factor is given by the number of accepted simulations generated by one model divided by the number of accepted simulations generated by the other models. For model choice, Beaumont (2008) proposed to apply a weighted multinomial logistic regression and, similarly as for parameter estimation, Blum and François (2010) proposed using feed-forward neural networks. Whether one uses the simple rejection method or one of the available rejection-regression methods, the sensitivity of the results to the choice of ϵ should always be verified.

Refining ABC

Several questions arise when applying ABC to more complex models or datasets. Firstly, as the dimensionality of the data increases, the probability of simulating a dataset close to the observed data decreases. To alleviate this problem and to allow the method to remain computationally efficient, data can be summarized by a set of values that capture the effect of the model parameters on the data but is of lower dimension than the observed data itself: the summary statistics. If the summary statistics are sufficient, that is if they capture completely the information contained in the data about the parameters of the model, no error will be induced by this simplification. In this case, the estimated posterior density will be identical to the theoretical posterior density that could be obtained from the whole dataset. However, except in few particular cases (Sunnåker *et al.* 2013), finding sufficient statistics is not possible. Therefore, several methods have been developed to choose summary statistics that reduce the dimensionality of the data while remaining informative (although potentially non-sufficient): PCA, PLS (Wegmann *et al.* 2009), regression (Fearnhead and Prangle 2012). These methods use different statistical tools but have the same underlying principle. They aim to find, in an extended set of summary statistics, the optimal combination of statistics that decrease the amount of redundant information or noise irrelevant to the estimation while minimizing the loss of information. In all cases, the informativeness of the statistics is defined for a given model and a given set of parameter range and prior distribution. The analysis is performed on a data set simulated within this parameter range and prior distribution. Therefore, if the model and parameter range are badly chosen, information that is relevant to the real demographic processes might be deemed irrelevant for the simulated processes and this information will be lost. This problem is not specific to ABC but inherent to all model based estimation methods and should be taken into account. A possible way to verify such adverse effects of summary statistics is to use the estimated parameters to generate data and compare it to the untransformed observed data or at least to the extended set of summary statistics before applying any reduction. This problem is even more crucial and proposed solutions scarcer in the case of model choice (Robert *et al.* 2011). The two main approaches proposed in that case are the computationally intensive automatic summary statistic selection proposed in Chu *et al.* (2013) and the logistic regression based methods from Prangle *et al.* (2014). More recently, the ABC random forest method proposed by Pudlo *et al.* (2016) circumvents this issue by being more robust to overfitting and uninformative summary statistics.

This leads to a second concern raised by ABC, as well as by any other Bayesian statistic method: the choice of the prior distributions and parameter range. If the choice of the parameter range or the prior distribution is too narrow, the resulting posterior distribution will not reflect the likelihood function. Conversely, choosing too wide parameter range with uninformative priors might lead to another dimensionality problem, namely, the parameter space will not be explored well enough within a realistic amount of simulations. In practice, it means that the model choice can be affected by the chosen priors, especially if models with different parameters are compared.

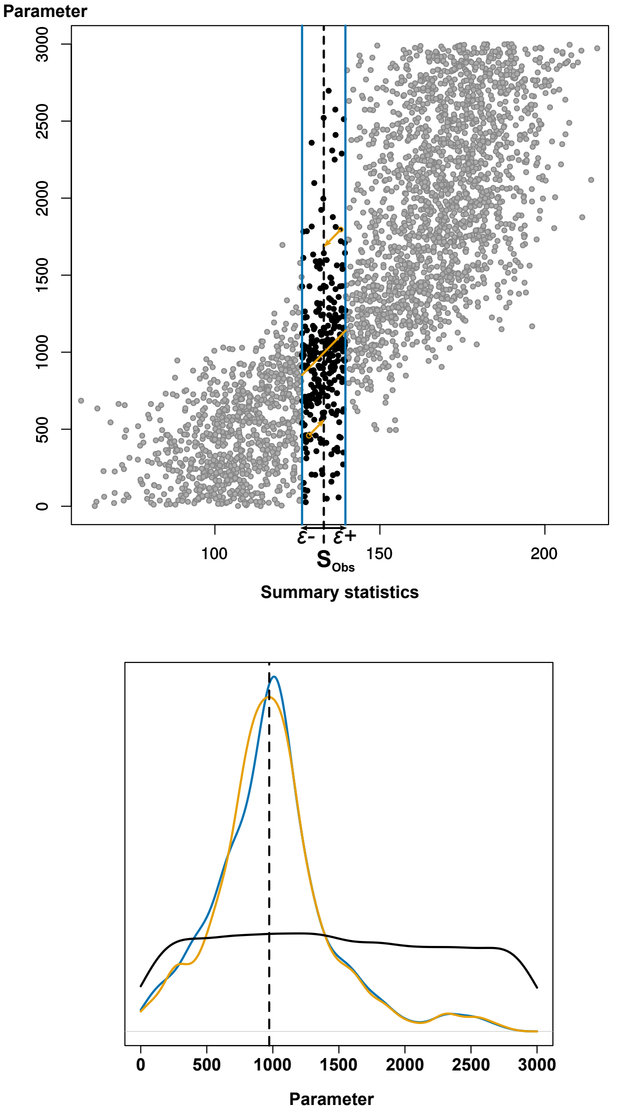


Figure 2.4: Schematic representation of a run of approximate Bayesian computations. (A) The simulated parameter value as a function of the resulting summary statistic. The gray points are the simulations rejected by the algorithm. Their summary statistic are too distant from the observed value marked by a black dashed line. The tolerance region ε around the observed value is delimited by the two blue lines. The orange line shows the slope of the local regression. As an example, arrows show how the regression would affect two points, circled in orange. (B) represents the uniform prior density from which the parameter is drawn (in black), the posterior density before the regression also called *truncated prior* (in blue), and the posterior density after the regression (in orange). The dashed line represents the mode of the posterior density, that is the estimated parameter value.

Part I

Plant domestication

Chapter 3

Rye population diversity and structure

3.1 Introduction

The archaeological evidence presented in the introduction already gives some insights into the history of rye. However, how genetic diversity was affected and how more recent populations fit in this history remains largely unknown as shown by the contradictions in the few previous genetic studies. In this chapter, we use microsatellite data to describe the diversity of 14 rye populations and investigate their structure and to shed light on the domestication and use diversification process.

3.2 Material and methods

3.2.1 Sampling and SSR genotyping

We chose 14 open-pollinated winter rye populations to represent rye growing areas in Europe and the Americas. Different gene banks and plant breeders provided the seeds that were propagated through several cycles of cross-pollination under isolation. We

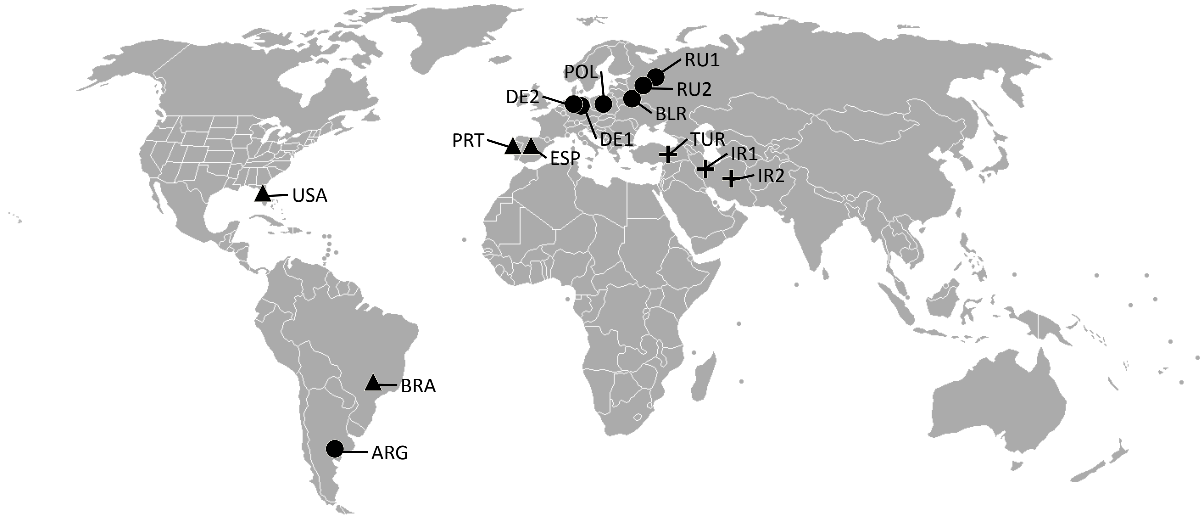


Figure 3.1: Geographical distribution of sampled rye populations. Crosses, circles, and triangles indicate weedy, grain and forage populations, respectively. For convenience, populations are named with the 3-letter ISO codes (International Organization for Standardization, ISO 3166) and in case of multiple populations per country we use the 2-letter ISO codes followed by a consecutive number.

randomly sampled between 37 and 45 S_0 plants per population, representing a total of 620 S_0 individuals (Figure 3.1; Table 3.1).

Genomic DNA of the 620 individuals was extracted from leaf samples as described in Rogowsky *et al.* (1991), and then all were genotyped with 32 unlinked and genome-wide distributed SSR markers following established protocols (Table 3.2). Briefly, separation of fragments by polymerase chain reaction (PCR) was carried out on a 3130xl Genetic Analyzer (Applied Biosystems Inc., Foster City, CA, USA). Alleles were assigned using the software GeneMapper v. 4.0 (Applied Biosystems Inc., Foster City, CA, USA). The software determined fragment lengths using size standards of a known length. All SSR-marker data necessary to reproduce the analysis as well as PCR primers and PCR conditions have been archived on Dryad (doi: http://dx.doi.org/10.5061/dryad.q0694).

Table 3.1: Populations under study and their genetic diversity. The level of genetic diversity of each population was estimated based on 32 SSR markers and described with the parameters number of alleles, number of private alleles (A_p), average number of effective alleles (A_e), observed heterozygosity (H_{obs}), gene diversity (expected heterozygosity, $\hat{H}$) and inbreeding coefficient (F_{IS}).

Code	Origin	Usage	Breeding level	Individuals	Alleles	A_p	A_e	H_{obs}	$\hat{H}$	F_{IS}	Population name
IR1	Iran	Weedy	Primitive rye	44	161	8	2.99	0.46	0.59	0.23	Altevogt 14160
IR2	Iran	Weedy	Primitive rye	45	233	17	3.89	0.52	0.67	0.21	IRAN GP.IX
TUR	Turkey	Weedy	Primitive rye	45	243	41	4.37	0.55	0.69	0.19	Türkischer Unkrautroggen
ESP	Spain	Forage	n.i.	37	174	8	3.24	0.47	0.61	0.22	R778 (‘Villablanca’*)
BRA	Brazil	Forage	n.i.	44	140	4	2.58	0.47	0.57	0.18	Centeio Branco
USA	USA	Forage	Variety	45	132	2	2.48	0.45	0.54	0.16	Florida Black
PRT	Portugal	Forage	Landrace†	45	199	9	3.22	0.52	0.63	0.15	R1008 (‘Malhadas’*)
ARG	Argentina	Forage	Landrace‡	45	87	0	1.81	0.29	0.45	0.35	Pico Gentario
DE1	Germany	Grain	Old variety	45	114	0	2.21	0.39	0.48	0.21	Carokurz
RU1	Russia	Grain	Landrace	45	125	5	2.62	0.43	0.57	0.23	Karelische Landsorte
RU2	Russia	Grain	Landrace	44	109	0	2.23	0.39	0.48	0.17	Leningrader Landsorte
BLR	Belarus	Grain	Variety	43	156	3	2.6	0.49	0.56	0.12	Belorusskaja
DE2	Germany	Grain	Old variety	45	118	2	2.27	0.39	0.48	0.18	Halo
POL	Poland	Grain	Improved landrace	45	120	0	2.19	0.4	0.49	0.16	Dankowskie Selekcyjne

n.i., no information. *Synonymous name. †Matos *et al.* (2001). ‡Stracke *et al.* (2003)

Most fragments displayed stepwise variation in length as expected from varying numbers of microsatellite repeats. The size of the rest of the fragments deviated from this expected pattern due to additional insertions or deletions outside of the microsatellite motif. To improve the SSR marker data quality, we checked allele assignments manually and set ambiguous results to ”missing data”. For the population genetics statistics, we considered null alleles as additional valid alleles for each marker. The global statistical trend remained similar whether null alleles were considered as valid or set to missing, for example, the populations with a higher expected heterozygosity conserved this characteristic (Table 3.3).

Table 3.2: Overview of the SSR markers used for genotyping of 14 rye populations. Information is given on source, chromosome, and parameters describing the variability of markers [number of alleles, PIC, private alleles (A_p), effective alleles (A_e), observed heterozygosity (H_{obs}), gene diversity (expected heterozygosity, Ĥ), population differentiation (F_{ST}, Jost's D)]

Marker	Source	Chrom.	Alleles	PIC	Ap	Ae	Hobs	Ĥ	F_{ST}	D
scm266	I	1R	6*	0.39	2	1.72	0.16	0.42	0.16	0.11
rms1280	K	1R	6	0.59	1	2.86	0.33	0.65	0.25	0.34
scm247	I	1R	6	0.70	1	3.93	0.30	0.75	0.19	0.41
rms1107	I	1R	21	0.68	6	3.30	0.43	0.70	0.16	0.28
rms1238	K	2R	8*	0.77	2	4.91	0.41	0.80	0.15	0.40
scm290	I	2R	7	0.57	1	2.63	0.43	0.62	0.17	0.24
rms1138	K	2R	19	0.62	5	2.86	0.53	0.65	0.17	0.26
scm276	I	2R	6	0.58	0	2.64	0.49	0.62	0.18	0.24
rms1230	K	2R	6*	0.73	0	4.21	0.22	0.76	0.23	0.46
rms1254	K	3R	9	0.59	1	2.72	0.42	0.63	0.22	0.29
rms1028	I	3R	22	0.78	7	4.95	0.65	0.80	0.19	0.45
rms1323	K	3R	20	0.52	9	2.18	0.46	0.54	0.17	0.18
scm294	I	3R	4	0.33	2	1.66	0.36	0.40	0.09	0.06
rms1026	I	4R	14	0.61	6	3.02	0.53	0.67	0.25	0.36
scm047	H	4R	2	0.30	0	1.60	0.28	0.37	0.20	0.11
rms1181	H	4R	4	0.41	0	2.10	0.33	0.52	0.13	0.14
rms1218	K	5R	6	0.67	1	3.57	0.86	0.72	0.11	0.23
rms1259	K	5R	8	0.73	2	4.19	0.23	0.76	0.25	0.48
scm260	I	5R	15	0.76	3	4.67	0.30	0.79	0.16	0.41
rms1205	K	5R	10*	0.65	3	3.16	0.27	0.68	0.21	0.34
rms1237	K	5R	13	0.79	3	5.17	0.55	0.81	0.24	0.54
rms1278	I	5R	13	0.65	3	3.29	0.53	0.70	0.21	0.34
rms1090	I	6R	14	0.65	4	3.27	0.53	0.69	0.19	0.33
rms1121	I	6R	33	0.91	9	11.30	0.73	0.91	0.16	0.67
scm107	H	6R	3	0.41	0	2.09	0.42	0.52	0.21	0.20
rms1197	K	7R	13*	0.67	4	3.37	0.23	0.70	0.16	0.30
scm322	I	7R	4	0.56	1	2.75	0.53	0.64	0.15	0.22

Table 3.2 – *Continued from previous page*

Marker	Source	Chrom.	Alleles	PIC	Ap	Ae	Hobs	$\hat{\mathrm{H}}$	F_{ST}	D
rms1018	I	7R	28	0.88	11	9.30	0.74	0.89	0.12	0.55
rms1187	K	7R	3	0.36	1	1.85	0.41	0.46	0.09	0.08
scm063	H	7R	3	0.54	0	2.60	0.53	0.62	0.18	0.24
rms1188	K	7R	8	0.57	2	2.68	0.20	0.63	0.20	0.27
rms1012	I	7R	40	0.92	9	12.80	0.70	0.92	0.15	0.70
		Sum:	374		99					
		Mean:	11.7	0.62	3.09	3.86	0.44	0.67	0.18	0.32
		Standard deviation:	9.2	0.16	3.10	2.61	0.17	0.14	0.04	0.16
		Min.:	2	0.30	0	1.60	0.16	0.37	0.09	0.06
		Max:	40	0.92	11	12.84	0.86	0.92	0.25	0.70

H: Hackauf and Wehling (2002); K: Khlestkina *et al.* (2004); I: Internal KWS LOCHOW GMBH;

*: occurrence of a null allele among the alleles of a marker.

Each marker was tested for deviations from the Hardy–Weinberg equilibrium (HWE) within populations using the X^2 goodness-of-fit test with a Benjamini–Hochberg correction for multiple testing (Table 3.4), and marker informativeness was measured as the polymorphism information content (PIC, Botstein *et al.* 1980). We defined the following groups of populations by rye main ancestral end use: grain, which assembled the populations of northeast European ancestry; forage, which assembled the populations of Mediterranean ancestry and weedy rye, for the weedy populations from the center of diversity (Table 3.1).

3.2.2 Genetic diversity and population structure

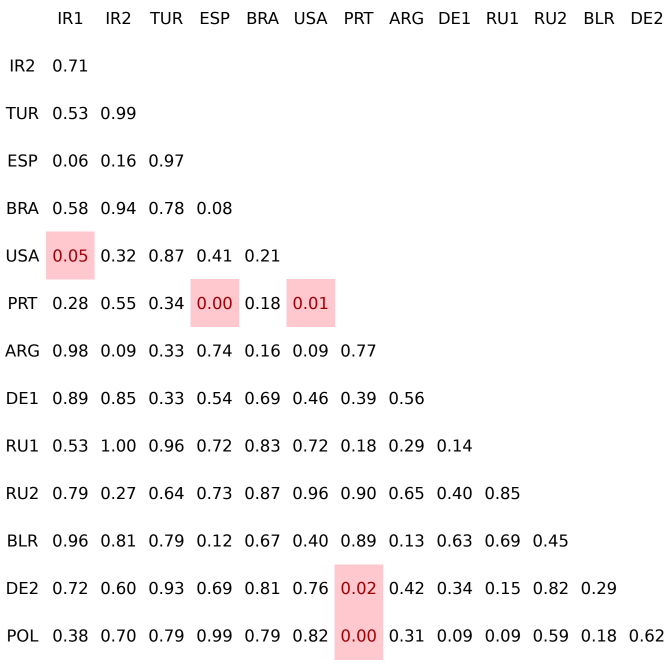
We computed for each population the total number of alleles, number of private alleles (A_p), number of effective alleles (A_e, Kimura and Crow 1964), observed heterozygosity (H_{obs}), gene diversity ($\hat{H}$, Nei 1987) and Garza–Williamson's M (Garza and Williamson 2001). The analysis of molecular variance (AMOVA, Excoffier *et al.* 1992) was computed with Arlequin v.3.5.1.3. The total variance was partitioned into components due

Table 3.3: Populations under study and their genetic diversity after removing null alleles. The level of genetic diversity of each population is estimated based on 32 SSR markers and described with the parameters number of alleles, number of private alleles (A_p), average number of effective alleles (A_e), observed heterozygosity (H_{obs}), gene diversity (expected heterozygosity, $\hat{H}$), inbreeding coefficient (F_{IS}), Garza-Williamson (GW) and the modified Garza-Williamson (GW*) statistics, calculated by using the range over all populations.

Code	Usage	# Individuals	Alleles	Ap	Ae	Hobs	$\hat{\mathbf{H}}$	$\mathbf{F}_{\mathrm{IS}}$	GW	GW*
IR1	Weedy	44	154	8	2.91	0.46	0.58	0.21	0.35	0.24
IR2	Weedy	45	229	17	3.81	0.53	0.67	0.21	0.39	0.31
TUR	Weedy	45	238	41	4.35	0.56	0.70	0.20	0.41	0.33
ESP	Forage	37	171	8	3.20	0.47	0.60	0.22	0.38	0.26
BRA	Forage	44	136	4	2.55	0.47	0.55	0.15	0.35	0.22
USA	Forage	45	128	2	2.42	0.46	0.53	0.13	0.40	0.21
PRT	Forage	45	196	9	3.21	0.52	0.62	0.16	0.38	0.28
ARG	Forage	45	84	0	1.81	0.30	0.40	0.25	0.41	0.17
DE1	Grain	45	110	0	2.13	0.39	0.47	0.17	0.36	0.20
RU1	Grain	45	121	5	2.56	0.44	0.56	0.21	0.36	0.20
RU2	Grain	44	104	0	2.19	0.41	0.46	0.11	0.36	0.18
BLR	Grain	43	153	3	2.59	0.50	0.56	0.11	0.39	0.23
DE2	Grain	45	115	2	2.24	0.40	0.47	0.15	0.36	0.19
POL	Grain	45	117	0	2.16	0.41	0.48	0.15	0.39	0.21

to differences among the three defined groups (V_{a}), differences among populations within those groups (V_{b}) and differences among individuals within populations (V_{c}). Variance components (V_{a}, V_{b} and V_{c}) were used to calculate the fixation indices (F-statistics; F_{CT}, F_{SC}, F_{ST}) according to Weir and Cockerham (1984). F-statistics were preferred to allele size-based measures of differentiation based on the results of the allele permutation test proposed by Hardy *et al.* (2003), which was performed across loci for each pair of populations using the software SPAGeDi 1.5 (Hardy and Vekemans 2002, Figure 3.2).

However, the allele size dataset, without null alleles, was used for the computation of Garza–Williamson's M. This statistic is used to detect recent bottlenecks and is based on the ratio of the number of microsatellite alleles to the observed range in allele size. Although its interpretation can vary depending on mutational models and demography, an M value lower than 0.68 would generally be considered significant (Garza and Williamson 2001). The genetic differentiation among weedy, forage and grain groups was denoted as



	IR1	IR2	TUR	ESP	BRA	USA	PRT	ARG	DE1	RU1	RU2	BLR	DE2
IR2	0.71												
TUR	0.53	0.99											
ESP	0.06	0.16	0.97										
BRA	0.58	0.94	0.78	0.08									
USA	0.05	0.32	0.87	0.41	0.21								
PRT	0.28	0.55	0.34	0.00	0.18	0.01							
ARG	0.98	0.09	0.33	0.74	0.16	0.09	0.77						
DE1	0.89	0.85	0.33	0.54	0.69	0.46	0.39	0.56					
RU1	0.53	1.00	0.96	0.72	0.83	0.72	0.18	0.29	0.14				
RU2	0.79	0.27	0.64	0.73	0.87	0.96	0.90	0.65	0.40	0.85			
BLR	0.96	0.81	0.79	0.12	0.67	0.40	0.89	0.13	0.63	0.69	0.45		
DE2	0.72	0.60	0.93	0.69	0.81	0.76	0.02	0.42	0.34	0.15	0.82	0.29	
POL	0.38	0.70	0.79	0.99	0.79	0.82	0.00	0.31	0.09	0.09	0.59	0.18	0.62

Figure 3.2: P-values of Hardy *et al.*'s (2003) allele permutation test across loci for each pair of population. A non-significant test means that allele identity-based statistics of population differentiation (e.g., F_{ST}) perform better than allele size-based ones (e.g., R_{ST} or $\delta\mu^2$). Significant values are indicated in red.

F_{CT}, among populations within groups as F_{SC} and among populations denoted as F_{ST}. The within-population fixation index F_{IS} was also computed. Jost's D (Jost 2008) was used as an alternative measure of population differentiation. Measures of genetic diversity and pairwise F_{ST} were aggregated over usage groups and levels of improvement (Table 3.6). A neighbor joining tree was drawn based on F_{ST} distances using the R-package "ape" v. 3.2 (Paradis *et al.* 2004). We calculated the pairwise distance between populations based on the proportion of shared alleles (Dps, Figure 3.3) as described in Bowcock *et al.* (1994). Population structure was further investigated by a Bayesian clustering approach implemented in the STRUCTURE software v. 2.2 (Pritchard *et al.* 2000). Burn-in period and Markov Chain Monte Carlo iterations were both set to 50,000. Ten runs were executed for each number of assumed subgroups K (K = 1, 2, ..., 15). Resulting membership

Table 3.4: P-values of χ^2 goodness-of-fit test for HWE within each population. Highlighted in gray are the p-values significant (HWE cannot be assumed) after a correction for multiple testing using the Benjamini and Hochberg procedure (1995) on each population independently, using a false discovery rate of 0.05. All population are significantly in disequilibrium using Fisher's method (conservative).

SSR marker	IR1	IR2	TUR	BRA	ESP	PRT	USA	ARG	BLR	DE1	DE2	POL	RU1	RU2
scm266	0.58	0.21	0.00	0.00	0.00	0.00	0.00	0.76	0.24	0.00	0.69	0.00	0.00	0.00
rms1280	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.11	0.19	0.92	0.46	0.69
scm247	0.00	0.00	0.38	0.00	0.00	0.41	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00
rms1107	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.69	0.00	0.19	0.00	0.32	0.42	0.17
rms1238	0.00	0.00	0.00	0.01	0.00	0.00	0.06	0.05	0.00	0.85	0.00	0.02	0.00	0.00
scm290	0.19	0.00	0.53	0.84	0.51	0.00	0.50	0.13	0.31	0.80	0.48	0.26	0.02	0.15
rms1138	0.58	0.04	0.06	0.11	0.44	1.00	0.01	0.04	0.00	0.57	0.28	0.25	0.70	0.20
scm276	0.51	0.30	0.12	0.01	0.92	0.76	0.99	0.00	0.82	1.00	0.88	0.90	0.00	0.75
rms1230	0.00	0.01	0.00	0.83	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00
rms1254	0.04	0.15	0.02	0.81	0.65	0.12	0.38	0.00	0.11	0.98	0.29	0.98	0.59	0.61
rms1028	0.02	0.03	0.67	0.04	0.52	0.53	0.77	0.07	0.03	0.93	0.00	0.02	0.32	0.54
rms1323	0.98	1.00	0.17	0.75	0.04	0.00	0.43	0.03	0.68	0.52	1.00	0.97	0.91	0.19
scm294	0.88	0.37	0.46	0.66	0.80	0.40	0.92	0.65	0.42	0.78	0.95	0.86	0.27	0.76
rms1026	0.11	0.15	0.82	0.25	0.33	0.00	0.45	0.01	0.10	0.66	0.83	0.46	0.07	0.75
scm047	0.41	0.84	0.37	0.01	0.06	0.17	0.30	0.01	0.51	0.29	0.92	0.42	0.63	0.75
rms1181	0.01	0.17	0.00	0.03	0.00	0.00	0.73	0.00	0.51	0.26	0.52	0.17	0.17	0.10
rms1218	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.07
rms1259	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.06	0.00
scm260	0.00	0.04	0.00	0.93	0.00	0.00	0.00	0.46	0.00	0.64	0.52	0.00	0.69	0.00
rms1205	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.44	0.02	0.00
rms1237	0.52	0.40	0.00	0.94	0.00	0.00	0.00	0.00	0.08	0.00	0.05	0.96	0.69	0.10
rms1278	0.01	0.41	0.04	0.47	0.67	0.95	0.73	1.00	0.94	0.13	0.92	0.62	0.17	0.18
rms1090	0.90	0.99	0.02	0.68	0.22	0.17	0.74	0.00	0.89	0.89	0.68	0.00	0.86	0.85
rms1121	0.28	0.36	0.68	0.33	0.41	0.87	0.00	0.01	0.19	0.29	0.59	0.98	0.33	0.80
scm107	0.02	0.28	0.70	0.03	0.18	0.03	0.44	0.00	0.28	0.46	0.50	0.11	0.94	0.69
rms1197	0.00	0.00	0.06	0.59	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.99	0.27	0.00
scm322	0.86	0.90	0.71	0.60	0.32	0.71	0.17	0.29	0.75	0.68	0.48	0.52	0.52	0.00
rms1018	0.10	0.56	0.00	0.60	0.56	0.83	0.60	0.06	0.98	0.04	0.98	0.87	0.69	0.62
rms1187	0.12	0.61	0.74	0.69	0.22	0.17	1.00	0.31	0.49	0.65	0.71	0.39	0.38	0.38
scm063	0.94	0.31	0.83	0.00	0.17	0.75	0.85	0.00	0.54	0.98	0.91	0.61	0.64	0.93
rms1188	0.00	0.00	0.95	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rms1012	0.20	0.30	0.91	0.00	0.98	1.00	0.05	0.00	0.04	0.00	0.58	0.71	0.04	0.00
# signif. tests	15	11	10	11	14	16	14	20	10	9	9	8	7	11

coefficients from each run were averaged by individual and visualized in a bar chart. The average log-likelihood (± standard deviation) was calculated for each K to deduce the most probable number of subgroups. Ten individual runs per K were plotted to check convergence. Additionally, ΔK was calculated according to Evanno *et al.* (2005) to determine the optimum K for the uppermost hierarchical level of structure (Figure 3.5).

3.3 Results

3.3.1 SSR genotyping

Genotyping of 620 individuals from 14 winter rye populations with 32 genome-wide SSR markers resulted in 374 alleles, including 99 private alleles (Table 3.1). One individual from each of the populations IR1, BRA and RU2 was excluded from the analyses due to more than 33% of missing data. On average, we observed 11.7 ± 9.2 (range: 2–40) alleles per locus, of which 7.07 ± 10.87 (range: 0–41) were private alleles (A_p). Across populations, the observed heterozygosity per SSR was 0.44 ± 0.17, and the expected heterozygosity $\hat{H}$ was 0.67 ± 0.14. H_{obs} was smaller than $\hat{H}$ for all of the 32 SSRs, except the locus rms1218 (Table 3.1). That deviation is expected due to the structure in populations. Individuals are more likely to reproduce with individuals of their own population increasing inbreeding. However, sites also showed deviation from Hardy–Weinberg equilibrium within populations (Table 3.4). Deviation from the Hardy–Weinberg equilibrium can be due to multiple factors among which population substructure and sampling bias are the most common. The latter might explain the high number of deviating sites in ARG and PRT.

3.3.2 Genetic diversity of populations

The weedy populations showed, as expected, a high genetic diversity with TUR having both the highest number of alleles per SSR (7.6 on average) and the highest heterozygosity ($H_{\text{obs}} = 0.55$ and $\hat{H} = 0.69$; Table 3.1). Contrary to expectations, we found no pattern of reduced diversity in varieties compared with landraces. Indeed, the five populations showing the lowest diversity (< 2.3 alleles per marker and $\hat{H} < 0.5$) included two varieties (DE1 and DE2) and three landraces (ARG, POL and RU2). Interestingly, four of the five populations were of northeast European ancestry and mainly used for grain production. The South American landrace ARG showed an especially low genetic diversity

($\hat{H} = 0.45$) and had the highest inbreeding coefficient ($F_{\mathrm{IS}} = 0.35$), indicating a possible small population size and strong bottleneck during its establishment or sampling bias. The lowest F_{IS} value was found for the grain population BLR ($F_{\mathrm{IS}} = 0.12$). AMOVA results showed that molecular variation was mainly (79.70%) found among individuals within populations as expected for cross-pollinated species, whereas variation observed among populations within groups explained 16.39% and the variance among groups only 3.91% of the total genetic variability (Table 3.5). Although variance among groups was small, permutation tests indicated that both populations and groups explained variance significantly better than random assignments ($P < 0.01$). We further investigated the differentiation among populations and groups below.

Table 3.5: AMOVA results including fixation indices $\mathbf{F_{CT}}$, $\mathbf{F_{SC}}$ and $\mathbf{F_{ST}}$ for the total population. The genetic differentiation among weedy/forage/grain groups is denoted as F_{CT}, among populations within groups as F_{SC} and among populations as F_{ST}.

Source of variation	Proportion of explained variation
Among groups	0.04
Among populations within groups	0.16
Within populations	0.80
Fixation indices	
FCT	0.04
FSC	0.17
FST	0.20

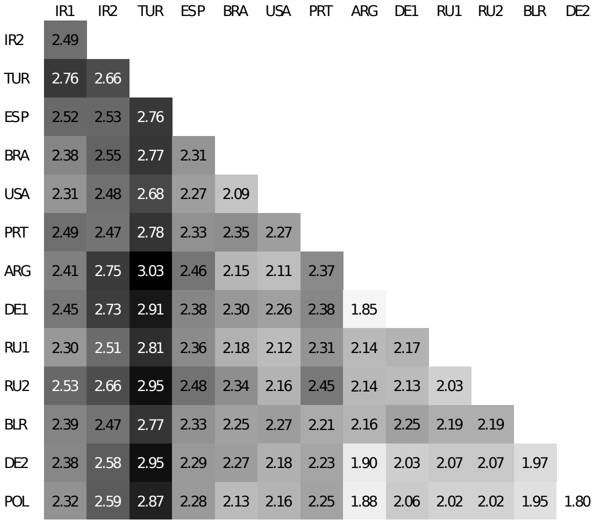
3.3.3 Genetic relationships between the populations

Pairwise F_{ST} and Jost's D values were calculated to indicate the level of differentiation between populations (Figure 3.6). As expected, we found a relationship among F_{ST}, Jost's D and genetic diversity, as pairwise population comparisons containing population PRT achieved lower values and comparisons containing DE1, RU2 or ARG revealed higher values than the other comparisons (Table 3.1; Figure 3.6 This resulted from the enhanced effect of drift in populations with small size that increased allele fixation and differentiation between populations (DE1, RU2 or ARG exhibited the smallest expected heterozygosity; Table 3.1). The relationship between diversity and ancestry appeared to be consistent overall as we observed low F_{ST} and Jost's D values within weedy and forage groups and among these two groups, while the grain group showed higher values within and among group differentiation. The neighbor joining tree (Figure 3.7) confirmed that these results showing a clear group containing the weedy populations, a relatively close but clearly separated group of three forage populations (BRA, ESP and USA) and distinct groups containing the grain populations as well as the two forage populations ARG and PRT. These three clusters could also be observed in the pairwise distance between populations based on the proportion of shared alleles (Figure 3.3). As expected for poorly differentiated populations with few private alleles, F_{ST} and Jost's D showed a similar ordering.

However, note that Jost's D indicated a generally higher differentiation than F_{ST} (Figure 3.6) because it was less biased by the high mutation rate of microsatellite markers (Jost 2008).

3.3.4 Population structure analysis

Individual-based grouping was investigated using STRUCTURE for values of K ranging from 2 to 15 (Figure 3.4) The log-likelihood curve for different K values did not show a clear plateau (Figure 3.5). However, the log-likelihood values started stabilizing around K =



	IR1	IR2	TUR	ESP	BRA	USA	PRT	ARG	DE1	RU1	RU2	BLR	DE2
IR2	2.49												
TUR	2.76	2.66											
ESP	2.52	2.53	2.76										
BRA	2.38	2.55	2.77	2.31									
USA	2.31	2.48	2.68	2.27	2.09								
PRT	2.49	2.47	2.78	2.33	2.35	2.27							
ARG	2.41	2.75	3.03	2.46	2.15	2.11	2.37						
DE1	2.45	2.73	2.91	2.38	2.30	2.26	2.38	1.85					
RU1	2.30	2.51	2.81	2.36	2.18	2.12	2.31	2.14	2.17				
RU2	2.53	2.66	2.95	2.48	2.34	2.16	2.45	2.14	2.13	2.03			
BLR	2.39	2.47	2.77	2.33	2.25	2.27	2.21	2.16	2.25	2.19	2.19		
DE2	2.38	2.58	2.95	2.29	2.27	2.18	2.23	1.90	2.03	2.07	2.07	1.97	
POL	2.32	2.59	2.87	2.28	2.13	2.16	2.25	1.88	2.06	2.02	2.02	1.95	1.80

Figure 3.3: Pairwise distance between populations based on the proportion of shared alleles (Dps). The proportion of shared allele is calculated as: $ps = \sum_{i=1}^{n} min\left(freq(i)_{Pop1}; freq(i)_{Pop2}\right)/n$, where n is the total number of alleles for all loci present in at least one of the two population samples. The corresponding distance is obtained as -ln(ps). Contrary to F_{ST} this measure of distance tends to increase with the diversity as can be seen here for BLR. Indeed the more alleles are present in a population the more likely many of them will not be shared.

10 indicating a possible optimal number of groups between 10 and 12. The differentiation of groups reflected the ancestral main end use and origin of the rye populations.

At K = 2, the optimal number of groups based on ΔK, the first subgroup (green) comprised the forage and weedy populations IR1, IR2, TUR, BRA, USA and ESP and was clearly separated from the second subgroup (red) of grain populations BLR, DE1, POL, DE2, RU1 and RU2. The forage population ARG was also found in this group of grain populations. Those subgroups were already observed in the pairwise F_{ST} and Jost's D statistics. The subgroup comprising forage and weedy populations showed a low level of differentiation, whereas the grain populations were highly differentiated from the forage and weedy as well as among themselves. The forage population PRT could not be assigned to either of the groups as the individuals exhibited mixed membership coefficients. When

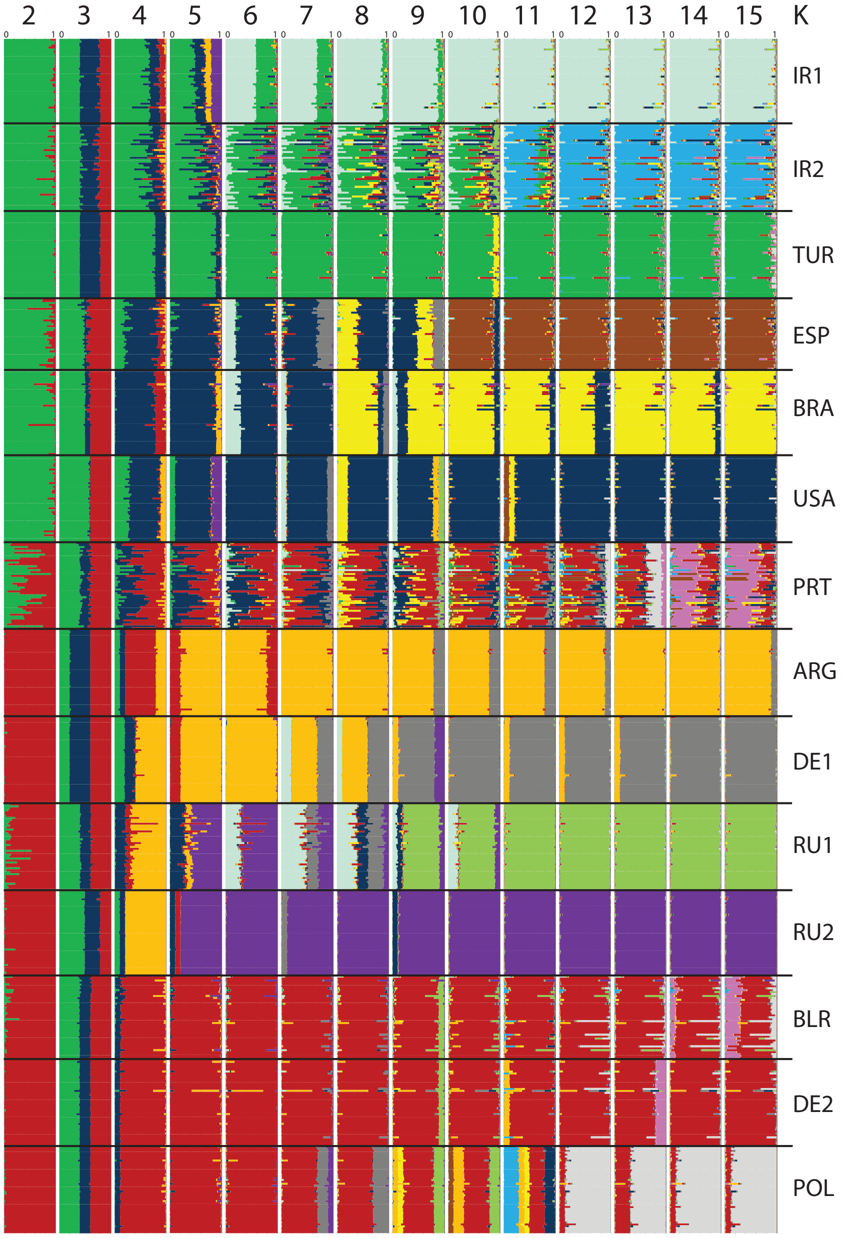


Figure 3.4: STRUCTURE results. Bar plots of tested numbers of subgroups are shown from left (K = 2) to right (K = 15). Individuals were plotted on the y-axis and have been sorted according to population assignment given on the right side. Each horizontal line represents the individual's proportion of membership to a given number of subgroups. The number of assumed subgroups is indicated on top.

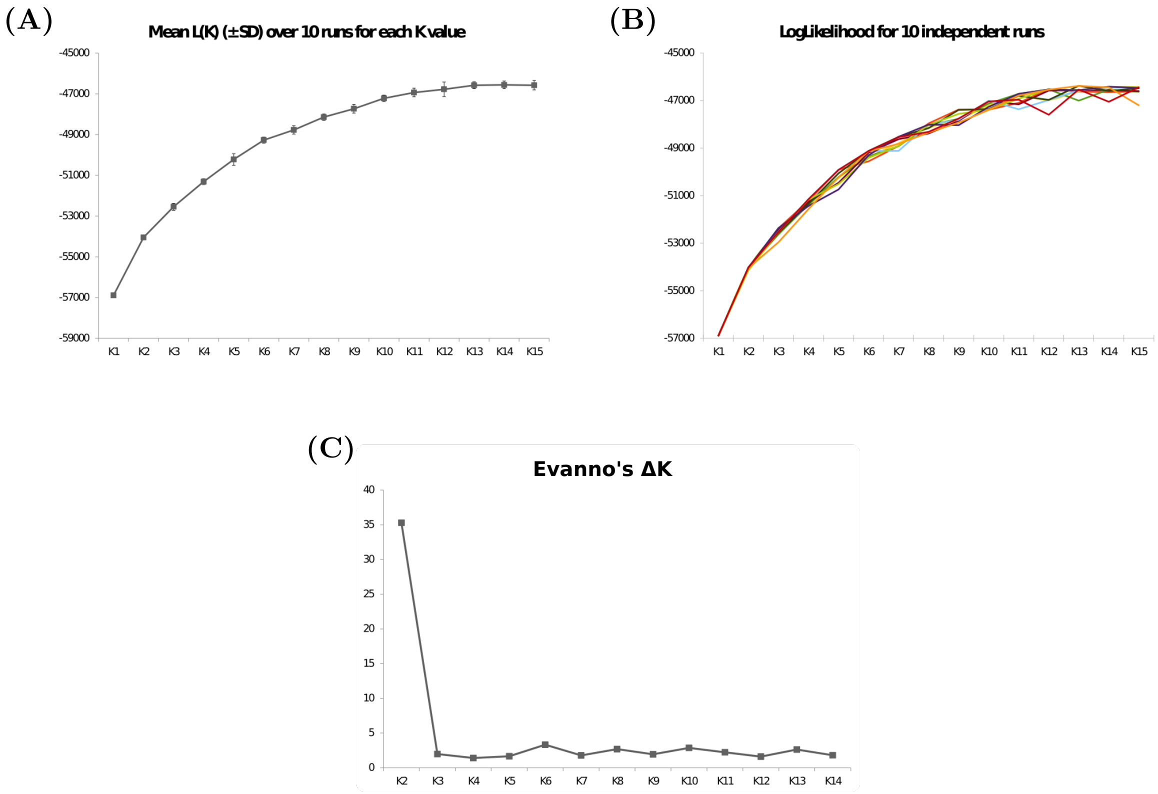
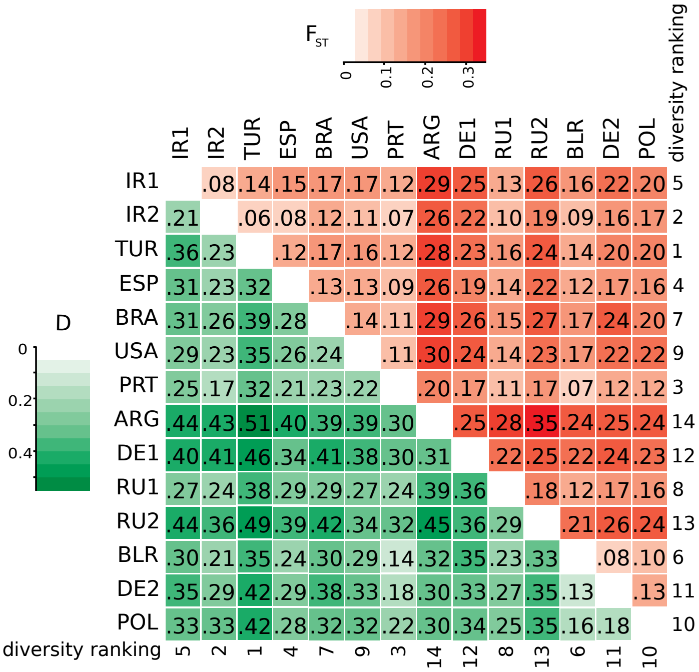


Figure 3.5: Graphical analyses of STRUCTURE's log-likelihood for numbers of assumed subgroups K from 1 to 15. (A) Average log-likelihood ± standard deviation. (B) Log-likelihood for ten individual runs to visualize convergence. (C) Evanno *et al.*'s (2005) ΔK (with $\Delta K = m|L''(K)|/s[L(K)]$) for K between 2 and 14.

increasing K from five to seven, several geographical subgroups appeared: (1) the three weedy rye populations IR1, IR2 and TUR, (2) the two Russian populations RU1 and RU2, (3) the eastern and central European populations BLR, POL and DE2, (4) the ESP population from Spain with BRA and USA and (5) the German population DE1 with the Argentinian population ARG. These clusters were similar to those observed in the neighbor joining tree (Figure 3.7). At K = 10, most of the populations appeared separated from each other except IR2 and TUR in the weedy group and BLR, DE2 and POL in the grain group. The high fragmentation of membership coefficients observed for the sampled individuals of IR2 and PRT reflected the very high diversity observed in these populations. At K = 12, only the two populations BLR and DE2 clustered together, agreeing with the low F_{ST} and Jost's D between these populations (Figure 3.6).



	IR1	IR2	TUR	ESP	BRA	USA	PRT	ARG	DE1	RU1	RU2	BLR	DE2	POL	diversity ranking
IR1		.08	.14	.15	.17	.17	.12	.29	.25	.13	.26	.16	.22	.20	5
IR2	.21		.06	.08	.12	.11	.07	.26	.22	.10	.19	.09	.16	.17	2
TUR	.36	.23		.12	.17	.16	.12	.28	.23	.16	.24	.14	.20	.20	1
ESP	.31	.23	.32		.13	.13	.09	.26	.19	.14	.22	.12	.17	.16	4
BRA	.31	.26	.39	.28		.14	.11	.29	.26	.15	.27	.17	.24	.20	7
USA	.29	.23	.35	.26	.24		.11	.30	.24	.14	.23	.17	.22	.22	9
PRT	.25	.17	.32	.21	.23	.22		.20	.17	.11	.17	.07	.12	.12	3
ARG	.44	.43	.51	.40	.39	.39	.30		.25	.28	.35	.24	.25	.24	14
DE1	.40	.41	.46	.34	.41	.38	.30	.31		.22	.25	.22	.24	.23	12
RU1	.27	.24	.38	.29	.29	.27	.24	.39	.36		.18	.12	.17	.16	8
RU2	.44	.36	.49	.39	.42	.34	.32	.45	.36	.29		.21	.26	.24	13
BLR	.30	.21	.35	.24	.30	.29	.14	.32	.35	.23	.33		.08	.10	6
DE2	.35	.29	.42	.29	.38	.33	.18	.30	.33	.27	.35	.13		.13	11
POL	.33	.33	.42	.28	.32	.32	.22	.30	.34	.25	.35	.16	.18		10
diversity ranking	5	2	1	4	7	9	3	14	12	8	13	6	11	10	

Figure 3.6: Pairwise $\mathbf{F_{ST}}$ (above the diagonal) and D (below the diagonal) values. Leading zeros of all values are not shown.

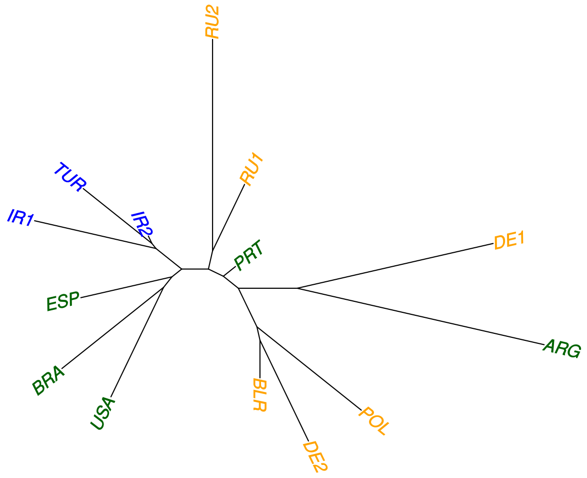


Figure 3.7: Neighbor joining tree based on F_{ST} distances. The populations are colored depending on their usage: blue for weedy, green for forage and orange for grain populations.

3.4 Discussion

In this study, 14 winter rye populations were genotyped with SSRs. These populations had different within population levels of diversity (number of alleles and heterozygosity) while F_{ST} and Jost's D values revealed differentiation among the 14 populations and confirmed the assumption of considerable population structure within cultivated rye. The STRUCTURE analysis revealed two main subgroups indicating a differentiation according to both geography and end use, which can be described as "southern European forage rye" vs. "northern European grain rye". The observed genetic diversity and population structure of a global collection of rye populations with different end uses and improvement levels suggest that (1) forage populations have reached a lower diversification level than grain ryes and (2) the strong structuration of populations according to geography and usage might explain discrepancies in previous studies.

3.4.1 Cultivated ryes are at different diversification stages

For any crop plant, it is assumed that genetic diversity decreases from its wild form over landraces to modern varieties (Feuillet *et al.* 2008; Yamasaki *et al.* 2005). Following this hypothesis, we tested whether the present sets of rye populations showed a decrease in genetic diversity concurrent with an increase in improvement status. We confirmed that primitive rye populations had a significantly higher level of genetic diversity than the landraces and varieties. Noticeably, the genetic diversity levels of landrace and variety populations were comparable; whereas differences in genetic diversity could be seen between grain and forage rye populations independently of the improvement status. Despite the high number of populations in this study, our statistical power to distinguish clearly between these two patterns was somewhat limited because the landraces and varieties were unequally distributed between grain and forage populations. There was a higher pairwise differentiation among grain than among forage rye populations (F_{ST} and Jost's D), which indicated that these groups probably followed different domestication and/or artificial selection paths. The lower diversity found in grain populations suggested that the differentiation might have been accelerated by successive bottlenecks that go hand in hand with domestication and selective breeding, especially in cases of adaptation to diverse environments as was the case between ARG and RU2 ($F_{\mathrm{ST}} = 0.35$). Therefore, we suggest that the early end use of rye explains its peculiarity compared with most other cereals. In northeast Europe, rye was used exclusively for grain early on and was, therefore, submitted to a longer and more intensive selection consisting of successive bottlenecks and diversification of populations for local adaptation. Conversely, in southern Europe, where wheat performed better, rye cultivation was neglected, and it was mainly used as forage or remained as a weed among other crops and selection was less intensive. This divergent end use and breeding is much more ancient than the distinction between landraces and varieties and had a much stronger impact on diversity patterns and among population differentiation. Our results might explain the discrepancies found in earlier studies on rye as highlighted in the introduction. In fact, studies sampling from either forage or grain populations or across both usages would obtain very different results when comparing genetic diversity. This is especially true when stratifying the data based on the landrace or variety status of the populations (Table 3.6).

Table 3.6: Genetic diversity and F_{ST} aggregated (A) over usage groups and (B) over levels of improvement.

(A)		Alleles/site	Hobs	$\hat{\mathrm{H}}$	Allelic Range	GW
Weedy	Mean	9.91	0.52	0.69	34.22	0.42
	s.d.	7.41	0.19	0.16	48.45	0.19
Forage	Mean	7.59	0.48	0.63	22.75	0.41
	s.d.	5.82	0.21	0.16	24.94	0.18
Grain	Mean	6.44	0.40	0.60	28.19	0.37
	s.d.	4.26	0.16	0.15	43.77	0.17

	Weedy	Forage
Forage	0.059	
Grain	0.093	0.062

(B)		Alleles/site	Hobs	$\hat{\mathrm{H}}$	Allelic Range	GW
Weedy	Mean	9.91	0.52	0.69	34.22	0.42
	s.d.	7.41	0.19	0.16	48.45	0.19
Landraces	Mean	7.03	0.41	0.61	29.84	0.38
	s.d.	4.94	0.18	0.16	44.74	0.18
Varieties	Mean	6.25	0.44	0.61	26.50	0.40
	s.d.	4.09	0.17	0.16	43.71	0.18

	Weedy	Landraces
Landraces	0.087	
Varieties	0.081	0.033

3.4.2 Impact of population management and the signature of bottlenecks

Our analyses highlighted possible issues concerning the sampling and management of these various populations. Despite generally high levels of diversity, all populations, including weedy ones, exhibited signs of recent bottlenecks with low GW values. The fact that the three groups have the same average value for the GW statistic (0.38) suggests that the detected bottlenecks might be due to the maintenance of accessions in seed banks rather than previous population history. The ARG sample showed a relatively low genetic diversity and comparatively high inbreeding. Both facts might indicate that a recent bottleneck effect occurred in the original production area of South America. We hypothesize, however, that the genetic bottleneck is an artifact, caused by small and biased sampling of the original South American population, or because of recurrent bottlenecks during the maintenance of accessions in seed banks. Similar conclusions could be drawn for DE1, an old German variety, and RU2, a landrace of northwest Russia, although the effects on genetic diversity were less severe than for ARG. Another point of interest in our study is the unexpected shared ancestry between populations. For example, the grouping of ARG with DE1 among the European grain rye populations in the STRUCTURE analysis was unexpected, as DE1 originates from northern Germany and ARG from Argentina. It is noteworthy that a common ancestry is not known between these populations. In addition, STRUCTURE suggests that a high admixture occurred between PRT and other populations possibly resulting from severe seed and/or pollen contamination. In summary, genetic bottlenecks can occur in population management due to few seeds collected from the original population, limitations during seed propagation in gene banks (Börner *et al.* 2005; Chebotar *et al.* 2003) or because of the small number of seeds distributed by gene banks and made available for research. Moreover, introgression/contamination likely explains unexpected genetic similarities among populations although this cannot be confirmed as the breeding history is often unknown.

3.4.3 High diversity in weedy ryes

The three weedy rye populations in our study represented an area of Turkey and Iran, a region considered to be within the cereals' domestication center (Badr *et al.* 2000; Behre 1992; Khush 1963; Nesbitt and Samuel 1998) and the center of maximum diversity for rye with regard to cytological and morphological aspects (Khush 1963). We found that, as expected, the three populations of weedy rye IR1, IR2 and TUR showed diversity parameters higher than the cultivated populations in this study since we used wild grown populations, which are more closely related to the wild rye ancestor and have most probably not been subjected to strong bottlenecks. The comparison of membership coefficients from the STRUCTURE analysis among IR2 individuals revealed high heterogeneity within this population. This observation is not surprising since the original seed collection from 1930 was composed of seeds from individuals growing in a wide area around Elburz-Karaj in Iran, and these were taken together as one population (Kranz 1957; Kuckuck 1956). In contrast, the weedy population TUR was collected from a single wheat field (Hartwig H. Geiger, personal communication). We expected individuals distributed over a wider area to be more genetically diverse and to deviate more from HWE compared with individuals from a population in close spatial proximity. Despite the sampling differences, IR2 and TUR showed comparable levels of diversity although IR2 exhibited an excess of rare alleles.

Chapter 4

Rye population demography inference

4.1 Introduction

In the previous chapter, we described the diversity and structure of 14 rye populations without making hypotheses about their demographic history. However, statistics such as F_{ST} and diversity are strongly influenced by population size and other demographic events through genetic drift. Moreover, we observed a strong impact of the end use on diversity but it is still unknown when the primary use diversification took place.

In this chapter, we formally model different demographic scenarios to investigate the demographic history of 12 of the previously introduced 14 rye populations. The Portuguese (PRT) and Argentinean (ARG) populations showed atypical characteristics of very high and very low diversity, respectively, suggesting a demographic history that could not be described by the same type of model.

4.2 Material and Methods

Different scenarios of population split were studied using ABC. This method allowed us to compare demographic scenarios and estimate population parameters using prior knowledge about rye history without evaluating the likelihood function analytically. Due to the large number of populations analyzed (12), the number of possible demographic models of population split was very large. Moreover, preliminary ABC analyses demonstrated that models allowing for complex relationships between populations had poor reliability (see Discussion and Figure 4.7). To maximize the statistical power of the ABC method, we studied a simplified model with three populations defined as one weedy and two derived populations (Figure 4.1). Each of these population trios was composed of one of three ancestral weedy populations (IR1, IR2 and TUR), one of the derived grain populations (BLR, DE1, POL, DE2, RU1 and RU2) and one of the forage populations (BRA, ESP, and USA). The weedy population of each trio was characterized by the population size Na, while the derived grain and forage populations were characterized by sizes Ng and Nf, respectively. The grain and forage populations were founded from the weedy population at times-to-merger T2Mg and T2Mf generations ago, respectively. The founding event was modeled as a bottleneck, characterized by its length BTLg (or BTLf) and by its strength BTNg (or BTNf), where we define the strength of a bottleneck as the ratio of the population size before and after the change. At times TSa, TSg and TSf ago, the weedy, grain and forage populations underwent a short bottleneck of strength PSa, PSg and PSf, respectively. This recent bottleneck was introduced to mimic the effects of repetitive sampling and gene bank conservation (Figure 4.1; Table 4.1).

All bottlenecks were modeled as stepwise population size changes. Our scenario modeled the origin of rye in the Fertile Crescent that was assumed to be a weedy population, and subsequent bottlenecks associated with the export and spread of rye as grain or forage to diverse parts of the world. Based on archaeological studies that show that rye was brought from the Fertile Crescent to northern Europe through an eastern migration

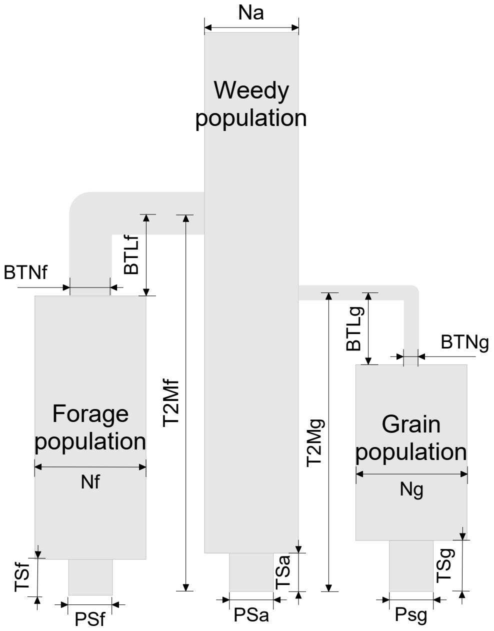


Figure 4.1: Graphical representation of the simulated demographic model. A forage (f) and a grain (g) population split "time-to-merger" (T2Mf and T2Mg) years ago, respectively, from a weedy population of constant population size Na. At the time of the split, the forage and the grain populations suffered a stepwise bottleneck (BTN) of strength BTNf and BTNg, respectively, over one generation, following which they have constant population sizes Nf and Ng, respectively.

route (via the Caucasus), we excluded models where the northern grain populations came indirectly from the ancestral weedy ones through the southern forage rye lineage (Zohary *et al.* 2013).

We used ABCtoolbox (Wegmann *et al.* 2010) to perform the ABC and 2,000,000 simulations were performed with the coalescent simulator fastsimcoal (Excoffier and Foll 2011). The model was defined by the 15 parameters described above for which we assumed prior distributions with wide bounds as information on rye domestication was scarce (Table 4.1). We simulated 32 independent SSR loci following a generalized stepwise model (GSM) with a mean mutation rate (MU) drawn from a log-uniform distribution between 10^{-5} and 10^{-2}. These priors are consistent with mutation rates obtained from plant species such as wheat species, *Triticum turgidum* (Thuillet *et al.* 2002) and maize, *Zea mays* (Vigouroux *et al.* 2002). The average proportion ($\mathrm{P_{GSM}}$) of mutations that affected the allele size by more than one step was drawn from a uniform distribution between 0.2 and 0.5. The mutation rate at each locus was drawn independently from a gamma distribution of mean MU and shape ALPHAMU, the latter varying between 2 and 30. This allowed for heterogeneity in mutation rates among loci (Excoffier *et al.* 2005; Xu *et al.* 2005). Similarly, $\mathrm{P_{GSM}}$ per locus was drawn from a gamma distribution of shape ALPHAP with uniform priors between 1 and 4. The range for the prior distribution of the times-to-merger was chosen based on archaeological findings on rye domestication (Behre 1992), namely that forage and grain populations were established no more recently than 1,000 but not more than 15,000 years ago. Since rye is an annual plant, years were simply considered equivalent to generations. The simulated data were summarized by the mean and standard deviation of genetic diversity and differentiation statistics over the 32 loci. The statistics for each population are the number of alleles, expected heterozygosity and private alleles; and over all populations, we used the average, sum and standard deviation of the number of alleles, heterozygosity, Jost's D over all populations and pairwise F_{ST} (detailed list can be found in Table 4.2).

All statistics were calculated with arlsumstat (Excoffier and Lischer 2010) except private alleles and Jost's D that were calculated with a custom Perl script.

Table 4.1: Definition and prior distribution of the model parameters.

Parameter name	Parameter definition	Prior distribution
ANCESTRAL SIZE (Na)	Diploid effective size of the ancestral population	Log-uniform [100; 30000]
GRAIN DEME SIZE (Ng)	Diploid effective size of the grain population	Log-uniform [100; 15000]
FORAGE DEME SIZE (Nf)	Diploid effective size of the grain population	Log-uniform [100; 15000]
GRAIN TIME TO MERGER (T2Mg)	Time since the split of the grain population from the ancestral one (years ago)	uniform [1000; 15000]
FORAGE TIME TO MERGER (T2Mf)	Time since the split of the forage population from the ancestral one (years ago)	uniform [1000; 15000]
GRAIN BOTTLENECK RATIO (BTNg)	Size of the grain deme during the one generation bottleneck, relative to its present size	Log-uniform [0.01; 0.5]
FORAGE BOTTLENECK RATIO (BTNf)	Size of the forage deme during the one generation bottleneck, relative to its present size	Log-uniform [0.01; 0.5]
GRAIN BOTTLENECK LENGTH (BTLg)	Span of the grain bottleneck after split from the ancestral population	Log-uniform [1; 2000]
FORAGE BOTTLENECK LENGTH (BTLf)	Span of the forage bottleneck after split from the ancestral population	Log-uniform [1; 2000]
ANCESTRAL SAMPLING RATIO (PSa)	Strength of the most recent bottleneck of the ancestral population	Log-uniform [1; 100]
GRAIN SAMPLING RATIO (PSg)	Strength of the most recent bottleneck of the grain population	Log-uniform [1; 100]]
FORAGE SAMPLING RATIO (PSf)	Strength of the most recent bottleneck of the forage population	Log-uniform [1; 100]
ANCESTRAL SAMPLING TIME (TSa)	Time since the start of the bottleneck of the ancestral population (years ago)	Log-uniform [1; 100]
GRAIN SAMPLING TIME (TSg)	Time since the start of the most recent bottleneck of the grain population	Log-uniform [1; 100]
FORAGE SAMPLING TIME (TSf)	Time since the start of the most recent bottleneck of the forage population	Log-uniform [1; 100]
MU	Mean mutation rate over loci	Log-uniform $[10^{-5}; 10^{-2}]$

Table 4.2: Summary statistics for model choice and parameter estimation.

Abbreviation	Description	
K 1	Mean number of alleles per site within each population (of a trio)	m,e
K 2		
K 3		
Ksd 1	Standard deviation of the number of alleles per site within each population	e
Ksd 2		m,e
Ksd 3		
mean K	Mean number of alleles per site over all 3 populations	e
sd K	Standard deviation of the mean number of alleles per site over all populations	e
tot K	Mean total number of alleles over all 3 populations	e
GW 1	Mean per site Garza-Williamson's M for each population	e
GW 2		m,e
GW 3		
GWsd 1	Standard deviation of the mean per site Garza-Williamson's M for each population	m,e
GWsd 2		
GWsd 3		
mean GW	Mean per site Garza-Williamson's M over all 3 populations	e
sd GW	Standard deviation of the mean per site Garza-Williamson's M over all 3 populations	m,e
tot GW	Mean total Garza-Williamson's M over all 3 populations	e
H 1	Mean per site heterozygosity for each population	m,e
H 2		
H 3		
Hsd 1	Standard deviation of the mean per site heterozygosity for each population	e
Hsd 2		m,e
Hsd 3		
mean H	Mean per site heterozygosity over all 3 populations	e
sd H	Standard deviation of the mean per site heterozygosity over all 3 populations	e
tot H	Mean total heterozygosity over all 3 populations	m,e
FST 2 1	Pairwise Fst for each possible pair of population	m,e
FST 3 1		
FST 3 2		
Jost D	Jost's D over all populations	m,e
Private K 1	Total number of private alleles within each population	m,e
Private K 2		
Private K 3		

m: model choice procedure, e: parameter estimation

4.2.1 "Scenario" choice

For each trio of populations, we performed a "model choice" procedure to distinguish between two scenarios: (1) the forage population split from the weedy population at an earlier time than the grain population so that T2Mf > T2Mg and conversely (2) the weedy-grain split occurred before the weedy-forage split so that T2Mg > T2Mf. We computed acceptance rates for each scenario based on the simulations associated with the smallest Euclidean distances to the observed data (Pritchard *et al.* 1999). For a given relative distance δ to the observed data, we computed the Bayes Factor (BF) for scenario 1 against scenario 2 by dividing the number of retained simulations in the respective scenarios. In very rare cases, the two split times were equal, and these were not considered in the model choice. The model choice procedure was done using a range of relative distances δ (from 0.005% to 1%) as indicated in Pritchard *et al.* (1999). For each trio of populations, we manually checked that the statistical support for one scenario did not depend on δ (Figure 4.2). We summarized the Bayes Factor for scenario 1 against scenario 2 for all trios of populations using the 5,000 simulations fitting best (namely, a relative distance of $\delta = 0.25\%$). The dimensionality of the summary statistics was reduced using a logistic regression on 500,000 simulations (Prangle *et al.* 2014). Statistics included in the regression were chosen based on significance and Bayesian information criterion (Table 4.2).

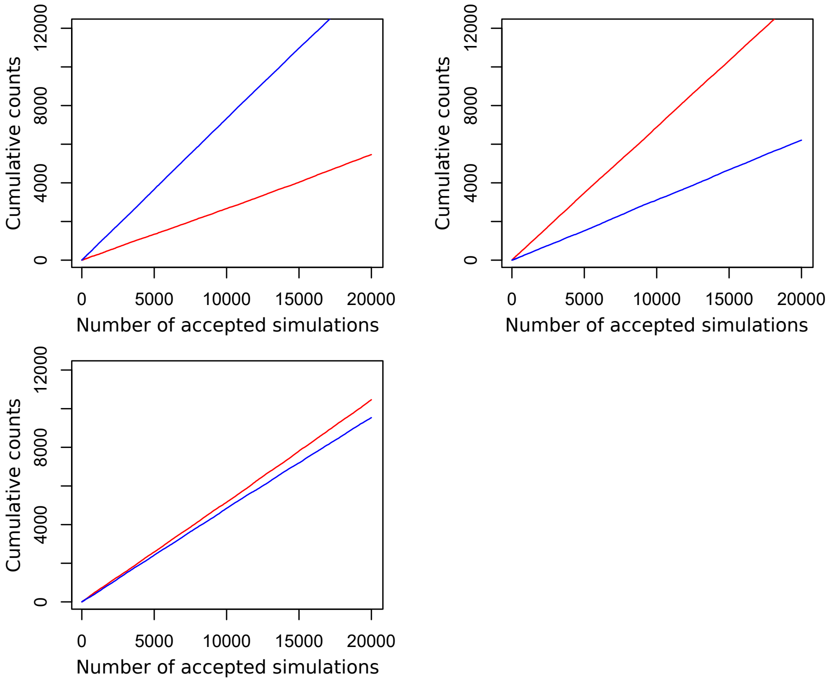


Figure 4.2: Cumulative counts of accepted simulations for each scenario depending on the total number of accepted simulations for four representative trios taken as examples. Blue line: number of accepted simulations from Scenario 1 (earlier forage split); red line: number of accepted simulations from Scenario 2 (earlier grain split).

Acceptance rate ratios do not seem to vary much depending on the total number of accepted simulations. Either the models are clearly differentiated or not (e.g., last figure down). In the last case the ratio will always stay close to 1. None of the trios presented curves that are apart for a small number of simulations but cross or come close together before 5000 accepted simulations.

4.2.2 Parameter estimation

For each trio of populations, we estimated all parameters of our model but reported the results for the five most important: the population sizes (Na, Nf and Ng) and the times-to-merger for the grain and the forage populations (T2Mg and T2Mf). The dimension of the summary statistics was reduced using a linear regression for each of the five estimated parameters as well as for the mutation rate per the method of semi-automatic approximate Bayesian computation proposed by Fearnhead and Prangle (2012). The coefficients of the linear regression were estimated using 500,000 simulations (Table 4.2). The Leuenberger and Wegmann (2010) post-sampling GLM adjustment was used to estimate the posterior probability of the parameters, based on the 0.25% of the simulated datasets (i.e., 5,000 simulations) closest to the observed data. Smaller relative distances (0.05% or 0.1%) gave similar results but less smooth densities. The shapes of the posterior distributions for each of these parameters are reported in Figure 4.5, and the mode and credibility intervals are given in Table 4.3.

4.2.3 Validation of the ”scenario” choice and estimation procedure

We evaluated the power of the ABC method to discriminate between scenarios and to estimate parameters by analyzing 1,500 pseudo-observed datasets (PODs). The PODs were sampled from the model and prior distribution as described above and transformed using the regression coefficient used for transforming the summary statistics of the observed data. We evaluated the number of times that the correct scenario was found for a trio of populations, the so-called confusion matrix (Bertorelle *et al.* 2010), and the difference between the estimated and the true parameter value (as the mean percent error and the root relative mean square error).

Table 4.3: Estimations, 50% and 95% highest probability density (HPD) of the demographic parameters. For each population, is given the average of the parameter estimation (mode) and ranges over all trios this population is part of.

Estimated parameter	**Pop.**	**Average marginal mode**	**Range of estimations over all trios**	**Min – max HPD50**	**Min – max HPD95**
Ancestral deme size	IR1	22958	18146 – 27130	13421 – 29998	6639 – 29998
	IR2	26558	20065 – 30001	15869 – 29998	8394 – 29998
	TUR	26715	21456 – 30001	16410 – 29998	8118 – 29998
Grain deme size	BLR	9505	6163 – 11151	4070 – 14136	1828 – 15000
	DE1	8767	7363 – 10509	5008 – 13723	2387 – 15000
	POL	9372	6540 – 11833	4448 – 14561	2058 – 15000
	DE2	9226	6540 – 11487	4448 – 14561	2120 – 15000
	RU1	9273	6163 – 12189	4192 – 14999	1939 – 15000
	RU2	8705	7148 – 10826	4862 – 13723	2317 – 15000
Forage deme size	BRA	10070	5983 – 12933	3951 – 15000	1774 – 15000
	USA	9611	5474 – 12189	3614 – 15000	1672 – 15000
Grain time to merger (years ago)	BLR	2698	2325 –2988	1249 – 4396	1000 – 8704
	DE1	2933	2491 – 3402	1414 – 5059	1000 – 9615
	POL	2629	2160 – 3402	1166 – 5059	1000 – 9533
	DE2	2698	2077 – 3568	1083 – 5142	1000 – 9615
	RU1	2339	1994 – 2657	1083 – 3899	1000 – 8787
	RU2	3099	2491 – 3734	1331 – 5556	1000 – 10527
Forage time to merger (years ago)	BRA	2993	2574 – 3651	1331 – 5556	1000 – 9947
	USA	2634	2160 – 3237	1166 – 4728	1000 – 8953

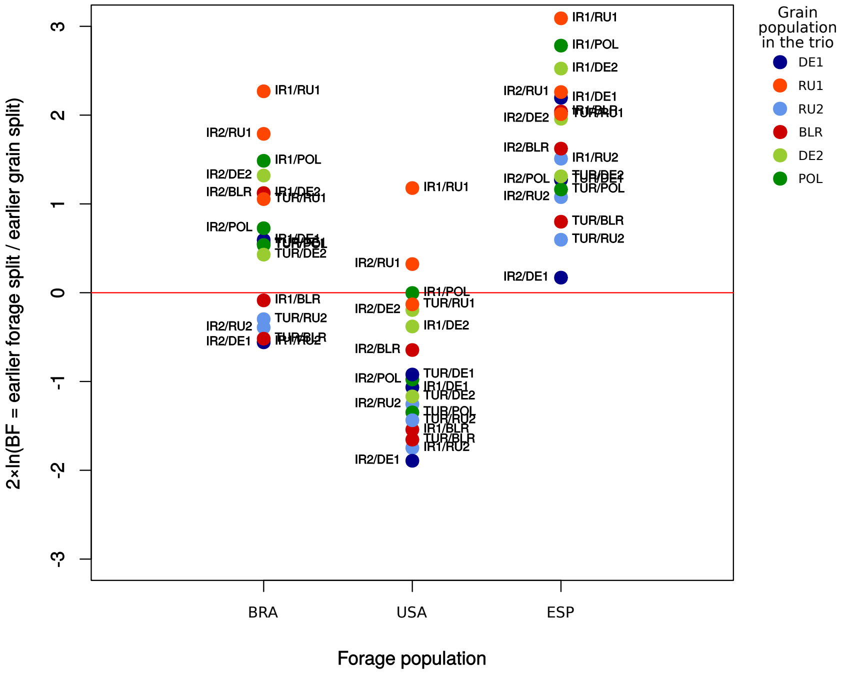


Figure 4.3: Graphical summary of the ABC results regarding forage and grain time-to-merger. Bayes Factor (BF) for Scenario 1: T2Mf > T2Mg over Scenario 2: T2Mf < T2Mg for each trio of populations (weedy, grain and forage). The BF has been estimated as the ratio of the number of accepted simulations for each scenario when taking 5000 simulations fitting best (d = 0.25 %). The dots above the horizontal (red) line represent a BF > 1, indicating that Scenario 1 (i.e., the split of the forage population is older than split of the grain population) is more strongly supported by the trio under consideration than Scenario 2 (i.e. the split of the grain population is older than split of the forage population). The distance to the BF = 1 horizontal line indicates the strength of evidence for the given scenario. Dots on or close to the BF = 1 line indicate a lack of support for one or the other order of split.

4.3 Results

As we used a common set of simulations for all population trios, the Spanish forage population ESP was kept in the model choice procedure but was not used for parameter estimation because of its smaller sample size (37 individuals). A smaller sample size could potentially lead to an overestimation of the time-to-merger due to incommensurable summary statistics (e.g., number of alleles, private alleles and F_{ST}). Our power analysis of the model choice procedure showed that we had moderate power to discriminate between the two scenarios (percentage of PODs attributed to the wrong scenario was 32%). Additionally, we achieved high power in only some of our parameter estimations as shown by low (8% for population sizes) to high (55% for times) mean percent error (Table 4.4 and Table 4.5; Figure 4.6). The ABC model selection indicated different split orders of the forage and grain populations depending on the considered trios, as indicated by the BF per trio in Figure 4.3 (the forage population is on the x-axis). Most of the trios involving the USA forage population had BF < 1; therefore, we concluded that the split of this population from the weedy group occurred more recently than splits of the other two forage populations. The grain populations DE1 and RU2 consistently split earlier than the three forage populations from the different weedy populations. In contrast, the grain population RU1 split later than all of the forage populations (Figure 4.3). We summarized the estimation results in Figure 4.4, by representing each mode of the marginal posterior distributions of the time-to-merger for the forage and grain populations. Table 4.3 indicates the minimum and maximum values of highest posterior density (HPD) 50% and 95% calculated over all posterior distributions of a given time split. Note that, as expected, the similar estimated values of times-of-split of the grain populations BLR, POL and DE2 were in line with the pairwise F_{ST}, Jost's D values and the STRUCTURE analysis in Chapter 3. Indeed, two populations showing a recent split from the ancestral weedy population would exhibit lower genetic differentiation between them than if they had diverged a longer time ago due to the action of genetic drift. These results would also fit a scenario where a single ancestral population deriving from the weedy population would have later on split to give rise to those three grain populations. The ancestral population sizes of the weedy group were estimated to be higher than those of the incipient populations. IR1 was estimated to be the smallest weedy population but with a higher effective population size of N = 23,000 than all grain and forage populations, whose effective population sizes were found to be in the range of 8,700 and 10,100. The population sizes matched the overall expectations based on heterozygosity, except for RU1 that had a lower population size than expected (Table 3.1 and Table 4.3).

Table 4.4: Confusion matrix for the model choice procedure. The values are in percentage of the total number of 1,250 pseudo-observed datasets (PODs) scenarios simulated for each order of split.

		Estimated	
		Forage older	Grain older
True	Forage older	0.69	0.31
	Grain older	0.33	0.67

Table 4.5: Error measures: mean percent error (MPE), relative mean square error (rMSE) and root relative mean square error (RrMSE) for the estimation of the demographic parameters. The values are calculated on the estimation of the logarithm of population sizes and grain and forage time to merger from 1,500 pseudo-observed datasets (PODs) randomly drawn from the prior distributions.

	Ancestral Size	Forage Size	Forage time to merger	Grain Size	Grain time to merger
MPE (%)	5.13	8.32	55.84	8.52	54.35
RMSE (%)	26.59	28.25	86.54	29.26	82.77

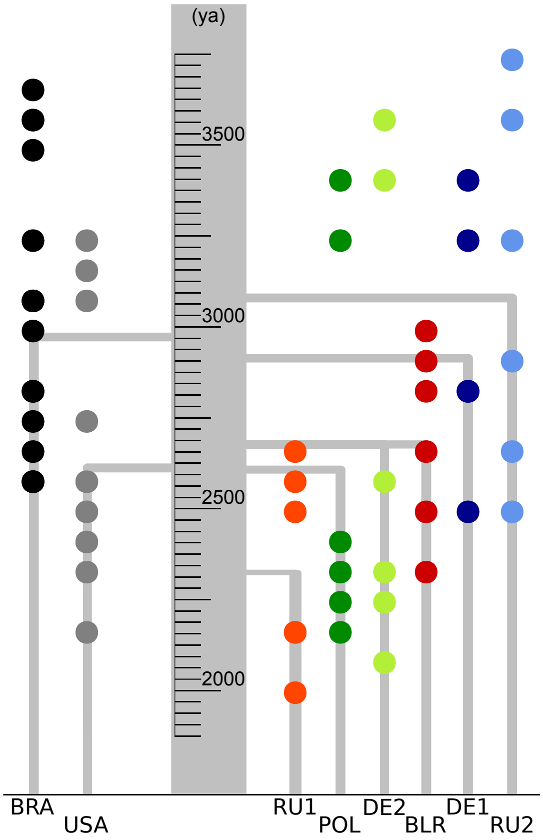


Figure 4.4: Graphical representation of the marginal mean and range of estimation of the time-to-merger from the ABC analysis. The branching represents the estimated time since the split from the weedy population (ya) for a given population average over all trios containing this population. The colored dots represent the estimations for each of these trios of populations.

4.4 Discussion

4.4.1 Several domestication events

The ABC results showed that some of the grain and forage populations may have split at different times from the weedy population. This differences can be attributed to different domestication events, possibly explaining the high genetic distances between geographically close populations. The estimated split times of the weedy populations were spread over 760 years from the Bronze Age (RU2) to the Iron Age (RU1), a period for which there is supporting archaeological evidence of rye populations in Europe. These results agree with the hypothesis of multiple domestication events proposed by archaeological studies in rye (Behre 1992; Burger *et al.* 2008; Khush 1963; Sencer and Hawkes 1980; Zohary *et al.* 2013). ABC performed using a single domestication event model with one forage and one grain ancestor splitting from the weedy population giving rise to our various rye populations, with or without migration between them (Figure 4.7), were also tested, but we could not reproduce the data (results not shown), corroborating the previous results. Different waves of breeding or domestication in northern European populations would explain the patterns of population differentiation and the STRUCTURE results (cf. $K =$ 5). For instance, despite their spatial proximity, the two Russian landraces RU1 and RU2 are differentiated from the eastern and central European populations BLR, POL and DE2 that cluster together. Ma *et al.* (2004) reported a similar distinction between Russian cultivars and those from Norway, Finland, Estonia, Ukraine and Poland. In contrast, we found similar times-of-split (360 years apart) for the two forage populations: USA and BRA from the American continent. USA is one of several of the southern North American varieties that originated from the Italian cultivar ”Abruzzi”, imported by the US Department of Agriculture in the early 1900s to be used for pasture or as a cover crop (Briggle 1959). This explains why, in our study, USA genetically clustered with the southern European population ESP. The low differentiation between USA and BRA and their high shared population membership at $K = 7$ were also expected as BRA ("Centeio Branco", which means white rye) was one of the first rye populations brought from the USA and introduced to Brazil in 1984 (De Mori *et al.* 2013). However, we cannot draw any definitive conclusions on the multiplicity of domestication events because of the variance in split time estimates from different trios and the fact that, despite their likely common origin, the difference between time-of-split estimates for USA and BRA is of the same order of magnitude as the differences among grain populations (Figure 4.4).

Some trios led to very different parameter estimations than most trios involving the same population, i.e., outliers, increasing the variance between estimations. For example, two trios of POL and DE2 had a very ancient split time compared with other estimates for these populations (Figure 4.4). In several cases, the outlier trios all contained the same TUR weedy population. We conclude that the three weedy populations are not identical and might relate in different ways to the true ancestor of the sampled cultivated rye populations. They exhibit an unknown history and possibly complex relationships with each other and to cultivated rye.

4.4.2 Bottlenecks and other complex demographic features

Our ABC results highlight the difficulty of capturing complex demographic domestication events using simplified modeling and summary statistics based on microsatellite data. As very little is known about the history of the populations studied here, we have built a simple model of a population split using the scarce information obtained from archaeological studies, our analyses of genetic diversity and preliminary ABC simulations. We have modeled strong bottlenecks associated with population splits (founding events) as well as very recent bottlenecks to mimic the effect of conservation in gene banks. However, several other bottlenecks may have occurred in the history of the populations explaining low values of the Garza–Williamson statistics (between 0.35 and 0.41; Table 3.1) and in some cases, admixture events that would explain the mixed membership coefficients obtained from STRUCTURE analyses for PRT and POL.

The bottlenecks detected mostly occurred in the last 500 generations, i.e., more recently than domestication, otherwise the populations would have recovered sufficiently and the GW statistics would be comparable to that of a population of constant size (Garza and Williamson 2001). Many cultivated species are known to have repeatedly experienced such complex demographic events in their history (Glémin and Bataillon 2009). Ultimately, the power of the ABC to study domestication (e.g., Cornille *et al.* 2012) depends on the system studied (population history, type of markers and generation time), prior knowledge and violation of the model hypotheses. This calls for caution when performing an ABC. While it is possible to use more complex models such as many populations with bottlenecks, splits and introgression/gene flow, statistical power for the estimation of parameters will be low. Typically, the model choice procedure will lead to a low rate of correct assignment and parameter estimation does not give credible posterior distributions. In this study, we chose as an alternative to derive sets of very simple models, each consisting of a limited number of populations and parameters. While it is clear that not accounting for these more complex features (i.e., repeated bottlenecks, admixture, substructure and migration) might bias our results, our model fitted the data reasonably well while giving sufficient power to compare scenarios (as shown in the moderate error rates in the confusion matrix; Table 4.4) and to infer the parameters of interest. Therefore, we concluded that the ABC can be assumed to give realistic estimates for relative split times between the grain/forage populations and the weedy populations, providing new insights into the history of these populations.

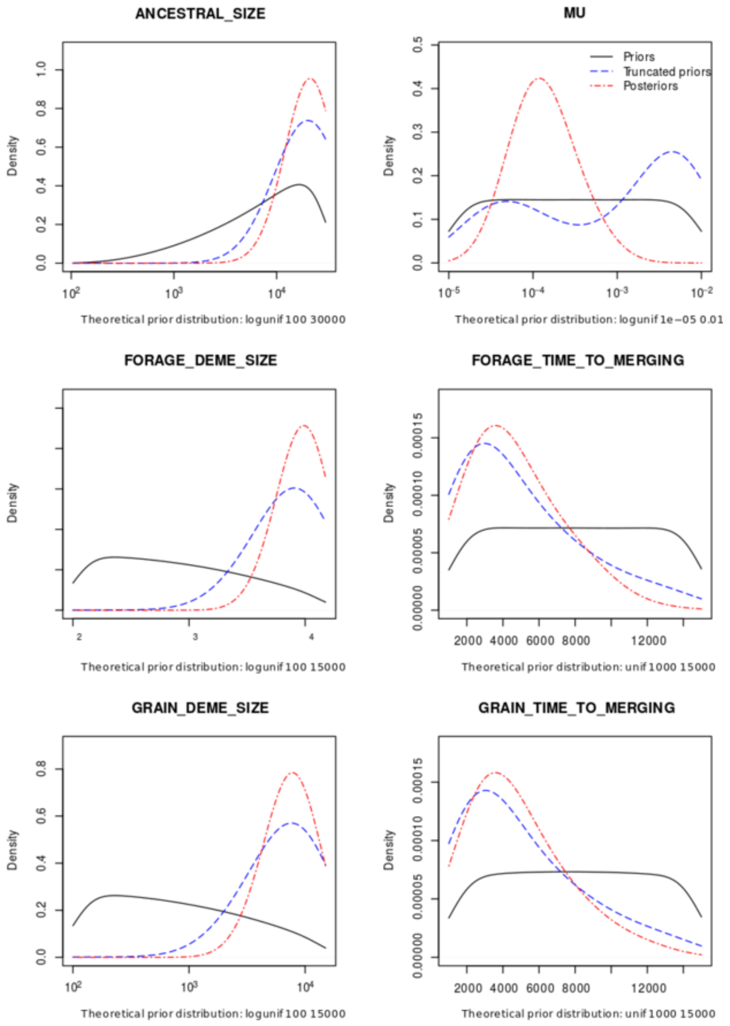


Figure 4.5: Prior, truncated prior (posterior before GLM), and posterior (after GLM) distributions for the five estimated demographic parameters and the mutation rate for one representative population trio. For all trios used for parameter estimation, the highest posterior density is at least twice the prior density and the estimated value is never on the extreme limit of the interval prior.

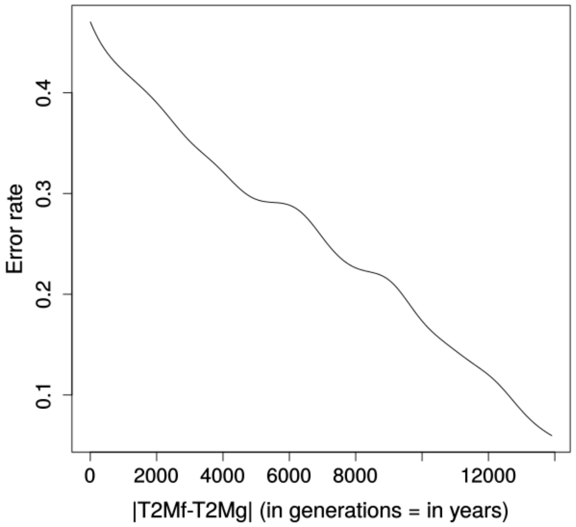


Figure 4.6: Model Choice error rate as a function of the absolute difference between the forage and the grain "times of split". The more the times of split of the forage and of the grain populations are different the higher is the power to distinguish the two models. As the two parameters T2Mf and T2Mg are i.i.d. sampled from U(1000 ; 15000), the density of their absolute difference is $f_z(z) = \frac{2}{14000} - \frac{2z}{14000^2}$. This means that the PODs are mostly sampled in the region of lesser power leading to an overestimation of the error rate.

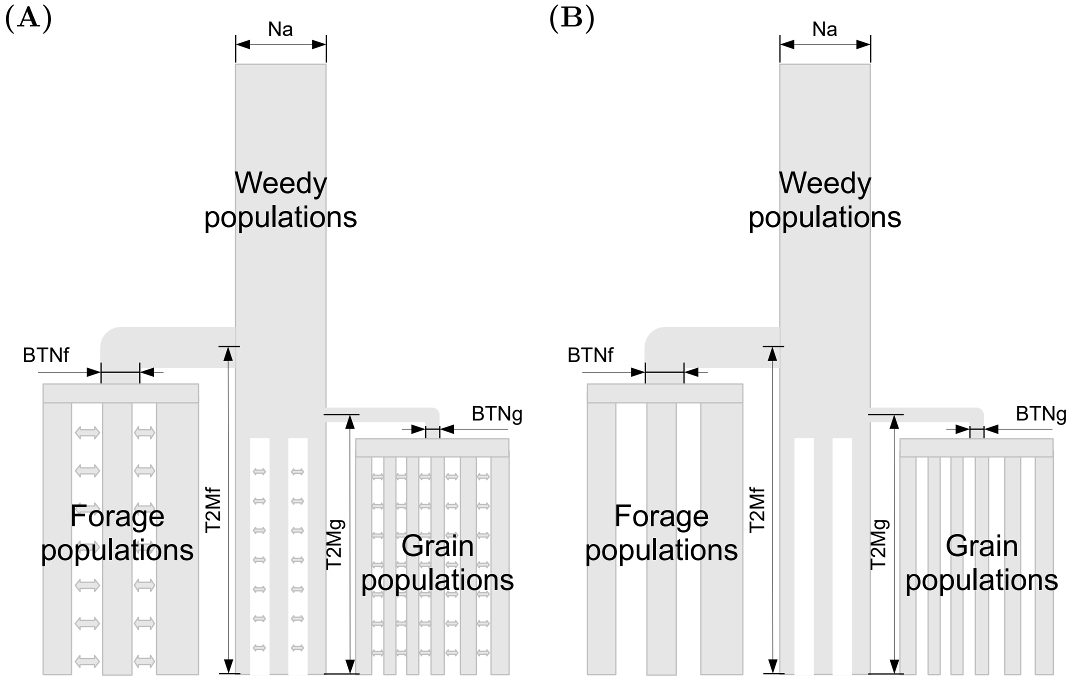


Figure 4.7: Graphical representation of demographic models with a single domestication event for all forage population and one for all grain populations with (A) or without (B) migration within groups. The simulations performed under these models and within a wide parameter space cannot at all reproduce the data and are therefore abandoned in the ABC. Furthermore, adding other features such as longer bottlenecks or one additional (recent) bottleneck per populations to these models does not improve reproducibility of the data.

Part II

Animal domestication

Chapter 5

Pedigree modelling and demographic inference

5.1 Introduction

In this chapter, we first present the model combining pedigree and genetic data and later describe how this model can be used to infer demographic and mutational parameters from that data.

5.2 The Model

To account for life history traits very common in animal populations, we extend the classic Wright-Fisher model to a diploid species with two sexes and overlapping generations. Our model, which is schematically depicted in Figure 5.1, assumes discrete generations consisting of N_f female and N_m male individuals. We further assume random mating; that is, in each generation, and going backward in time, each individual picks at random a female and a male of the previous generation as mother and father, respectively.

We model overlapping generations by allowing individuals to pick their parents not

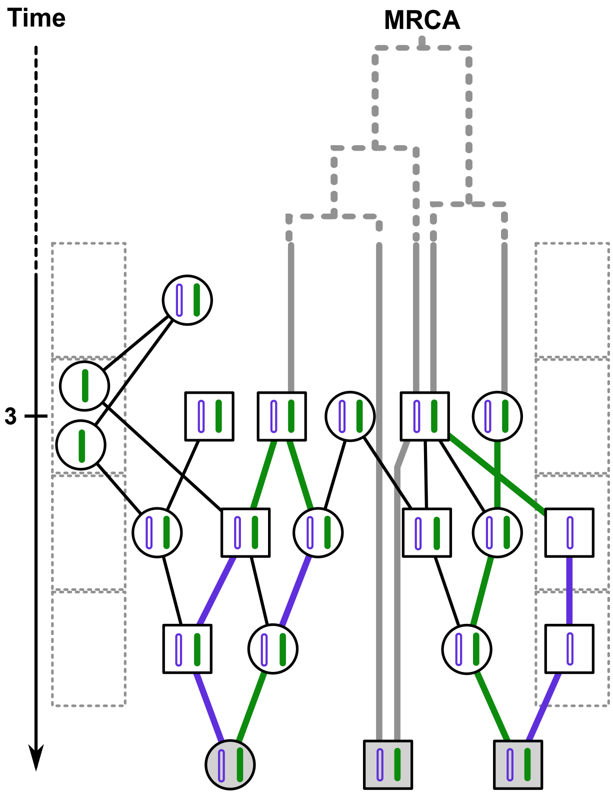


Figure 5.1: Graphical representation of the proposed two-sex model with overlapping generations. Shown are female (circles) and males (squares) individuals along with their parent-offspring relationships (black edges). The doted boxes on each side represent the female and male gamete storages populated with gametes of individuals from the pedigree. For the individuals of generation 0 (grayed) genetic data is available and the thick lines represent the genealogy of these individual. Dotted lines indicate the part of the genealogy before g_{Max} for which the continuous time approximation is used. The tick mark on the time scale represent the depth of this pedigree, $d = 3$.

from the directly preceding generation but from an earlier one with probabilities b_f and b_m for female and male parents, respectively. In this case, however, the choice of the actual distant parent is delayed and the lineage is just stored. In biological terms, these stored lineages thus represent gametes of a defined sex from previous generations, and we refer to this compartment as “gamete storage” in the following. At the beginning of a generation, the so stored gametes then pick a parent in the current generation with respective probabilities $1 - b_f$ and $1 - b_m$, and otherwise remain in storage, which implies that the number of generations between parents of a given sex and their offspring are exponentially distributed (with parameter $-\log(b_f)$ and $-\log(b_m)$ respectively).

For simplicity, we only considered here the case of constant population sizes N_f, N_m and probabilities to jump a generation b_m, b_f.

Derivation of the coalescent

Two-sex models were previously shown (Möhle 1998b) to be approximated accurately by the time-changed Kingman coalescent (1982a). Similarly, Blath *et al.* (2013) recently showed that models with overlapping generations due to seed banks, or in our case gamete storage, also result in a simple scaling of the classic coalescent if the average time lineages spend in storage is relatively small compared to the waiting time between coalescent events. Here we derive the appropriate scaling for the model introduced above.

We begin with the rate of coalescent and note that only the lineages that are in individuals of the active populations can coalesce, whereas lineages currently stored in the dormant gamete storage have first to re-enter the active populations. We describe in Figure 5.2, the four resulting compartments: the active female (A_f) and male (A_m) populations as well as the female (D_f) and male (D_m) gamete storages, and the rate at which lineages move between compartments. This system is at equilibrium when, for each compartment, the same number of lines are expected to enter and exit the compartment (i.e., Δ, the change of number of lineages in each compartment is null, $\Delta A_f = \Delta D_f = \Delta A_m = \Delta D_m = 0$). To obtain the fraction of lineages that can coalesce with each other when the system is at equilibrium, we solve the following system of difference equations:

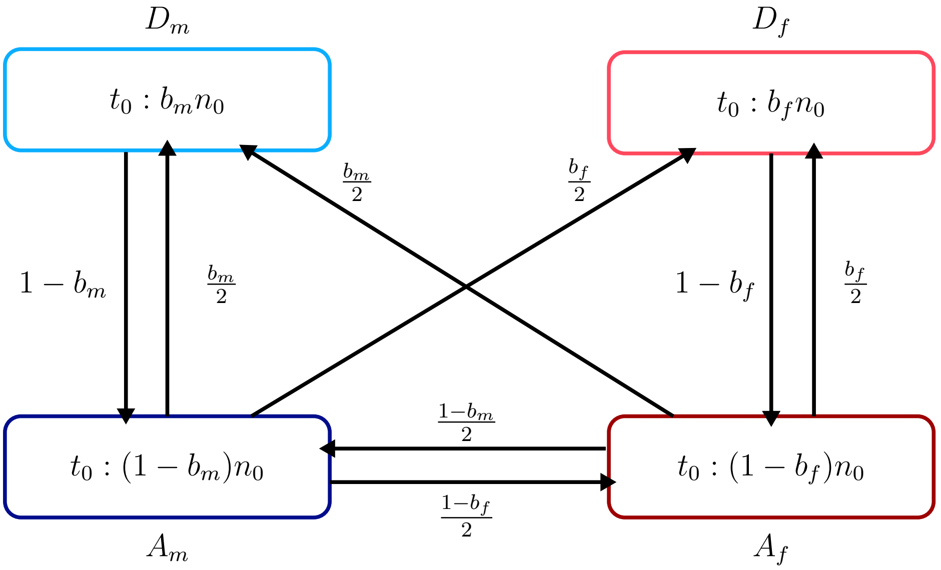


Figure 5.2: Diagram of the transition probabilities between the different sub-populations of the model: the active female (A_f) and active male (A_m) populations, and the female (D_f) and male (D_m) gamete storages. The number of lineages at t_0 is shown in each compartment, where n_0 is the initial number of individuals genetically sampled (i.e, sequenced).

$$\begin{cases} \Delta D_m = A_m \frac{b_m}{2} + A_f \frac{b_m}{2} - D_m(1 - b_m) \\ \Delta A_m = D_m(1 - b_m) + A_f \frac{1-b_m}{2} - A_m \frac{b_m}{2} - A_m \frac{b_f}{2} - A_m \frac{1-b_f}{2} \\ \Delta D_f = A_f \frac{b_f}{2} + A_m \frac{b_f}{2} - D_f(1 - b_f) \\ \Delta A_f = D_f(1 - b_f) + A_m \frac{1-b_f}{2} - A_f \frac{b_f}{2} - A_f \frac{b_m}{2} - A_f \frac{1-b_m}{2} \ . \end{cases}$$

The global rate of coalescence $\mathbb{P}(\text{Coal})$ is then given by sum of the rates per compartment weighted by the fraction of lineages residing in them. Since the coalescent rates are zero in the gamete storages and $1/2A_f$ and $1/2A_m$ in (A_f) and (A_m), respectively, we have

$$\mathbb{P}(\text{Coal}) = \frac{1}{2N_f}\left[\frac{(1-b_f)^2}{(1-b_m)+(1-b_f)}\right]^2 + \frac{1}{2N_m}\left[\frac{(1-b_m)^2}{(1-b_m)+(1-b_f)}\right]^2 .$$

If $b_f = b_m = 0$, the obtained rate reduces to

$$\mathbb{P}(\text{Coal}) = \frac{1}{2}\frac{N_m+N_f}{4N_mN_f} ,$$

as previously found for two-sex models (Möhle 1998b). If $b_f = b_m = b$ and $N_f = N_m = N/2$ the rate reduces to

$$\mathbb{P}(\text{Coal}) = \frac{(1-b)^2}{2N} ,$$

in accordance with the results of the monoecious seed bank model of Kaj *et al.* (2001).

Following Kingman's approach, the distribution of time of coalescent under our model is $T_i \sim \exp\binom{i}{2}$ with time scaled in $2N_e$ with

$$N_e = \frac{N_fN_m(2-b_m-b_f)^2}{N_f(1-b_m)^4+N_m(1-b_f)^4} . \tag{5.1}$$

We next derive the rate of novel mutations in the presence of overlapping generations. Importantly, the number of germline mutations may not scale linearly with time. Indeed, in females, most of these mutations occur during early development, and in males age effect on the germline mutation load is not necessarily proportional to the age, as has been shown for humans (Campbell and Eichler 2013). We model this effect using two mutation rates: μ per generations spend in the active populations and $\mu^* = \varepsilon\mu$ per generation spent in the gamete storage. From the compartment model introduced above, we obtain the average fraction of time t_b that linages spend in one of the gamete storages as

$$t_b = \frac{D_f + D_m}{A_f + A_m + D_f + D_m} = \frac{b_f - 2b_f b_m + b_m}{(1 - b_m) + (1 - b_f)} \ ,$$

which results in the average effective mutation rates per generation

$$\begin{aligned} \bar{\mu} &= (1 - t_b)\mu + t_b \varepsilon \mu \\ &= \mu - \frac{b_f - 2b_f b_m + b_m}{(1 - b_m) + (1 - b_f)}(1 - \varepsilon)\mu \ . \end{aligned}$$

When $\varepsilon < 1$, that is, when the mutation rate in the gamete storage is lower than the mutation rate in the active population, the average mutation rate per generation decreases with increasing values of b_f and b_m. Conversely, the time between coalescent event increases with b_f and b_m. When time is measured in terms of number of mutations occurring between two coalescent events, those two phenomena partially compensate each other. In other terms, the length of the branches measured in number of mutations increases slower with increasing b_f and b_m, than the length of the branches in generations. In the following we use the term ”mutational time” to describe coalescent time in generations corrected by the effect of ε on the mutation rate; that is, a lower average mutation rate if $\varepsilon < 1$ and a higher one if $\varepsilon > 1$.

5.3 Inference

We introduce here a Maximum Likelihood (MLE) method to infer jointly the demographic $\theta_d = \{N_f, N_m, b_f, b_m\}$ and mutational $\theta_m = \{\mu, \varepsilon\}$ parameters of the model introduced above. This estimation is based on genetic data summarized by the site frequency spectrum (SFS) and available pedigree information in terms of child-parent relationships (filiation) that form one or several connected networks spanning two or more generations ($\mathcal{P}$). The relevant likelihood function can be decomposed as

$$\begin{aligned} L(\mathcal{M}) &= \mathbb{P}(\mathrm{SFS}|\mathcal{P},\theta_d,\theta_m)\mathbb{P}(\mathcal{P}|\theta_d) \\ &= \sum_G \left[\mathbb{P}(\mathrm{SFS}|G,\mu)\mathbb{P}(G|\mathcal{P},\theta_d,\varepsilon)\right]\mathbb{P}(\mathcal{P}|\theta_d)\ , \end{aligned} \tag{5.2}$$

where the sum runs over the unknown genealogies G representing the genetic relationships between all sampled individuals up to the the most recent common ancestor (MRCA). While the pedigree and the genealogies share similar features, they should not be confused.

In the following sections, we first derive each term of the likelihood function individually, and then give a detailed description of an inference framework under this model.

5.3.1 The Pedigree

Let $\mathcal{P}_g$ be the way in which the individuals of generation $g-1$ in the pedigree are assigned to their parents in generation g. Note that generations as well as the choice of the mother and the father are considered as independent events thus their likelihood can be multiplied and we obtain

$$\mathbb{P}(\mathcal{P}|\theta_d)=\prod_{g\geq 1}\mathbb{P}(\mathcal{P}_g|\theta_d)=\prod_{g\geq 1}\mathbb{P}(\mathcal{P}_{f,g}|\theta_d)\mathbb{P}(\mathcal{P}_{m_{,g}}|\theta_d)\ , \tag{5.3}$$

where $\mathcal{P}_{f,g}$ and $\mathcal{P}_{m_{,g}}$ represent the assignment of individuals to their mothers and fathers, respectively.

The pedigree spans between the generation of the most recent individual $g = 0$ and the generation of the last known parent that we call g_{Max}. To derive the probability of the pedigree, we consider all individuals at generation 0 in the pedigree as numbered (i.e., identifiable). These individuals then choose their parents from the previous generation, but are constrained in their choices by the pedigree. A parent chosen by an identifiable individual becomes automatically identifiable itself as it is the unique mother, father respectively, of a unique, identifiable individual. The chosen, now identifiable, parents choose in turn their own parents from the previous generations according to the pedigree information and this process continues until the top of the pedigree is reached.

Here we derive $\mathbb{P}(\mathcal{P}_{f,g}|\theta_d)$ for this process for the individuals of generation $g-1$, of which exactly $B_{f,g-1}$ will enter the gamete storage as their mother is from a distant generation, and the remaining individuals, $\bar{B}_{f,g-1}$, will choose a mother from generation g. Within these $\bar{B}_{f,g-1}$ individuals, the first individual of each of the M_g groups of siblings chooses a distinct mother from the population, which they do in turn with probabilities $1, \frac{N_f-1}{N_f}, \dots \frac{N_f-M_g}{N_f}$. The M_g so chosen mothers, which have become identifiable themselves, are chosen by their remaining offspring with probability $\frac{1}{N_f}$ each. The resulting probability of this process is

$$\mathbb{P}(\mathcal{P}_{f,g}|\theta_d) = \alpha_{fg}\left(\frac{1}{N_f}\right)^{\bar{B}_{f,g-1}-M_g} {b_f}^{B_{f,g-1}}(1-b_f)^{\bar{B}_{f,g-1}} \ , \tag{5.4}$$

where we used the notation

$$\alpha_{fg} = \frac{N_f!}{N_f^{M_g}(N_f-M_g)!} \ .$$

The same holds true analogously for $\mathbb{P}(\mathcal{P}_{m,g}|\theta_d)$ by replacing the subscript f by m and using F_g, the number of fathers in generation g, instead of M_g.

A maximum likelihood estimate of N_f, N_m, b_f and b_m is easily obtained by taking the first derivative of the logarithm of eq. 5.3. For b_f, this yields

$$\frac{\mathrm{d}}{\mathrm{d}b_f}\log\mathbb{P}(\mathcal{P}|\theta_d) = \frac{\sum_{g=1}^{g_{\mathrm{Max}}} B_{f,g-1}}{b_f} - \frac{\sum_{g=1}^{g_{\mathrm{Max}}} \bar{B}_{f,g-1}}{(1-b_f)} \ ,$$

which admits the maximum likelihood estimate

$$\hat{b_f} = \frac{\sum_{g=1}^{g_{\mathrm{Max}}} B_{f,g-1}}{\sum_{g=1}^{g_{\mathrm{Max}}} \bar{B}_{f,g-1} + B_{f,g-1}} \ ,$$

and analogously for b_m. For N_f, the first derivative is

$$\frac{\mathrm{d}}{\mathrm{d}N_f} \log \mathbb{P}(\mathcal{P}|\theta_d) = \sum_{g=1}^{g_{\mathrm{Max}}} \mathcal{F}(N_f) - \mathcal{F}(N_f - M_g) - \frac{\bar{B}_{f,g-1}}{N_f} \ , \tag{5.5}$$

where $\mathcal{F}$ is the digamma function defined as the logarithmic derivative of the factorial function. This probability function is defined for $N_f > M_g$. Its maximum, if it exists, can be found numerically.

5.3.2 Genetic data

Coalescence is the merging of two or more genetic lineages. In a diploid population, an offspring may inherit one of two possible chromosomes of each parent. There are thus 2^l ways in which l offspring lineages can be assigned to the two chromosomes of a single parent (Figure 5.1). Enumerating all possible genealogies constrained by even a small but fully resolved pedigree, as done for two lineages in Wakeley *et al.* (2012), is computationally already very challenging for large sample sizes, and easily becomes prohibitive if the pedigree is only partially known. We thus chose to turn to simulations to evaluate the sum in eq. 5.2, as is commonly done in the absence of pedigree information (e.g., Excoffier *et al.* 2013; Nielsen 2000; Nelson *et al.* 2012):

$$\sum_G \left[\mathbb{P}(\mathrm{SFS}|G,\mu)\mathbb{P}(G|\mathcal{P},\theta_d,\varepsilon)\right] \approx \frac{1}{N_\gamma} \sum_\gamma \mathbb{P}(\mathrm{SFS}|G=\gamma,\mu) \ ,$$

where the genealogies $\gamma \sim \mathbb{P}(G|\mathcal{P},\theta_d,\theta_m)$ are simulated under model parameters $\mathcal{M}$ and constrained by the pedigree $\mathcal{P}$.

Simulating genealogies inside a pedigree is straight forward and only requires binary choices when following lineages backward in time through the pedigree. Simulations within the pedigree also allows for more complex sampling schemes such as sampling (sequencing) individuals over several generations, even within families (e.g., father-son sequencing). In a pedigree without missing parents, the topology of genealogies is constraint enough to permit the efficient simulation of many genealogies at once by propagating through the pedigree the number of genealogies that makes a specific binary gamete choice. In case of only partial pedigree information, lineages reaching parents of which only one or none of the parents are known choose their unknown parents randomly from the whole population, or enter the gamete storage (Figure 5.1). Lineages that exit the pedigree due to missing parents can also reenter the pedigree in a later generation by chance. Due to the extremely large number of possible topologies, in this case, the process needs to be simulated per generation (discrete process) and for each genealogy separately. Simulations are therefore time consuming in case of limited pedigree information. However, at a certain generation in the past that we term g_{Max}, the pedigree does not contain any information about ancestors anymore and the genealogy is then only constraint by the parameters of the model. The expected time to the MRCA of a large sample of diploid individuals is approximately $4 \times N_e$ generations where N_e is usually larger than 200. For most available pedigrees, g_{Max} is thus reached long before the MRCA. We therefore make use of the appropriately scaled coalescent approximation introduced above to simulate the genealogies from g_{Max} backwards to the MRCA (Figure 5.1).

To calculate $\mathbb{P}(\text{SFS}|G = \gamma, \mu)$, the probability of the genetic data summarized by the SFS given a genealogy γ, we use the classic infinite site mutation model with Poisson distributed mutations at rate μ per site. Under this model, and assuming that sites are independent, the probability that a mutation results in a derived sample allele frequency of i is given by the summed length L_i of all branches with i leaves and the probability of the SFS is thus given by a multinomial distribution

$$\mathbb{P}(\text{SFS}|G = \gamma, \mathcal{M}) = e^{-\mu L} \frac{\mu^S L_1^{S_1} \ldots L_{n-1}^{S_{n-1}}}{S_1! \ldots S_{n-1}!} \ , \tag{5.6}$$

where S_i is the number of segregating sites being shared by i chromosomes in the sample of size n and L the total length of the genealogy γ (Fu 1998). We note that the branch lengths L_i is measured in mutational time, that is a generation spent in the gamete storage only adds ε to the branch length.

The maximum likelihood estimate of μ can be obtained analytically by differentiating the logarithm of eq. 5.6, which yields the estimator

$$\hat{\mu} = \frac{S}{L} \ ,$$

where L is the total length of the genealogy in mutational time. In the absence of pedigree information, for instance for the part of the genealogy simulated under the coalescent approximation, the total length of the genealogy is only available measured in generations. In this case, the time spent in gamete storage is averaged over the tree branches and the ML estimate becomes

$$\hat{\mu} = \frac{S}{4\hat{N}_e(1 - t_b + t_b\varepsilon)L_c} \ , \tag{5.7}$$

where L_c is the total length of the genealogy in coalescent time (i.e., in θ generations) t_b is the average time spend in the gamete storage, equal to $\hat{b}$ in our model, and $\hat{b}$ and $\hat{N}_e$ are the ML estimates of b and N_e, respectively.

5.3.3 Inference algorithm

An exact analytical or numerical solution for the joint maximum likelihood of all parameters is not available. We therefore combine some of our analytical derivations with numerical evaluations using MCMC in the following inference algorithm:

Algorithm 1

1. *We sample vectors of demographic parameters* $\theta_d^{(i)} \sim \mathbb{P}(\theta_d|\mathcal{P}), i = 1, \ldots, I$ *from their joint posterior distribution given the pedigree using an MCMC framework.*

2. *For each sampled vector of parameters* $\theta_d^{(i)}$*, we simulate* $G = 100$ *genealogies constraint by the pedigree* $\mathcal{P}$*.*

3. *For each $\theta_d^{(i)}$ we then compute the MLE estimate $\hat{\mu}^{(i)}$ according to eq. 5.7 using the sampled genealogies.*

4. *Finally, we compute the joint likelihood of all model parameters for each pair of $\theta_d^{(i)}$ and $\mu(i)$ according to eq. 5.2, again using the simulated genealogies.*

Our inference scheme is thus closely related to a grid search on the model parameters where we make use of the pedigree information to conduct the simulation-based likelihood evaluation only at promising locations of the parameter space. The proposed combination of MCMC sampling and MLE is possible because the population size is constant and the maximum likelihood of $\bar{\mu}$ does only depend on the total length of the genealogies and not on their topology. As shown in Chapter 6, this method is an efficient compromise between speed and accuracy. Further advantages and limitations are discussed below, in section 5.4.

The MCMC sampling in step 1 is implemented using a standard Metropolis algorithm (Metropolis *et al.* 1953) in which a single parameter is updated per iteration using a Gaussian proposal kernel mirrored at prior limits. We use uniform priors on all parameters except the population sizes for which we use log-uniform priors and propose updates on the logarithmic scale during the MCMC to account for their prior easily spanning several orders of magnitude. The implementation of the method in C++ is available upon request.

5.4 Discussion

In this chapter, we developed a model explicitly accounting for two sexes and overlapping generations. Under this model, genealogies follow a standard coalescent provided that time is rescaled appropriately and that the expectation of the age distribution (i.e., the number of generations between parent and offspring) is finite and in particular small compared to the effective population size. This is generally true in our model, under realistic parameter values. This new model allowed us to infer parameters jointly from genetic data and pedigree information.

5.4.1 Random sampling

Random sampling of the most recent individuals of the pedigree (i.e., individuals without known offspring) is an implicit assumption in the construction of this model and related inference method. For example, individuals sampled based on the amount of pedigree information recorded for their ancestry, are more likely to be related within that pedigree than individuals chosen at random. In this case, the number of siblings in the pedigree increases leading to an underestimation of the population size. Moreover, random sampling of the current generation does not guarantee that any missing information occurs at random in the pedigree, potentially leading to ascertainment bias.

This phenomenon is not specific to the pedigree model and can occur in classical population genetics inference without a pedigree. If individuals within a population are more likely to be sequenced if they are unrelated, for example with the aim of maximizing the diversity of the sequenced panel, an ascertainment bias is introduced and the population size might be overestimated. However, the effect might be stronger when using the pedigree in inference. Indeed, Wakeley *et al.* (2012) showed that the coalescent process is robust to demographic events that strongly affect the pedigree.

5.4.2 Multi-generational sampling

In our inference method, the genealogies of the genetic samples are drawn within the known pedigree. Therefore, sequences of individuals that are from different generations within the time of the pedigree can be used to calculate the SFS, whether these individuals are known to be related or not. If the relationship between the sampled individuals is recorded in the pedigree, the true probability of coalescence between their lineages will be more accurately represented in the simulations. Sequencing data from sire-son pairs, a common practice in cattle sequencing projects (e.g., Hayes *et al.* 2012), can therefore readily be used within our framework.

In conclusion, we presented here a new model and some theoretical results on how to combine pedigree and genetic information for the inference of demographic and mutational process and showed that these processes can be disentangled if sufficient pedigree information is available. This is widely unexplored territory as most methods use individual or genealogy based models. But the availability of both pedigree and genetic data for many species, in particular domesticated animals, motivates the development of methods that combines such data. While an application to real data may pose additional challenges, our work is a first step towards such a method and extensions of our approach to more complex demographies and other features of real populations are readily possible. If done properly, the application of these to real data has the potential to give us deep insight into the mutational process in natural populations.

Chapter 6

Application to simulated and real data

6.1 Introduction

The model and inference framework developed in the previous chapter allows the estimation of male and female population sizes, rate of overlapping generations and mutation rate but the quality of the estimation depends on the amount of information contained in the pedigree and the genetic data. We therefore perform simulations, to explore some of the possible variables that might impact the accuracy of the estimation such as, the number of the generations over which the pedigree is known, the number of individuals and the current generation for which ancestry has been recorded, or the true population sizes. This chapter describes the inference performed over these simulations as well as inference performed on a real pedigree from a cattle population.

6.2 Simulations

To test the performances of our inference method, we used a custom R script to simulate pseudo observed datasets (PODs) consisting of a pedigree and a corresponding SFS for a sample of 50 individuals, unless specified differently. The pedigree includes all ancestors of the sampled individuals until the predefined depth d as well as the parents of all lineages in gamete storage at generation d (Figure 5.1). Thus, the generation of the oldest individual contained in the pedigree g_{Max} is such that $g_{\text{Max}} \geq d$. We set $b_f = b_m = b$ for all simulations and generated an SFS by simulating 2000 loci of 10 kb each with $\mu = 5 \times 10^{-9}$ and $\varepsilon = 0$.

For each simulation presented here, the MCMC in step 1 of Algorithm 1 was run for 4.5×10^6 steps thinned out to keep only every 500th parameter combination, of which the first 200 were discarded as a burn-in (resulting in 8800 sampled parameter vectors θ_d). We use a normal distribution for the kernel of all three estimated demographic parameters (N_f,N_m and b). The MCMC parameters relative to each of these demographic parameters can be found in Table 6.1.

We first used simulations to assess the benefit of having pedigree data as a function of the pedigree depth across 10 independently generated PODs with demographic parameters realistic for domesticated breeds ($N_f = 5000, N_m = 500$ and $b = 0.2$). As shown in Figure 6.1, our method is capable of accurately disentangling the effects of the mutation rate and population sizes on genetic diversity already if limited pedigree information is available. Indeed, reliable estimates are obtained for all parameters including sex-specific population sizes, the frequency of overlapping generations as well as the mutation rate if a pedigree of depth four or more is used (Figure 6.1).

Interestingly, the rate of overlapping generations b is estimated well across the whole parameter range in the presence of sufficient pedigree data (Figure 6.2), but smaller population size seems to be consistently estimated more accurately than larger sizes. This is visible as a reduced accuracy in the inference of the female compared to male size in Figure 6.1, but also occurs if the population sizes of both sexes are equal (Figure 6.2

Table 6.1: MCMC parameters used to estimate parameter on simulated data.

Parameter	Value
Burnin	10000
Total number of steps (incl. burnin)	4500000
Number of genealogies per steps	100
Parameters relating to b:	
Initial value	0.5
Minimum value	0.001
Maximum value	0.8
Variance of the normal jump kernel	0.02
Parameters relating to both N_f and N_m:	
Initial value	10000
Minimum value	25
Maximum value	50000
Variance of the normal jump kernel	0.5

and 6.3). We explain this as follows. The information about the population size of a pedigree is mostly contained in individuals sharing parents (i.e., half or full siblings). If the population is large but the number of individuals in the pedigree relatively small, few to no siblings are observed and the power to estimate the population size decreases. Indeed, when there are no siblings in the pedigree, the likelihood of the population size increases monotonously but reaches a kind of plateau before the true value is reached (eq. 5.5). This leads to an overestimation of the population size in the absence of genetic information and the inability to disentangle N_e from μ if such data is available. As an example, consider the posterior distributions shown in Figure 6.4 for the case of $N_f = 5,000$ and a pedigree depth of one.

We next quantified the effect of pedigree depth and width (number of individuals at $g = 0$) on the accuracy of inferring population sizes. Maybe not surprisingly we found that much more information is contained in small but deep compared to large but shallow pedigrees (Figure 6.3). Indeed, increasing the width beyond just a handful of individuals seems to hardly increase estimation accuracy except for very small population sizes, probably due to the oversampling effect described by Wakeley and Takahashi (2003).

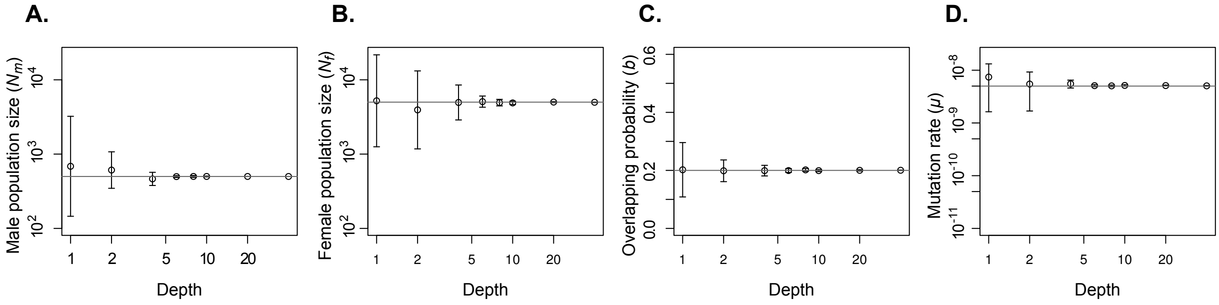


Figure 6.1: Parameter inference as a function of pedigree depth (d). Shown are the mean and standard deviation over 10 simulated datasets. The true parameter values used for all simulations (A) $N_m = 500$, (B) $N_f = 5000$, (C) $b = 0.2$, and (D) $\mu = 5 \times 10^{-9}$ are indicated by gray horizontal lines.

The reason for this is that the number of individuals included in a completely resolved pedigree is growing rapidly with each generation going further back into the past (Derrida *et al.* 2000, Figure 6.5), and so are the number of observed parent-offspring relationships informative about population size. Indeed, around 80% of the whole population is included in a complete pedigree of width 50 individuals at only few generations in the past, depending on the population size. At a depth of four, which we found to result in good estimates, about 7.5% or 750 individuals are part of the complete pedigree of 50 individuals from a population of 10,000 individuals (Figure 6.5).

6.3 Application to a real cattle pedigree

Bovine races are a good example of domesticated species with sex-biased populations and sex-biased overlapping generation rate for which pedigree records are kept. By fitting the presented model to the pedigree of a Fleckvieh population (Jansen *et al.* 2013), graciously provided by Prof. Dr. Ruedi Fries, we infer the female and the male population sizes as well as overlapping generation parameters.

The Fleckvieh pedigree is constituted of 4822 individuals from which 88 are the root individuals with no recorded descent, 3094 are dams and 1640 are sires. Generations have

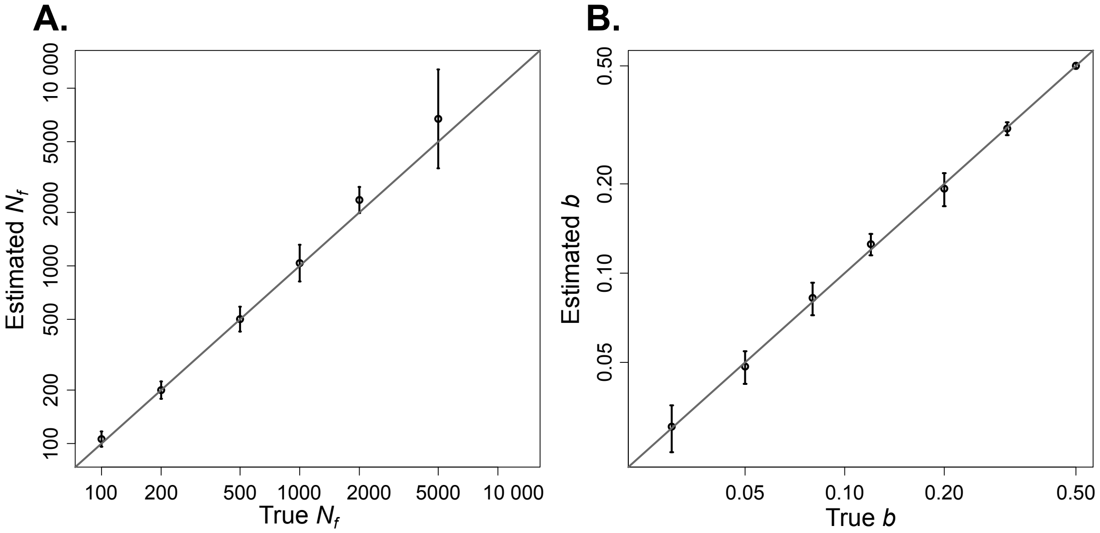


Figure 6.2: Estimated (A) $\mathbf{N_e}$ in function of simulated $\mathbf{N_e}$ and (B) b in function of simulated b, in log-log scale. The vertical bares represent the standard deviation over 10 datasets. The gray diagonal marks the identity line.

been fixed based on the date of birth when available, assuming 4 year generations on average. When the date of birth was unknown, the generation number was given using relation information with other recorded individuals. In total, the pedigree contains records over 22 generations.

In the model, the individuals are considered as uniquely identified (not exchangeable) and the choice of parents independent across generations. The pedigree can therefore be summarized in a table recording the number of individuals in the six different compartments: female and male parents (M for mother and F for father), female and male gametes choosing a parent in the active population ($\bar{B}_f$ and $\bar{B}_m$ respectively), female and male in gamete storage (B_f and B_m, respectively), at each generation (Table 6.2).

The geometric model chosen to model overlapping generations seems to fit the observed pedigree (Figure 6.6). The maximum likelihood estimates (MLE) for the probability to move to, or stay, in gamete storage $\hat{b}$ is 0.0324 for maternal gametes and 0.438 for paternal gametes. Assuming a constant population size during the generations recorded in the

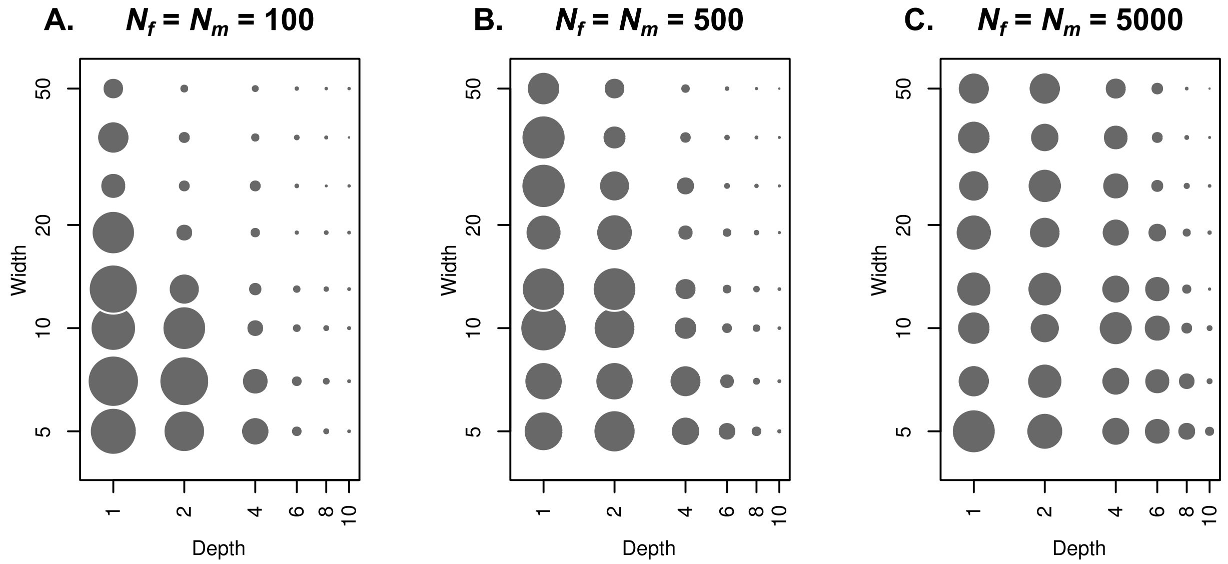


Figure 6.3: Power to infer population sizes as a function of pedigree width and depth. The surface of each dot represents the root mean squared errors (RMSE) over 10 simulations with population sizes (A) $N_f = N_m = 100$, (B) $N_f = N_m = 500$ and (C) $N_f = N_m = 5,000$. The RMSE is comprised between 1.681×10^{-3} for depth = 40 and width = 36 in (C) and 3.716 for depth = 1 and width = 7 in (A).

pedigree, the MLE of the female and male population sizes are 7151 and 252 respectively. The log-likelihood contour plot in Figure 6.7 gives additional information about the shape of the likelihood function around the MLE values, showing a greater uncertainty (flatter likelihood surface) around the estimation of the female population size than around the male one, probably due to the larger population size of females and lower connectivity in the pedigree (i.e., less siblings from the same mother).

The effective population size (Ne) of the sampled Fleckvieh population is 4513 when calculated from eq. 5.1 with the MLE. Using the classical two sex formula from Wright (1931), as in Eq. 2.1 ignoring the overlapping generation coefficients, the effective population would only be 974.

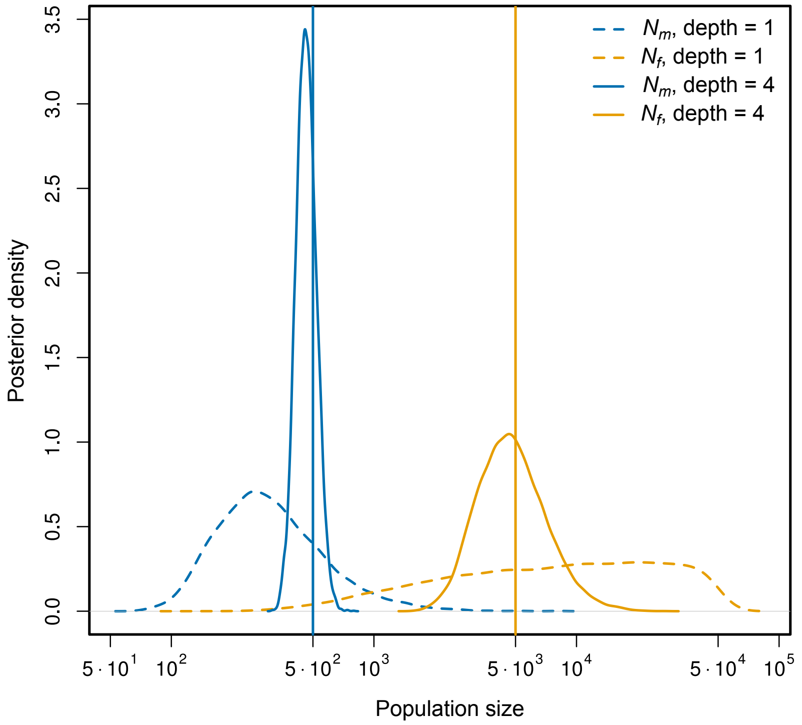


Figure 6.4: Examples of posterior distributions of population sizes based solely on pedigree information. Shown are posterior distributions of the female (yellow) and male (blue) population sizes of simulations conducted with 5000 female and 500 male individuals (vertical solid lines) calculated using MCMC. Distributions obtained from pedigrees of depth four are plotted as solid lines, those obtained from pedigrees of depth 1 as dotted lines.

6.4 Discussion

In this chapter, we show using simulations that including pedigree information not only improves the estimates of demographic parameters, but also allows to disentangle the effects of demographic and mutational processes on genetic diversity and hence to estimate these processes jointly.

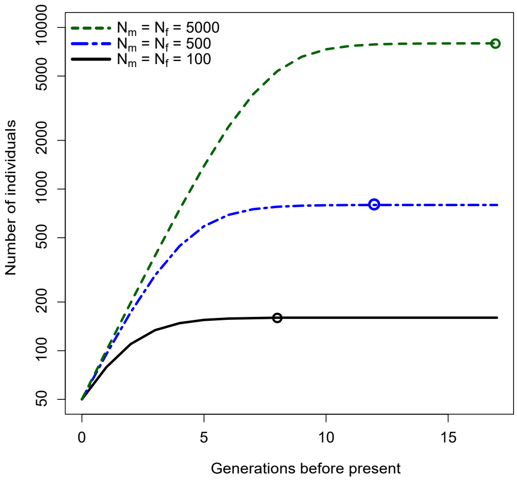


Figure 6.5: Average number of individuals entering the pedigree per generation starting with 50 individuals at generation 0. In each generation, each individual picks a female and male individual at random as parents, leading to an initial increase in the number of individuals in a pedigree. The probability to have $M_{g+1} = i$ distinct mothers in generation $g + 1$ given M_g distinct mothers in generation g is given by

$$\mathbb{P}(M_{g+1} = i|M_g) = \frac{1}{N^{M_g}}\binom{N}{i}\sum_{j=0}^{i-1}(-1)^j\binom{i}{j}(i-j)^{M_g},$$

and hence

$$\mathbb{E}[M_{g+1}|M_g] = N - \frac{(N-1)^{M_g}}{(N)^{M_g-1}}$$

and analogously for the expected number of fathers $\mathbb{E}[F_{g+1}|F_g]$. The total number of individuals in the pedigree at generation g is thus $\mathbb{E}[\Phi_{g+1}|\Phi_g] = \mathbb{E}[M_{g+1}|M_g] + \mathbb{E}[F_{g+1}|F_g]$.

When the curve reaches a plateau, marked by the dots, approximately 80 % of the population is sampled.

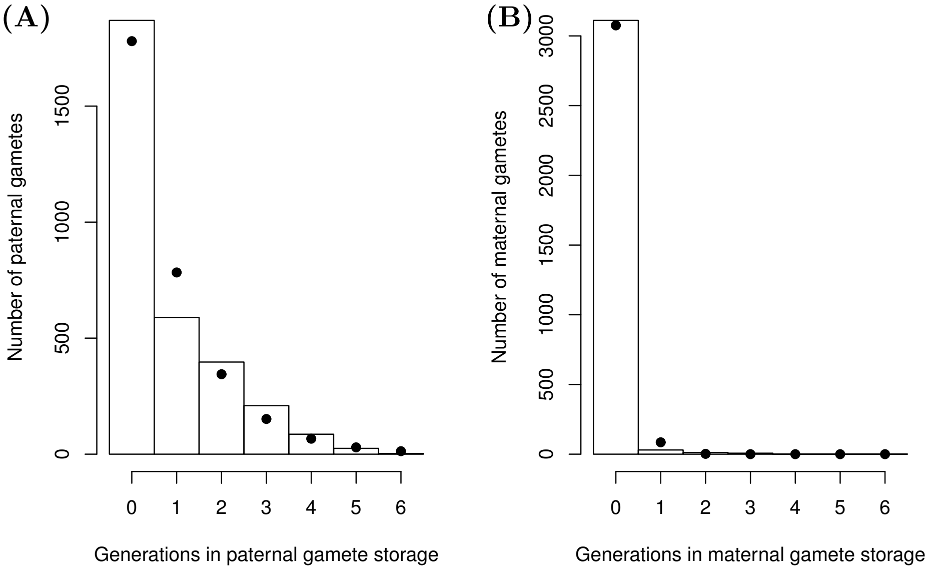


Figure 6.6: Empirical (histogram) and fitted geometric (dots) distribution of the time gametes spend in storage when choosing (A) a father or (B) a mother.

6.4.1 Amounts of data needed and implications for real applications

Our simulations show that the pedigree information of 50 individuals tracing back four generations is sufficient to obtain accurate joint estimates of the female and male effective population sizes, the proportion of overlapping generations and the mutation rate. Importantly, obtaining this amount of pedigree information is realistic for many populations of interest. For example, such pedigrees are available for several human populations (e.g., Hussin *et al.* 2015), for many domesticated animals breed of cattle and horses (e.g., Cunningham *et al.* 2001; Mc Parland *et al.* 2007) and for some wild animals (e.g., Clutton-Brock *et al.* 1982; Ellegren 1999).

Unfortunately, we could not find a dataset with both sequencing and pedigree data for a population. For many populations with a known pedigree, very low coverage sequence data filled using imputation or SNP array data are available but SFS build from such

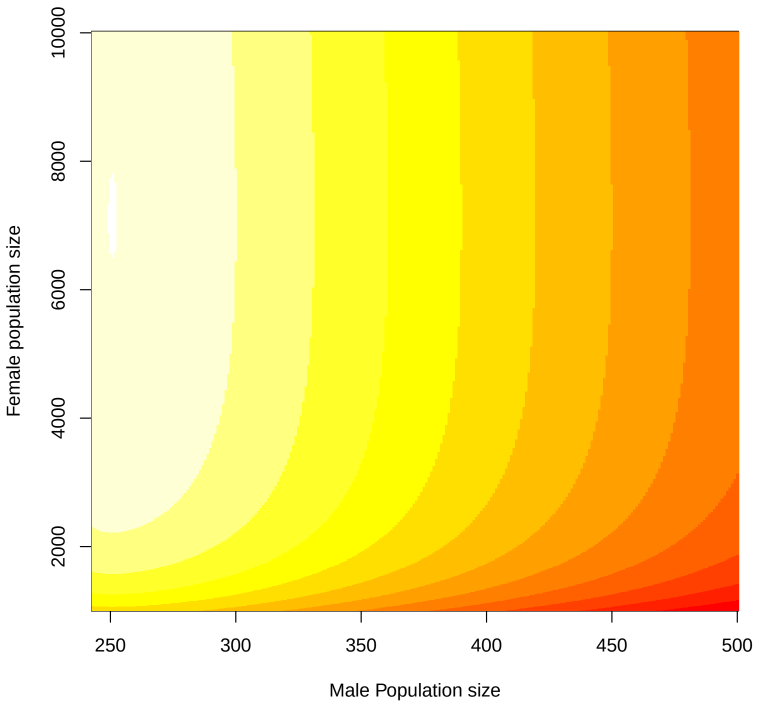


Figure 6.7: Log-likelihood surface for a range of female and male population sizes of the Fleckvieh population. The box probabilities are fixed to their MLE values $\hat{b}_f = 0.0324$ and $\hat{b}_m = 0.438$. The heat map indicates the highest log-likelihood values with white and lower values with darker red.

data are often of poor quality due to ascertainment bias. We expect the estimation of the mutation rate to be particularly sensitive to such bias.

However, we note that the amount of pedigree information required for accurate inference does depend on the population size with more data being required for larger populations. This stems from the fact that most of the information about the population size contained in a pedigree depends on the number of individuals sharing common ancestors. As a random sample is expected to contain less such individuals in a large population than in a smaller one, it will contain less information. Since the number of common ancestors increases more rapidly with the depth than the width of a pedigree, deep pedigrees of a few individuals contain much more information than shallow pedigrees of many individuals. As we discussed, the number of distinct ancestors in previous generations rapidly decreases with depth and reaches about 80% of the population within only few generations (Derrida *et al.* 2000, Figure 6.5). However, these results consider a complete pedigree and are expected to be mitigated in presence of missing information.

Having only little pedigree data available will make it difficult to disentangle the effect of mutation and drift. A particular characteristic of such a situation is that the posterior distribution of the population sizes given the pedigree data alone will be very flat and often extend to very large population sizes. In such cases, the samples generated with our MCMC will likely not be distributed densely enough around the joint MLE to warrant accurate inference. In the extreme case of no pedigree information, the joint likelihood surface of the mutation rate and population sizes will form a ridge and the estimate produced by our stochastic inference method will single out a random combination not necessarily reflective of the true parameters. However, as we have shown, already limited pedigree information of a few individuals over a few generations is sufficient to result in accurate inference.

6.4.2 Limitations of the model and possible extensions

While our theoretical and simulation results are very promising, we note that its application to real data may present some challenges. Firstly, the concept of generation, while convenient, is an artificial construct to discretize time that has little biological meaning for many long lived species. As a consequence, attributing the individuals of a pedigree to specific generations can be difficult. However, it is possible to extend our inference framework to also integrate over the attribution of individuals to generations as

$$L(\mathcal{M})=\int\sum_{G}[\mathbb{P}(\mathrm{SFS}|G,\mathcal{M})\mathbb{P}(G|\mathcal{P},\mathcal{M})]\mathbb{P}(\mathcal{P}|\mathcal{M})\mathbb{P}(\mathcal{P}^*|\mathcal{P})d\mathcal{P},$$

where we denote by $\mathcal{P}^*$ the pedigree data without generation information (hence only relationships). Here, $\Pr(\mathcal{P}^*|\mathcal{P}) = 1$ if the pedigree $\mathcal{P}$ is compatible with $\mathcal{P}^*$, that is, if all parents are from an older generation than all of their offspring and the most recently born individual is from generation 0, and $\Pr(\mathcal{P}^*|\mathcal{P}) = 0$ otherwise. Unfortunately, none of the parameters' MLE is trivial to derive because finding the maximum of this likelihood function implies finding the optimal set of pedigrees $\mathcal{P}$. However, an MCMC method sampling such pedigrees can be envisioned to infer parameters under such an extended model.

Secondly, our model assumes a geometric age distribution that seems to apply relatively well to the Fleckvieh population studied here but might not be realistic for other species. Our approach can be extended to other age distributions as long as they converge toward some form of rescaled coalescent. Example of such seed or gamete storage models have been given by Kaj *et al.* (2001) and Blath *et al.* (2013). Our results also show that cattle males are overlapping generations more often than females. This is expected knowing breeding practices and from biology, indeed the reproduction period of males is longer than the one of females. However this effect might be much stronger in the most recent generations due to the use of artificial insemination. This could bias our model and change the scaling of genealogy branches over time in an unexpected way.

Thirdly, demographic events such as population size changes, migration between populations or complex mating systems, e.g., monogamy or harem models, may be needed to describe real populations. The introduction of such demographic events in the discrete generation pedigree model is fairly easy. For example, population size changes can be directly implemented in eq. 5.4 by using generation specific values of N_f and N_m. The way demographic events shape coalescent processes is well described for many cases and they apply to our model if appropriately scaled.

Complex mating systems or reproductive skew are well described for generation by generation models (Gasbarra *et al.* 2005) however in a continuous setting, some nonstandard coalescent models are known to arise in some cases (Eldon and Wakeley 2006). In the pedigree part of the model, we assume that individuals choose their parents at random and that each choice is independent. However, we consider that parents become identified when chosen by an offspring and therefore, only the relation between the total number of parents chosen and the number of offspring choosing among them matters. For example, having one parent with 3 offspring and one with 5 leads to the same outcome in terms of the population size probability than having two parents with 4 offspring each. It might however not lead to the same genealogy shapes, as these are constraint by the pedigree. Depending on the type of true offspring distribution observed and how it deviates from the expected Poisson distribution, discrepancies could appear between the effective population size defined for the pedigree and for genealogies. These discrepancies might also be different if they arise from randomly over-dispersed offspring distributions or from selection. Further investigation is needed to find such offspring distributions and evaluate the robustness of the model to these deviations.

In less extreme cases, specific mating systems can be well approximated by strongly skewed sex ratio (Nunney 1993) which our model already incorporates in its current form. For example, in the Fleckvieh population we estimate the population size to be much larger for females than for males (Figure 6.7), as expected in cattle. Similarly, selection at the phenotypic level might result in a few individuals being more successful. This leads to a reduction of the effective population size that is accounted for by our method. However, at the genetic level the sequenced loci are considered neutral.

Table 6.2: Tabular representation of the Fleckvieh pedigree.

g	M	F	B_f	B_m	B_f	B_m
0	0	0	0	1	2	1
1	2	1	2	4	12	11
2	12	7	0	32	62	32
3	59	23	7	50	94	83
4	91	46	7	115	145	79
5	140	46	8	129	183	170
6	179	66	6	188	247	185
7	241	104	12	241	328	282
8	322	138	16	349	422	314
9	413	163	11	411	477	409
10	470	230	13	457	451	405
11	443	242	7	288	305	482
12	300	215	7	141	186	333
13	180	132	2	41	100	201
14	96	86	3	14	50	83
15	50	50	2	8	34	37
16	33	30	2	5	20	24
17	20	21	1	1	26	28
18	25	22	0	0	12	12
19	11	11	0	0	5	5
20	5	5	0	0	1	1
21	1	1	0	0	1	1
22	1	1	0	0	0	0

Chapter 7

Conclusion

The overall goal of this thesis is to investigate the demography of domesticated species using genetic data by adapting and applying population genetics methods to the specificities of such species.

In the first part, focused on plants, we analyze the diversity and structure of 14 rye populations and further infer the demographic history of 11 populations among them. We thereby increase our knowledge about this crop and its genetic history but we also challenge expectations from population genetics theory with a real complex dataset of a domesticated species. Thus explaining discrepancies between previous studies on rye and highlighting confounding effects that might contribute to discrepancies described in other species. As expected wild, weedy rye were more diverse than other accessions but not all landraces were more genetically diverse than varieties due to differences in selection pressure relating to the confounded effects of usage (grain or forage) and geography. The relation between these confounding factors can itself be explained by the performance of rye compared to competing crops (e.g., wheat) in different environments. We therefore advise to take into account the origin and use of the crop to choose populations to sample for later studies. This study also highlighted that it is likely that rye domestication has occurred several times independently and that these different events can still be observed in the genome. Using more markers or adding SNP markers that have a lower mutation rate and are therefore more informative about distant past would be needed to get a better picture of early domestication of rye. Finally, all population including wild material seem to have suffered recent bottlenecks that probably arose due to conservation methods and reproduction within seed banks.

In the second part of the thesis, focusing on animal domesticates for which pedigree data is readily available, we create a model that allows the joint use of pedigree and genetic information. We show that the availability of pedigree information allows to both estimate parameters that are confounded in genetic data such as the effective population size, (effective) sex ratio, and mutation rate. The described model and inference method allow to estimate with relatively high accuracy the demographic parameters of interest on simulated data. When applied to a real pedigree of a Fleckvieh population, the method shows results conform to expectations such as a higher rate of overlapping generation and smaller reproductive population size of males compared to females.

In conclusion, demographic history and life history traits of domesticated species are very complex but the methods developed for natural populations can be adapted and used to infer demography from its impact on genetic patterns. Combining those methods to specifically designed tools using the additional information recorded for breeding and conservation purposes increases inference accuracy and gives access to previously elusive population parameters.

Limitation of this work and implications for future research relating to specific results are discussed per chapter but all chapters of this thesis highlight the importance of recent past events in the inference of the demographic history of domesticated species. Recent demographic history can have a strong influence on the genetic makeup of the population and thereby obscure its ancient history. Highly mutable markers such as SSR and pedigree data are both great tools to infer recent past events. However, as shown in chapter 3 and 4, events that induce a loss of diversity, such as bottlenecks, lead to irreversible information loss. This information cannot be recovered even if the recent history is perfectly known. This underlines further the importance of appropriate population management and conservation in seed banks.

More genetic data such as provided by whole genome sequencing would mitigate the effect that the loss of information due to bottlenecks has on inference. Moreover, linkage between sites observed in sequence data can itself be used to infer demographic history of populations (Li and Durbin 2011; Harris and Nielsen 2013). Combining full genome sequencing data with SSR markers would therefore offer more power to infer demography on a broader time scale from the most recent conservation bottlenecks to the beginning of domestication.

The model and inference method proposed in chapter 5 could as well be adapted to take advantage of linkage information from whole genome data. In its current state, the method considers all sites independent and does not model recombination. However, the way the pedigree constrains genealogies extends to linked loci. Linkage is only broken through recombination that occurs randomly at each reproduction event. Instead of simulating genealogies independently through the pedigree, it is also possible to simulate recombination and possibly estimate recombination rate.

Selection at the scale of the individual is hard coded into the pedigree. Individuals that carry favorable alleles have more reproducing offspring and will therefore be over-represented in the pedigree. Genealogies of neutral sites whether they are linked or not to selected alleles will be constrained by that pedigree. Conversely, chromosome segregation and recombination are the processes that lead to different genealogies within a fixed pedigree. Modeling linkage within the known pedigree could therefore help finding sites under selection with higher accuracy than methods that do not account for the pedigree.

Bibliography

Abu Awad, D., S. Billiard, and V. C. Tran, 2016 Perenniality induces high inbreeding depression in self-fertilising species. Theoretical Population Biology **112**: 43–51.

Allaby, R. G., R. L. Ware, and L. Kistler, 2019 A re-evaluation of the domestication bottleneck from archaeogenomic evidence. Evolutionary Applications **12**: 29–37.

Badr, A., K. Muller, R. Schafer-Pregl, H. El Rabey, S. Effgen, H. H. Ibrahim, C. Pozzi, W. Rohde, and F. Salamini, 2000 On the origin and domestication history of barley (*Hordeum vulgare*). Mol Biol Evol **17**: 499–510.

Baer, C. F., M. M. Miyamoto, and D. R. Denver, 2007 Mutation rate variation in multicellular eukaryotes: causes and consequences. Nature Reviews Genetics **8**: 619–631.

Balter, M., 2007 Seeking Agriculture's Ancient Roots. Science **316**: 1830–1835.

Beaumont, M. A., 2008 Joint determination of topology, divergence time, and immigration in population trees. In *Simulation, Genetics, and Human Prehistory*, edited by S. Matsumura, P. Forster, and C. Renfrew, pp. 135–154, McDonald Institute for Archaeological Research, Cambridge.

Beaumont, M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian Computation in Population Genetics. Genetics **162**: 2025–2035.

Behre, K.-E., 1992 The history of rye cultivation in Europe. Vegetation History and Archaeobotany **1**: 141–156.

Benjamini, Y. and Y. Hochberg, 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society:

Series B (Methodological) **57**: 289–300.

Bertorelle, G., A. Benazzo, and S. Mona, 2010 ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Molecular Ecology **19**: 2609–2625.

Blath, J., A. G. Casanova, N. Kurt, and D. Spanò, 2013 The ancestral process of long-range seed bank models. Journal of Applied Probability **50**: 741–759.

Blum, M. G. B. and O. François, 2010 Non-linear regression models for Approximate Bayesian Computation. Statistics and Computing **20**: 63–73.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet **32**: 314–31.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368**: 455–457.

Briggle, L. W., 1959 *Growing rye*. United States. Department of Agriculture.

Burger, J. C., M. A. Chapman, and J. M. Burke, 2008 Molecular insights into the evolution of crop plants. American Journal of Botany **95**: 113–122.

Busch, J. D., P. M. Waser, and J. A. DeWoody, 2007 Recent demographic bottlenecks are not accompanied by a genetic signature in banner-tailed kangaroo rats (*Dipodomys spectabilis*). Molecular Ecology **16**: 2450–2462.

Börner, A., M. Röder, S. Chebotar, R. K. Varshney, and A. Weidner, 2005 Molecular tools for genebank management and evaluation. Czech J Genet Plant Breed **41**: 122–127.

Campbell, C. D. and E. E. Eichler, 2013 Properties and rates of germline mutations in humans. Trends in genetics : TIG **29**: 575–584.

Cannings, C., 1974 The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models. Advances in Applied Probability **6**: 260.

Cannings, C., 1975 The Latent Roots of Certain Markov Chains Arising in Genetics: A

New Approach, II. Further Haploid Models. Advances in Applied Probability **7**: 264.

Cauvin, J., 2000 *The Birth of the Gods and the Origins of Agriculture*. New studies in archaeology, Cambridge University Press, Cambridge, UK.

Charlesworth, B. and D. Charlesworth, 2010 *Elements of evolutionary genetics*. Roberts and Co. Publishers, Greenwood Village, Colo.

Chebotar, S., M. S. Röder, V. Korzun, B. Saal, W. E. Weber, and A. Börner, 2003 Molecular studies on genetic integrity of open-pollinating species rye (*Secale cereale* L.) after long-term genebank maintenance. Theor Appl Genet **107**: 1469–76.

Chen, H. and K. Chen, 2013 Asymptotic Distributions of Coalescence Times and Ancestral Lineage Numbers for Populations with Temporally Varying Size. Genetics **194**: 721–736.

Chu, J.-H., D. Wegmann, C.-F. Yeh, R.-C. Lin, X.-J. Yang, F.-M. Lei, C.-T. Yao, F.-S. Zou, and S.-H. Li, 2013 Inferring the Geographic Mode of Speciation by Contrasting Autosomal and Sex-Linked Genetic Diversity. Molecular Biology and Evolution **30**: 2519–2530.

Clutton-Brock, T. H., F. E. Guinness, and S. D. Albon, 1982 *Red Deer: Behavior and Ecology of Two Sexes*. University of Chicago Press.

Cornille, A., P. Gladieux, M. J. M. Smulders, I. Roldán-Ruiz, F. Laurens, B. Le Cam, A. Nersesyan, J. Clavel, M. Olonova, L. Feugey, I. Gabrielyan, X.-G. Zhang, M. I. Tenaillon, and T. Giraud, 2012 New Insight into the History of Domesticated Apple: Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties. PLoS Genet **8**: e1002703.

Crow, J. F. and M. Kimura, 1970 *An introduction to population genetics theory*. Harper & Row, New York.

Cunningham, E. P., J. J. Dooley, R. K. Splan, and D. G. Bradley, 2001 Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. Animal Genetics **32**: 360–364.

De Mori, C., A. Nascimento Junior, and M. Z. d. Miranda, 2013 Aspectos econômicos e conjunturais da cultura do centeio no mundo e no Brasil.

Derrida, B., S. C. Manrubia, and D. H. Zanette, 2000 On the Genealogy of a Population of Biparental Individuals. Journal of Theoretical Biology **203**: 303–315.

Donnelly, P. and S. Tavare, 1995 Coalescents and genealogical structure under neutrality. Annual review of genetics **29**: 401–421.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148**: 1667–1686.

Eldon, B. and J. Wakeley, 2006 Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed. Genetics **172**: 2621–2633.

Ellegren, H., 1999 Inbreeding and Relatedness in Scandinavian Grey Wolves Canis Lupus. Hereditas **130**: 239–244.

Engen, S., T. H. Ringsby, B.-E. Sæther, R. Lande, H. Jensen, M. Lillegård, and H. Ellegren, 2007 Effective Size of Fluctuating Populations with Two Sexes and Overlapping Generations. Evolution **61**: 1873–1885.

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology **14**: 2611–2620.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theoretical Population Biology **3**: 87–112.

Ewens, W. J., 1974 A note on the sampling theory for infinite alleles and infinite sites models. Theoretical Population Biology **6**: 143–148.

Ewens, W. J., 1982 On the concept of the effective population size. Theoretical Population Biology **21**: 373–378.

Ewens, W. J., 2004 *Mathematical Population Genetics: I. Theoretical Introduction*. Number 27 in Interdisciplinary Applied Mathematics, Springer, second edition.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust Demographic Inference from Genomic and SNP Data. PLoS genetics **9**: e1003905.

Excoffier, L., A. Estoup, and J.-M. Cornuet, 2005 Bayesian Analysis of an Admixture Model With Mutations and Arbitrarily Linked Markers. Genetics **169**: 1727–1738.

Excoffier, L. and M. Foll, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics .

Excoffier, L. and H. E. L. Lischer, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources **10**: 564–567.

Excoffier, L., P. E. Smouse, and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**: 479–91.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman.

Fay, J. C. and C.-I. Wu, 2000 Hitchhiking Under Positive Darwinian Selection. Genetics **155**: 1405–1413.

Fearnhead, P. and D. Prangle, 2012 Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society Series B **74**: 419–474.

Feuillet, C., P. Langridge, and R. Waugh, 2008 Cereal breeding takes a walk on the wild side. Trends Genet **24**: 24–32.

Fisher, R. A., 1922 On the Dominance Ratio. Bulletin of Mathematical Biology **52**: 297–318.

Fisher, R. A., 1930a The Distribution of Gene Ratios for Rare Mutations. Proceedings of the Royal Society of Edinburgh **50**: 205–220.

Fisher, R. A., 1930b *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

Fu, Y.-X., 1998 Probability of a Segregating Pattern in a Sample of DNA Sequences. Theoretical Population Biology **54**: 1–10.

Fu, Y. X. and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Fuller, D. Q. and R. Allaby, 2009 Seed Dispersal and Crop Domestication: Shattering, Germination and Seasonality in Evolution under Cultivation. In *Annual Plant Reviews Volume 38: Fruit Development and Seed Dispersal*, edited by L. Østergaard, pp. 238–295, Wiley-Blackwell.

Garza, J. C. and E. G. Williamson, 2001 Detection of reduction in population size using data from microsatellite loci. Molecular Ecology **10**: 305–318.

Gasbarra, D., M. J. Sillanpää, and E. Arjas, 2005 Backward simulation of ancestors of sampled individuals. Theoretical Population Biology **67**: 75–83.

Gillespie, J. H., 2000 The neutral theory in an infinite population. Gene **261**: 11–18.

Glémin, S. and T. Bataillon, 2009 A comparative view of the evolution of grasses under domestication. New Phytologist **183**: 273–290.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman, 1995 An evaluation of genetic distances for use with microsatellite loci. Genetics **139**: 463–471.

Griffiths, R. and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Communications in Statistics. Stochastic Models **14**: 273–295.

Griffiths, R. C. and S. Tavare, 1994 Sampling Theory for Neutral Alleles in a Varying Environment. Philosophical Transactions of the Royal Society B: Biological Sciences **344**: 403–410.

Grodzicker, T., J. Williams, P. Sharp, and J. Sambrook, 1974 Physical Mapping of Temperature-sensitive Mutations of Adenoviruses. Cold Spring Harbor Symposia on Quantitative Biology **39**: 439–446.

Gross, B. L. and K. M. Olsen, 2010 Genetic perspectives on crop domestication. Trends in plant science **15**: 529–537.

Gutiérrez, J. P., I. Cervantes, A. Molina, M. Valera, and F. Goyache, 2008 Individual increase in inbreeding allows estimating effective sizes from pedigrees. Genetics, selection, evolution: GSE **40**: 359–378.

Hackauf, B. and P. Wehling, 2002 Identification of microsatellite polymorphisms in an expressed portion of the rye genome. Plant Breeding **121**: 17–25.

Haldane, J. B. S., 1939 The Equilibrium Between Mutation and Random Extinction. Annals of Eugenics **9**: 400–405.

Hammer, K., 1984 Das Domestikationssyndrom. Die Kulturpflanze **32**: 11–34.

Hardy, G. H., 1908 Mendelian proportions in a mixed population. Science **28**: 49–50.

Hardy, O. J., N. Charbonnel, H. Fréville, and M. Heuertz, 2003 Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. Genetics **163**: 1467–1482.

Hardy, O. J. and X. Vekemans, 2002 SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes **2**: 618–620.

Harris, K. and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS genetics **9**: e1003521.

Hartl, D. L. and A. G. Clark, 2007 *Principles of population genetics*. Sinauer Associates, Sunderland, Mass, fourth edition.

Hastings, W. K., 1970 Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika **57**: 97–109.

Hayes, B., R. Fries, M. Lund, D. A. Boichard, P. Stothard, R. Veerkamp, C. Van Tassell, C. Anderson, I. Hulsegge, B. Guldbrandtsen, D. Rocha, D. Hinrichs, A. Bagnato, M. Georges, R. Spelman, J. Reecy, A. L. Archibald, M. Goddard, and B. Gredler-Grandl, 2012 1000 bull genomes consortium project.

Hedrick, P. W., 2005 A Standardized Genetic Differentiation Measure. Evolution **59**: 1633–1638.

Hill, W. G., 1974 Prediction and Evaluation of Response to Selection with Overlapping Generations. Animal Production **18**: 117–139.

Hillman, G. C. and M. S. Davies, 1990 Measured domestication rates in wild wheats and barley under primitive cultivation, and their archaeological implications. Journal of World Prehistory **4**: 157–222.

Hodgkinson, A., E. Ladoukakis, and A. Eyre-Walker, 2009 Cryptic Variation in the Human Mutation Rate. PLOS Biol **7**: e1000027.

Hunter, R. L. and C. L. Markert, 1957 Histochemical Demonstration of Enzymes Separated by Zone Electrophoresis in Starch Gels. Science **125**: 1294–1295.

Hussin, J. G., A. Hodgkinson, Y. Idaghdour, J.-C. Grenier, J.-P. Goulet, E. Gbeha, E. Hip-Ki, and P. Awadalla, 2015 Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nature Genetics **47**: 400–404.

Jaenicke-Després, V., E. S. Buckler, B. D. Smith, M. T. P. Gilbert, A. Cooper, J. Doebley, and S. Pääbo, 2003 Early Allelic Selection in Maize as Revealed by Ancient DNA. Science **302**: 1206–1208.

Jansen, S., B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pagès, E. Graf, T. Wieland, T. M. Strom, T. Meitinger, and R. Fries, 2013 Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics **14**: 446.

Jost, L., 2008 GST and its relatives do not measure differentiation. Molecular Ecology **17**: 4015–4026.

Kaj, I., S. M. Krone, and M. Lascoux, 2001 Coalescent Theory for Seed Bank Models. Journal of Applied Probability **38**: 285–300.

Khlestkina, E. K., M. H. M. Than, E. G. Pestsova, M. S. Röder, S. V. Malyshev, V. Korzun, and A. Börner, 2004 Mapping of 99 new microsatellite-derived loci in rye (*Secale cereale* L.) including 39 expressed sequence tags. Theoretical and Applied Genetics **109**: 725–732.

Khush, G., 1963 Cytogenetic and evolutionary studies in Secale III. Cytogenetics of weedy ryes and origin of cultivated rye. Economic Botany **17**: 60–71.

Kimura, M., 1953 ”Stepping-stone” models of population. Annual Report of the National Institute of Genetics **3**: 62–63.

Kimura, M., 1964 Diffusion Models in Population Genetics. Journal of Applied Probability **1**: 177–232.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893–903.

Kimura, M. and J. F. Crow, 1963 The Measurement of Effective Population Number. Evolution **17**: 279.

Kimura, M. and J. F. Crow, 1964 The Number of Alleles That Can Be Maintained in a Finite Population. Genetics **49**: 725–738.

Kimura, M. and G. H. Weiss, 1964 The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics **49**: 561–576.

Kingman, J., 1982a The coalescent. Stochastic Processes and their Applications **13**: 235–248.

Kingman, J. F. C., 1982b On the Genealogy of Large Populations. Journal of Applied Probability **19**: 27–43.

Klekowski, E. J. and P. J. Godfrey, 1989 Ageing and mutation in plants. Nature **340**: 389–391.

Kranz, A., 1957 Populationsgenetische Untersuchungen am iranischen Primitivroggen. Ein Beitrag zur Systematik, Evolution und Züchtung des Roggens. Zeitschrift für Pflanzenzüchtung **38**: 101–146.

Kuckuck, H., 1956 Report to the government of Iran: The distribution and variation of cereals in Iran (including their related wild species). Report, Food and Agriculture Organization of the United Nations.

Lapierre, M., A. Lambert, and G. Achaz, 2017 Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. Genetics **206**: 439–449.

Leuenberger, C. and D. Wegmann, 2010 Bayesian Computation and Model Selection Without Likelihoods. Genetics **184**: 243–252.

Li, H. and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature **475**: 493.

Lynch, M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster, 2016 Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics **17**: 704–714.

Ma, R., T. Yli-Mattila, and S. Pulli, 2004 Phylogenetic relationships among genotypes of worldwide collection of spring and winter ryes (*Secale cereale* L.) determined by RAPD-PCR markers. Hereditas **140**: 210–221.

Malécot, G., 1948 *Les mathématiques de l'hérédité*. Masson, Paris.

Matos, M., O. Pinto-Carnide, and C. Benito, 2001 Phylogenetic Relationships among Portuguese Rye Based on Isozyme, RAPD and ISSR Markers. Hereditas **134**: 229–236.

Mc Parland, S., J. F. Kearney, M. Rath, and D. P. Berry, 2007 Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. Journal of Animal Science **85**: 322–331.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953 Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics **21**: 1087–1092.

Michalakis, Y. and L. Excoffier, 1996 A Generic Estimation of Population Subdivision Using Distances Between Alleles With Special Reference for Microsatellite Loci. Genetics **142**: 1061–1064.

Miedaner, T., 2010 *Grundlagen der Pflanzenzüchtung*. DLG-Verlag, Frankfurt/Main.

Moran, P. A. P., 1958 Random processes in genetics. Mathematical Proceedings of the Cambridge Philosophical Society **54**: 60–71.

Möhle, M., 1998a Coalescent Results for Two-Sex Population Models. Advances in Applied Probability **30**: 513–520.

Möhle, M., 1998b A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. Advances in Applied Probability **30**: 493–512.

Möhle, M., 1998c Robustness results for the coalescent. Journal of Applied Probability **35**: 438–447.

Möhle, M., 1999 Weak convergence to the coalescent in neutral population models. Journal of Applied Probability **36**: 446–460.

Möhle, M., 2001 Forward and backward diffusion approximations for haploid exchangeable population models. Stochastic Processes and their Applications **95**: 133–149.

Möhle, M. and S. Sagitov, 2001 A Classification of Coalescent Processes for Haploid Exchangeable Population Models. The Annals of Probability **29**: 1547–1562.

Nei, M., 1973 Analysis of Gene Diversity in Subdivided Populations. Proceedings of the National Academy of Sciences **70**: 3321–3323.

Nei, M., 1987 *Molecular evolutionary genetics*. Columbia University Press, New York, NY, USA.

Nei, M. and W. H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America **76**: 5269–5273.

Nei, M., T. Maruyama, and R. Chakraborty, 1975 The Bottleneck Effect and Genetic Variability in Populations. Evolution **29**: 1–10.

Nelson, M. M. R., D. Wegmann, M. G. M. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14002 people. Science **337**: 100–104.

Nesbitt, M. and D. Samuel, 1998 Wheat domestication: archaeobotanical evidence. Science **279**: 1431.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154**: 931–42.

Nordborg, M., 2001 Coalescent theory. In *Handbook of Statistical Genetics*, edited by D. J. Balding, M. J. Bishop, and C. Cannings, pp. 179–212, John Wiley & Sons, Ltd.

Nordborg, M. and S. M. Krone, 2002 Separation of time scales and convergence to the coalescent in structured populations. In *Modern developments in theoretical population genetics: the legacy of Gustave Malécot*, pp. 194–232, Oxford University Press.

Nunney, L., 1993 The Influence of Mating System and Overlapping Generations on Effective Population Size. Evolution **47**: 1329–1341.

Ohta, T. and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetics Research **22**: 201–204.

Ossowski, S., K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, and M. Lynch, 2010 The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science (New York, N.Y.) **327**: 92–94.

Paradis, E., J. Claude, and K. Strimmer, 2004 APE: Analyses of phylogenetics and evolution in R language. Bioinformatics **20**: 289–90.

Parat, F., G. Schwertfirm, U. Rudolph, T. Miedaner, V. Korzun, E. Bauer, C.-C. Schön, and A. Tellier, 2016 Geography and end use drive the diversification of worldwide winter rye populations. Molecular Ecology **25**: 500–514.

Persson, K., O. Díaz, and R. Von Bothmer, 2001 Extent and Patterns of RAPD Variation in Landraces and Cultivars of Rye (*Secale cereale* L.) from Northern Europe. Hereditas **134**: 237–243.

Persson, K. and R. Von Bothmer, 2002 Genetic diversity amongst landraces of rye (*Secale cereale* L.) from northern Europe. Hereditas **136**: 29–38.

Pitman, J., 1999 Coalescents With Multiple Collisions. The Annals of Probability **27**: 1870–1902.

Polanski, A., A. Bobrowski, and M. Kimmel, 2003 A note on distributions of times to coalescence, under time-dependent population size. Theoretical Population Biology **63**: 33–40.

Prangle, D., P. Fearnhead, M. P. Cox, P. J. Biggs, and N. P. French, 2014 Semi-automatic selection of summary statistics for ABC model choice. Statistical Applications in Genetics and Molecular Biology **13**: 67–82.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution **16**: 1791–1798.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945–59.

Pudlo, P., J.-M. Marin, A. Estoup, J.-M. Cornuet, M. Gautier, and C. P. Robert, 2016 Reliable ABC model choice via random forests. Bioinformatics **32**: 859–866.

Ribeiro, M., L. Seabra, A. Ramos, S. Santos, O. Pinto-Carnide, C. Carvalho, and G. Igrejas, 2012 Polymorphism of the storage proteins in Portuguese rye (*Secale cereale* L.) populations. Hereditas **149**: 72–84.

Rienzo, A. D., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proceedings of the National Academy of Sciences **91**: 3166–3170.

Rindos, D., 1984 *The Origins of Agriculture: An Evolutionary Perspective*. Academic Press.

Robert, C. P., J.-M. Cornuet, J.-M. Marin, and N. S. Pillai, 2011 Lack of confidence in approximate Bayesian computation model choice. Proceedings of the National Academy of Sciences **108**: 15112–15117.

Rogowsky, P. M., F. L. Y. Guidet, P. Langridge, K. W. Shepherd, and R. M. D. Koebner, 1991 Isolation and characterization of wheat-rye recombinants involving chromosome arm 1ds of wheat. Theor Appl Genet **82**: 537–544.

Sagitov, S., 1999 The General Coalescent with Asynchronous Mergers of Ancestral Lines. Journal of Applied Probability **36**: 113–1125.

Sano, A., A. Shimizu, and M. Iizuka, 2004 Coalescent process with fluctuating population size and its effective size. Theoretical Population Biology **65**: 39–48.

Schaibley, V. M., M. Zawistowski, D. Wegmann, M. G. Ehm, M. R. Nelson, P. L. St Jean, G. R. Abecasis, J. Novembre, S. Zöllner, and J. Z. Li, 2013 The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome research **23**: 1974–84.

Schweinsberg, J., 2000 Coalescents with Simultaneous Multiple Collisions. Electronic Journal of Probability **5**: 1–50.

Sencer, H. A. and J. G. Hawkes, 1980 On the origin of cultivated rye. Biological Journal of the Linnean Society **13**: 299–313.

Sjödin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg, 2005 On the Meaning and Existence of an Effective Population Size. Genetics **169**: 1061–1070.

Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**: 457–462.

Slatkin, M., 2005 Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. Molecular Ecology **14**: 67–73.

Slatkin, M. and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**: 555–562.

Smith, B. D., 2006 Documenting Domesticated Plants in the Archaeological Record. In *Documenting Domestication: New Genetic and Archaeological Paradigms*, edited by M. A. Zeder, D. G. Bradley, E. Emshwiller, and B. D. Smith, pp. 15–24, University of California Press.

Stracke, S., A. G. Schilling, J. Forster, C. Weiss, C. Glass, T. Miedaner, and H. H. Geiger, 2003 Development of PCR-based markers linked to dominant genes for male-fertility restoration in Pampa CMS of rye (*Secale cereale* L.). Theoretical and Applied Genetics

106: 1184–90.

Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz, 2013 Approximate Bayesian Computation. PLoS Comput Biol **9**: e1002803.

Tajima, F., 1983 Evolutionary Relationship of DNA Sequences in Finite Populations. Genetics **105**: 437–460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring Coalescence Times from DNA Sequence Data. Genetics **145**: 505–518.

Tellier, A. and C. Lemaire, 2014 Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Molecular Ecology **23**: 2637–2652.

Tenaillon, M. I., J. U'Ren, O. Tenaillon, and B. S. Gaut, 2004 Selection Versus Demography: A Multilocus Investigation of the Domestication Process in Maize. Molecular Biology and Evolution **21**: 1214–1225.

Thuillet, A.-C., D. Bru, J. David, P. Roumet, S. Santoni, P. Sourdille, and T. Bataillon, 2002 Direct Estimation of Mutation Rate for 10 Microsatellite Loci in Durum Wheat, Triticum turgidum (L.) Thell. ssp durum desf. Molecular Biology and Evolution **19**: 122–125.

Vigouroux, Y., J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. D. Beavis, J. S. C. Smith, and J. Doebley, 2002 Rate and pattern of mutation at microsatellite loci in maize. Molecular biology and evolution **19**: 1251–1260.

Villa, T. C. C., N. Maxted, M. Scholten, and B. Ford-Lloyd, 2005 Defining and identifying crop landraces. Plant Genetic Resources **3**: 373–384.

Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts & Company Publishers.

Wakeley, J., L. King, B. S. Low, and S. Ramachandran, 2012 Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman's Coalescent. Genetics **190**: 1433–1445.

Wakeley, J. and O. Sargsyan, 2009 Extensions of the Coalescent Effective Population Size. Genetics **181**: 341–345.

Wakeley, J. and T. Takahashi, 2003 Gene genealogies when the sample size exceeds the effective size of the population. Molecular biology and evolution **20**: 208–213.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7**: 256–276.

Weber, J. L. and C. Wong, 1993 Mutation of human short tandem repeats. Human Molecular Genetics **2**: 1123–1128.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood. Genetics **182**: 1207–1218.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier, 2010 ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics **11**: 116.

Wehrhahn, C. F., 1975 The Evolution of Selectively Similar Electro-Phoretically Detectable Alleles in Finite Natural Populations. Genetics **80**: 375–394.

Weinberg, W., 1908 Über den nachweis der vererbung beim menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg **64**: 368–382.

Weir, B. S. and C. C. Cockerham, 1984 Estimating F-Statistics for the analysis of population structure. Evolution **38**: 1358–1370.

Wilkins, A. S., R. W. Wrangham, and W. T. Fitch, 2014 The "Domestication Syndrome" in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. Genetics **197**: 795–808.

Wolfe, K. H., W.-H. Li, and P. M. Sharp, 1987 Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs. Proceedings of the National Academy of Sciences of the United States of America **84**: 9054–9058.

Wright, S., 1922 Coefficients of Inbreeding and Relationship. The American Naturalist

56: 330–338.

Wright, S., 1931 Evolution in Mendelian Populations. Genetics **16**: 97–159.

Wright, S., 1938a The distribution of gene frequencies under irreversible mutation. Proceedings of the National Academy of Sciences **24**: 253–259.

Wright, S., 1938b Size of population and breeding structure in relation to evolution. Science **87**: 425–431.

Wright, S., 1943 Isolation by distance. Genetics **28**: 114–138.

Wright, S., 1949 The genetical structure of populations. Annals of Eugenics **15**: 323–354.

Wright, S., 1978 *Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations*, volume 4. University of Chicago Press, Chicago.

Xu, H., R. Chakraborty, and Y.-X. Fu, 2005 Mutation Rate Variation at Human Dinucleotide Microsatellites. Genetics **170**: 305–312.

Yamasaki, M., M. I. Tenaillon, I. V. Bi, S. G. Schroeder, H. Sanchez-Villeda, J. F. Doebley, B. S. Gaut, and M. D. McMullen, 2005 A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell **17**: 2859–72.

Zeder, M. A., 2015 Core questions in domestication research. Proceedings of the National Academy of Sciences of the United States of America **112**: 3191–3198.

Zohary, D., M. Hopf, and E. Weiss, 2013 *Domestication of plants in the Old World*. Oxford University Press, Oxford, UK, paperback 4th edition.