**PhD Thesis**[1]

# Genetic variation in natural populations: a modeller's perspective

**Marina Rafajlović**


GÖTEBORGS UNIVERSITET

Department of Physics
University of Gothenburg
Göteborg, Sweden 2014

[1]The thesis is available at http://physics.gu.se/∼rmarina/Marina_Rafajlovic/Home_files/PhD.pdf

**Front cover** Panel **a**: the expansion of genetic variation from a large source (mainland) into islands situated at increasing distances from the mainland. The genetic variation exhibits both temporal and spatial fluctuations. Bursts of high genetic variation coming from the mainland are supported by a high level of multiple paternity in the population. Refer to Fig. 4.3**A** in Chapter 4 for details. Panels **b-c**: the sexual structure of a colonising population in which each individual reproduces both sexually and asexually. Reproduction occurs locally in vicinity of parental individuals. Since sexual reproduction is possible if both sperms and eggs are present locally, sexual reproduction is hindered in newly colonised areas, and clonal colonies expand over the habitat during colonisation. The spread and the persistence time of the clonal colonies are larger when the rate of clonal reproduction is larger (**c**) than when it is smaller (**b**), all else being the same. Refer to Fig. 5.2**e-f** in Chapter 5 for further details.

ISBN 978-91-628-9069-8

Printed by Ineko AB
Göteborg 2014

# Genetic variation in natural populations: a modeller's perspective

Marina Rafajlović
Department of Physics
University of Gothenburg
SE-412 96 Göteborg, Sweden

**Abstract**

Thanks to advances in genome sequencing, empirical patterns of within- and between-species genetic variation are readily available. By studying these patterns much has been learned about the evolutionary histories of species. But the causes and consequences of different evolutionary histories are still difficult to tell apart. To this end, comparative analyses of genetic variation under different models are required. This thesis analyses genetic variation under specific models that are relevant for a number of biological species.

Firstly, this thesis discusses a method for inferring the population-size history of the population in question using simulated, as well as empirically observed frequency spectra of mutations. The method performs well when applied to simulated data, provided that a large number of mutations is sampled. However the estimation based on empirical data is biased. Secondly, the thesis studies a mainland-island colonisation model. The model allows for different levels of multiple paternity in the population. Multiple paternity promotes genetic variation. This effect is much larger during colonisation than on the long run. Therefore, multiple paternity may facilitate the establishment of species in new areas. Thirdly, this thesis analyses a colonisation model for species that reproduce both sexually and asexually, and have limited dispersal capabilities. Due to limited dispersal capabilities, sexual reproduction may be hindered locally, especially during colonisation. Unless the individuals are highly sexual, a few clones establish the front of the colonisation forming wide clonal colonies. Finally, this thesis analyses a joint effect of migration, selection and random genetic drift during adaptation in subpopulations subject to different environments. When divergent adaptation is driven by mutations, the frequency at which mutations appear, as well as how strongly they are selected for are the decisive parameters for whether or not subpopulations can adapt to their respective environments despite migration and drift. This remains to be analysed further.

**Keywords:** *coalescent process, site frequency spectrum, multiple paternity, dominant clone, divergent selection.*

This thesis consists of an introductory text and the following three appended research papers, henceforth referred to as [I], [II], and [III]:

[I] M. Rafajlović, A. Klassmann, A. Eriksson, T. Wiehe, and B. Mehlig, *Demography-adjusted tests of neutrality based on genome-wide SNP data*, Theoretical Population Biology **95**, 1–12, (2014).

[II] M. Rafajlović, A. Eriksson, A. Rimark, S. Hintz-Saltin, G. Charrier, M. Panova, C. André, K. Johannesson, and B. Mehlig, *The effect of multiple paternity on genetic diversity of small populations during and after colonisation*, PLos ONE **8**(10): e75587 (2013).

[III] M. Rafajlović, D. Kleinhans, C. Gulliksson, J. Fries, D. Johansson, A. Ardehed, L. Sundqvist, R. Pereyra, B. Mehlig, P. R. Jonsson, and K. Johannesson, *A neutral model can explain geographic patterns of sexual and asexual reproduction during colonisation and long thereafter*, in manuscript.

Two additional papers co-authored by the author of this thesis ([1, 2]) were discussed in the Licentiate thesis [3].

Specific contributions of the thesis' author (referred to as MR below) to the papers [I, II, III]:

- Ref. [I]: MR wrote the first version of the manuscript, derived theoretical results, performed computer simulations, executed demography estimation for the data simulated, and for the empirical data from ten Human populations sampled.

- Ref. [II]: MR wrote the first version of the manuscript, constructed the mating model, fitted the parameters of the mating model to the empirical data, derived theoretical results, performed computer simulations.

- Ref. [III]: MR contributed to writing the first version of the manuscript, participated in designing the study, planned the modelling part of the project, derived theoretical expectations.

# Acknowledgments

I should like to express my deep gratitude to my supervisor Bernhard Mehlig for his guidance, and optimism throughout this work. Bernhard also tried to teach me to simplify my wordings, but I am not completely sure if he was successful :).

Several parts of this thesis resulted from fruitful discussions that I had with Kerstin Johannesson. Thank you, Kerstin, for introducing me to the wonderful species *Littorina saxatilis* and *Fucus radicans*, and for posing interesting questions. I believe this thesis contains answers to some of them. I am also very grateful to Kerstin for her enormous support and encouragement during the past years.

I am very grateful to Anna Emanuelsson, Christian Gulliksson, Johan Fries, Fengchong Wang, Elke Schaper and Anna Rimark.

I would also like to thank Roger Butlin, Anna Godhe, Per R. Jonsson, Marina Panova, Carl André, Helen Nilsson Sköld, Anders Eriksson, and Serik Sagitov for interesting and helpful discussions on the topic.

Big thanks to Matteo Bazzanella, whom I continuously 'bothered' with my results. Thanks also to Erik Werner for proofreading a part of the thesis, as well as to Jonas Einarsson for helping me with Bibliography. I am also thankful to all colleagues from 'the third floor' for contributing to a friendly working environment.

I am extremely grateful to my sisters and parents for their love and encouragement, and for teaching and showing me that one's family is the most valuable treasure in one's life.

Finally, my greatest gratitude belongs to my son Ilija, my daughter Lena, and my husband Stevan for their understanding and patience, for not letting me give up, and most importantly, for their invaluable love.

I acknowledge the support from the Department of Physics at the University of Gothenburg, Gothenburg, Sweden.

Marina Rafajlović
Göteborg
October 6, 2014

# Contents

# 1

# Introduction

Genetic variation within and between biological species is a result of the interplay of a number of evolutionary processes such as mutation, recombination, random genetic drift, population-size fluctuations, migration, and natural selection. Despite numerous theoretical advances made in the field of population genetics, the mechanisms that allow the existing species to evolve in response to temporally and spatially changing environments (adaptation), and possibly give rise to new species are not fully understood [4–10]. The questions in population genetics that are still hotly debated include the following. Why do most species reproduce sexually, whereas some reproduce asexually or have both modes of reproduction [11–14]? Under which conditions do individuals in sexually reproducing species exhibit mate preferences [15, 16]? Specifically, do mate preferences promote or inhibit the tendency of a species to produce new ones through the process of speciation [7]? Which genome regions initialise the process of speciation [7–10]? When during speciation should one expect to observe 'concentrated genetic architectures' of genes that drive speciation, and what is the mechanism behind their establishment [17]? Due to the heritability of genetic variation, the answers to these and related questions can be gained by analysing empirical genome-wide patterns of genetic variation from within and between species to search for signatures of adaptation, and/or speciation [7–10]. Needless to say, the interpretation of empirical genetic data relies on a theoretical understanding of how different evolutionary processes contribute to establishing genetic variation. In this and the following chapters of this thesis, a number of past and present theoretical advances in understanding the patterns of genetic variation in natural populations are outlined.

Genetic variation is established via random mutations, and, in sexu-

ally reproducing organisms, via recombination. Mutations alter genome sequences (also called *loci*) by changing one or more nucleotides along the locus. The process of recombination, instead, re-arranges pairs of maternal or paternal genome sequences, thus producing individuals with unique arrangements of genetic sequences along genomes, that is, unique genotypes.

However, because natural populations have finite sizes, an individual in the population in question may by chance fail to give rise to offspring. This effect is referred to as *random genetic drift*. Thus, random genetic drift limits genetic variation. This effect is larger in smaller populations.

In addition to loss by chance, demographic processes such as colonisation of new habitats, expansion or contraction of population size, mating patterns (sexual and/or asexual reproduction, locally confined reproduction due to spatial structures of populations, etc.) can also affect the extent of population genetic variation. For example, severe reductions of population size limit the number of genotypes preserved in the population.

The extent of genetic variation is also influenced by the process of migration between geographically structured populations. On the one hand, migrants can bring new genetic variants into populations, and hence increase the within-population genetic variation. On the other hand, migrants can decrease the between-population genetic variation.

Finally, natural selection acts in such a way that the better adapted individuals have a higher chance of surviving, and establishing their offspring than those poorly adapted [18]. However unlike the processes listed above which affect all genome-wide regions in a mutually similar manner, natural selection acts locally on the genome regions which determine the degree of individual's adaptation to the environment, and on the closely linked neutral regions (*hitchhiking*) [19, 20]. Natural selection removes deleterious mutations at the loci targeted by selection (*negative selection*) or increases the frequency of beneficial mutations (*positive selection*) [21]. Thus, natural selection is expected to reduce the within-population genetic variation along regions targeted by selection, and their closely linked neighbourhood by selecting against the genotypes that are poorly adapted in the environment in question. However, natural selection may also favour individuals with different genetic variants (*alleles*) at a given locus over the individuals with the same alleles at the locus. This type of selection is known as *balancing selection* [21]. Moreover, when populations are exposed to different environmental conditions, natural selection is expected to favour different genotypes in the different populations, thus increasing the between-population genetic variation along the genome regions subject to selection [6, 7, 9, 22–24]. This is the so-

called *divergent selection*.

In summary, the contributions of the processes described above in establishing genetic variation are in general difficult to analyse jointly, because they may differ over time but also between different genome regions. But is it necessary to account for all these processes when analysing empirical patterns of genetic variation? Which mechanism are important for establishing and maintaining genetic variation in natural populations? To answer these questions, more or less complex models of the evolution of genetic variation must be analysed, and the underlying model predictions must be compared against each other. This thesis provides an advance in the endeavour in answering these questions by considering a number of models relevant for biological species.

The basic models, such as the Wright-Fisher model [11, 25] or the Moran model [26] neglect the effect of selection. These models are very important because empirical studies [27–31] suggest that the majority of genome-wide genetic variation is neutral or under weak selection. Moreover, neutral genetic variation provides a background for finding the signatures of selection along the genome of the population in question [7, 9, 10, 32].

In order to understand the patterns of neutral genetic variation using population-genetics models, one commonly makes a number of simplifying assumptions: 1) populations are of constant size, 2) mating is random, 3) populations are well mixed [11, 25, 33–35]. Under these assumptions, the coalescent process [33] provides a powerful method for generating ancestral gene genealogies of a sample of alleles at one or more loci from a given population. Consequently, it provides a complete description of the expected patterns of neutral genetic variation. However, it is likely that neither of the three assumptions listed above is fulfilled in natural populations [36–40]. Therefore, it must be understood: what are the consequences on the patterns of genetic variation upon relaxing one or more of these assumptions? Under which conditions is the coalescent process not appropriate?

In Refs. [41, 42] it was suggested that the coalescent process describes well gene genealogies under a varying population size, provided that the timescale of population-size changes is much shorter or much longer than the corresponding coalescent timescale. When the two timescales are of the same order, the approximation based on the coalescent process fails. The single-locus gene genealogies under varying population sizes are fully determined by the results in Ref. [1] (their Eq. (20)). The effect of population-size variations on two-locus gene genealogies is in general more difficult to analyse than the effect on one-locus genealogies, and simplifying assumptions concerning the population-size history need to

be made in models. In Ref. [2] it was shown that severe reductions of the population size during recurrent bottlenecks can promote the degree of association between pairs of physically distant loci in comparison to that expected using the coalescent process. This was discussed in more detail in the Licentiate thesis [3].

Using the results of Ref. [1], the moments of the site frequency spectrum of mutations at a neutral locus under a given demographic history can be computed [I]. This was used in Ref. [I] to estimate the demographic histories of Human populations using empirical genome-wide data gathered in the 1000 Genomes Project [43]. As shown in Ref. [I], empirical distributions of commonly used tests of neutrality such as Tajima's $D$ [32], Fay & Wu's $H$ [20] and others, differ substantially between the different populations. The question arises: since the empirical test distributions are different, how can one compare the extents of selection at candidate loci under selection between the different populations? This problem is resolved by integrating the estimated demography of the population in question into the tests of neutrality, yielding *demography-adjusted tests* [I]. Indeed, the empirical distributions of demography-adjusted tests are found to be similar between the different Human populations [I]. However the demographies estimated using empirical data are inevitably biased due to the assumptions made to facilitate the estimation. It remains to be understood how this bias influences the distributions of demography-adjusted tests. The results obtained in Ref. [I] are further discussed in Chapter 3.

Apart from the effect of population-size fluctuations, it must be understood how different mating patterns, with or without limited movement or dispersal capabilities of individuals influence the shape of gene genealogies and hence the patterns of genetic variation in natural populations. Under which conditions are the assumptions that the population is well-mixed and that its individuals exhibit random mating inappropriate? In the model analysed in Ref. [II], mating is not random. Instead, mating is allowed to result in higher or lower levels of multiple paternity. Multiple paternity is observed in many species, including e. g. the marine snail *Littorina saxatilis* [15], and a number of fish and invertebrate species [44–47]. The population in the model is also assumed not to be freely mixing. Instead, the population inhabits a geographically structured habitat with a large source population (mainland) and a number of islands that are assumed to be empty initially. While individuals in each patch are assumed to mix freely, migration is allowed to occur only between closest neighbouring patches. The analysis in Ref. [II] shows that when the population establishes a steady state, the gene genealogies are well described by the coalescent process, but with a coalescent

timescale that depends on an effective population size [48]. The effective population size depends on the level of multiple paternity in the population [II]. However on short timescales, that is, during the establishment of the individual island populations, the resulting gene genealogies and the corresponding genetic variation cannot be described in terms of the effective population size alone [II]. These results are further discussed in Chapter 4.

The model presented and analysed in Ref. [III] considers the spatial and genetic structure of a colonising population in which each individual reproduces both sexually and asexually. Examples of species that have the capacity for both sexual and asexual reproduction are the seaweed *Fucus radicans* [49], aquatic plant *Butomus umbellatus* [50] and others. Despite the fact that most species reproduce sexually [13], some species are highly asexual, especially in young habitats or during expansions [50–56]. The dominance of asexual reproduction has been argued for by a number of selection-based hypotheses [53, 57–61]. But these hypotheses have been difficult to prove empirically [62, 63]. The question is: under which conditions can asexuals dominate over sexuals assuming that genetic differences between them are selectively neutral? An important difference between sexual and asexual reproduction is that the former requires both sperms and eggs. Therefore, limited dispersal capabilities and the underlying local sexual structure of the population can be important confounding factors for sexual reproduction, as suggested in Ref. [64]. This was tested in Ref. [III]. The results in Ref. [III] show that clonal colonies establish the front of colonisation as long as the rate of production of clonal propagules is not too low. On the long run the population establishes a homogeneous sex ratio, and the overall frequency of sexual over asexual reproduction remains constant for a long time. But due to the limited dispersal capabilities and locally confined reproduction, the overall genotypic variation in the population differs from that expected under models of well-mixed populations with mixed sexual and asexual reproduction, such as the model analysed in Ref. [65]. The results obtained in Ref. [III] are further discussed in Chapter 5.

Finally, the effect of natural selection must be taken into account. Basic models of natural selection consider a single well-mixed population exposed to a given fixed environment [11, 25, 66, 67]. However, it is well understood that when a species is subject to spatially changing environmental conditions, its subpopulations can diverge, and hence initialise speciation [6, 7, 9, 22–24]. Therefore, it is necessary to analyse models of geographically structured populations subject to divergent selection. The joint effect of migration and divergent selection has been extensively studied in the past [17, 68–76]. These studies showed that

migration can limit or prevent divergence between subpopulations that are exposed to opposing environments. In Ref. [17] it was shown that divergent subpopulations tend to establish 'concentrated genetic architectures'. However it is still not well understood: how does the joint effect of selection and migration change during the course of adaptation? Under which conditions can two diverged subpopulations diverge further despite the effect of migration and random genetic drift? When during adaptation is the population expected to establish 'concentrated genetic architectures' [17]? What is the mechanism behind? These questions are discussed in Chapter 6.

This thesis presents the background to the methods used, and of findings discussed in Refs. [I, II, III]. It also presents a number of unpublished results that are summarised in Chapter 6. The thesis is organised as follows.

The basic models and concepts used in population genetics are introduced in Chapter 2. This chapter is essentially Chapter 2 in the Licentiate thesis [3]. It provides an introduction to the Wright-Fisher model of reproduction [11, 25], an introduction to the coalescent process [33], then to a number of common models of the processes of mutation and recombination, and to modelling natural selection [11, 25, 66, 67]. Chapters 3-5 discuss the models analysed and the results obtained in the papers [I, II, III], respectively. Chapter 6 outlines selected unpublished results on local adaptation in two partly isolated subpopulations. Finally, Chapter 7 summarises and discusses the main findings of this thesis. Selected calculations are given in appendices.

# 2

# Modelling population genetics

This chapter explains the basic models used in population genetics. The chapter is essentially Chapter 2 in the Licentiate thesis [3]. It is organised as follows. Section 2.1 explains the Wright-Fisher model of reproduction [11, 25]. Section 2.2 summarises the idea behind and the main results of the coalescent process, a powerful method for tracing the ancestry of a sample of individuals from the population in question [33]. Modelling of neutral mutations, and of recombination are covered in Section 2.3 [21, 35]. Section 2.4 explains common models of the process of natural selection [11, 25, 66, 67, 77, 78].

## 2.1 Wright-Fisher model

The Wright-Fisher model [11, 25] for the population consisting of $N$ haploid[1] individuals is based on the following three assumptions:

- generations are discrete and non-overlapping,
- the population size $N$ is constant, independent of time,
- the number of offspring of an individual is binomially distributed with the parameters $N$, and $1/N$. Here $N$ is the number of trials. For each trial the success probability that this individual establishes an offspring is equal to $1/N$.

[1] In a cell of a haploid organism one finds a single copy of each chromosome. In diploid organisms, by contrast, only sex cells carry a single copy of each chromosome (thus, sex cells are haploid), whereas somatic cells carry paired chromosomes. These are diploid cells. Two copies of a single chromosome in a diploid cell typically differ in their genetic sequences.

The first assumption listed above implies that the members of the parental generation produce progeny simultaneously, and that they are replaced immediately afterwards. This assumption may be relaxed. For example, in the Moran model [26] a single randomly chosen individual gives rise to a child in each time step. At the same time, a single randomly chosen individual dies. This individual may be the one that gave rise to a child, but it may also be some other individual. In this model, one generation is assumed to be equal to $N$, that is, to the average number of time steps needed for an individual to be replaced by an offspring. The generations are, thus, overlapping. Most of analyses in this thesis assume non-overlapping generations, but a model introduced and discussed in Chapter 5 accounts both for overlapping and non-overlapping generations.

In order to understand how genetic variation under the Wright-Fisher model evolves in time, each individual is characterised by its genetic sequence at the locus of interest (hereafter referred to as *allele*). Under the three assumptions listed above, the Wright-Fisher population can be generated as follows. The population in generation $\ell + 1$ is obtained by sampling at random with replacement $N$ alleles from the alleles in the population in generation $\ell$. Each of the alleles from generation $\ell$ can be a parent to an allele in generation $\ell + 1$ with probability $1/N$. Similarly, one generates the population in generation $\ell + 2$ by sampling from the individuals (alleles) in generation $\ell + 1$, and so on.

Due to random sampling in the Wright-Fisher population of a finite size $N$, a given allele may become lost by chance. This effect is referred to as *random genetic drift*. The effect of genetic drift is regulated by the population size $N$: drift is stronger when $N$ is smaller. This can be explained as follows. Consider a locus with two possible alleles, denoted by $A_1$ and $A_2$. Assuming that in generation $\ell$ there are $i$ copies of $A_1$, the probability that there are $j$ copies of $A_1$ in generation $\ell + 1$ is:

$$p_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j} . \tag{2.1}$$

The process defined with transition probabilities (2.1) inevitably reaches an absorbing state that is characterised by complete loss of $A_1$ (or $A_2$) and fixation of $A_2$ (or $A_1$) [79]. The average number of generations $\ell_{\text{loss}}$ that the population needs to experience complete loss of genetic variation at a given locus (also known as the mean fixation time) depends on the initial frequency $p_0$ of allele $A_1$ according to [79]

$$\ell_{\text{loss}}(p_0) \approx -2N \left[p_0 \ln(p_0) + (1 - p_0)\ln(1 - p_0)\right] . \tag{2.2}$$

Here it is assumed that $N \gg 1$. It follows that fixation occurs faster when populations are of smaller size. This implies that the effect of random genetic drift is stronger for smaller populations. The fixation of $A_1$ occurs with probability $p_0$ [79]. Note also that in the limit of infinite population size the allele frequencies are expected to remain unchanged. This is commonly referred to as *Hardy-Weinberg equilibrium* [80, 81].

The Wright-Fisher model can be extended to account for sexually reproducing diploid organisms. As an example consider a well-mixed population of $N_\mathrm{f}$ females and $N_\mathrm{m}$ males that mate randomly. Since the individuals are diploid, the population contains $2(N_\mathrm{f} + N_\mathrm{m})$ alleles. Assuming that Mendelian inheritance[2] applies, one finds that the probability that two alleles sampled at random from $2(N_\mathrm{f} + N_\mathrm{m})$ alleles stem from a single allele in the previous generation is equal to $(2N_\mathrm{e})^{-1}$, where [48, 79]

$$N_\mathrm{e} = \frac{4N_\mathrm{f}N_\mathrm{m}}{N_\mathrm{m} + N_\mathrm{f}} \, . \tag{2.3}$$

Here, $N_\mathrm{e}$ stands for an effective population size.

In summary, the Wright-Fisher and Moran model provide a method for tracing the ancestry of the population or a given sample of the population generation by generation. But instead of doing this generation by generation, the ancestry of the sample can be obtained much faster using the coalescent process [33, 35]. This is discussed next.

## 2.2 Coalescent process

The coalescent process provides a fast method for tracing backwards the ancestry of alleles sampled at the present time until the most recent common ancestor (MRCA) of the sample is found. In what follows, the ancestry of a given sample is called the gene genealogy (Fig. 2.1). This section outlines the basic concepts behind the coalescent process. The following is based on the results in Refs. [33, 35]

Consider a Wright-Fisher population of $N$ haploid individuals. A gene genealogy of $n$ sequences sampled from this population at the present time can be inferred using the standard coalescent theory. The derivation of Eqs. (2.4)-(2.9) given below was described in Refs. [33, 35].

The probability $P(n, 1)$ that $n$ alleles sampled have $n$ different ances-

[2] According to Mendelian inheritance, a child inherits at random one of the two maternal alleles, and at random one of the two paternal alleles.

tors one generation back in time is

$$P(n,1) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{N}\right) \ . \tag{2.4}$$

Assuming that the population size is much larger than the sample size ($N \gg n$), $P(n,1)$ becomes

$$P(n,1) \approx 1 - \sum_{i=1}^{n-1} \frac{i}{N} = 1 - \frac{n(n-1)}{2N} \ . \tag{2.5}$$

The probability $P(n,\ell)$ that the sample has $n$ distinct ancestors $\ell$ generations back in time satisfies $P(n,\ell) = P(n,1)^{\ell}$. In the case $N \gg n$, $P(n,\ell)$ can be approximated by

$$P(n,\ell) \approx \left(1 - \frac{n(n-1)}{2N}\right)^{\ell} \ . \tag{2.6}$$

When $n^2 \ll N$, Eq. (2.2) reduces to

$$P(n,\ell) \approx \mathrm{e}^{-\ell \frac{n(n-1)}{2N}} \ . \tag{2.7}$$

Therefore, $\ell + 1$ generations back in time, the number of ancestors of a sample is less than $n$ with probability $P_{\mathrm{c}}(n,\ell+1)$ given by

$$P_{\mathrm{c}}(n,\ell+1) \approx \frac{n(n-1)}{2N} \mathrm{e}^{-\ell \frac{n(n-1)}{2N}} \ . \tag{2.8}$$

In other words, in generation $\ell + 1$ at least two sequences find their common ancestor with probability $P_{\mathrm{c}}(n,\ell+1)$. Note that, under the assumption $N \gg 1$, the probability that more than two sequences find their MRCA in a single generation is negligible and can be ignored. In this case, thus, $P_{\mathrm{c}}(n,\ell+1)$ stands for the probability that a pair of sequences, among the $n$ sequences sampled, find their MRCA in generation $\ell+1$ back in time. An event in which two sequences find their MRCA is called a coalescent event.

From Eq. (2.8) it follows that the number of generations to the first coalescent event in a sample of $n$ alleles ($\tau_n$) is approximately exponentially distributed with mean:

$$\langle \tau_n \rangle = \frac{N}{\binom{n}{2}} \ . \tag{2.9}$$

Thus, the average number of generations for obtaining a coalescent event between any pair of lineages scales linearly with the population size,

and it is inversely proportional to the total number of possible pairs of the lineages in question (that is, $\binom{n}{2}$), each pair being equally likely to coalesce. Noting that each coalescent event reduces the number of ancestral lines to be traced back by one (Fig. 2.1), the time to the MRCA of the entire sample is given by $\sum_{i=2}^{n} \tau_i$, where $\tau_i$ $(i = 2, \ldots, n)$ are independent random variables, distributed approximately exponentially with mean $N/\binom{i}{2}$ [35]. The total branch length $T_n$ of a gene genealogy of sample size $n$ satisfies $T_n = \sum_{i=2}^{n} i\tau_i$.

The coalescent process provides a method for generating an ensemble of gene genealogies of sample size $n$ much more efficiently than by tracing the ancestry generation by generation. When a gene genealogy is obtained, neutral mutations may be superimposed on it, and hence the patterns of neutral genetic variation are fully described by the coalescent process [35]. The number of mutations along a branch of length $\tau_i$ is Poisson distributed with mean $\theta\tau_i/2$, where $\theta = 2\mu N$, and $\mu \ll 1$ is the mutation probability per generation, allele, individual.

Apart from being efficient, the coalescent process is also robust, and difficult to reject. It can be proven that the standard coalescent is not only valid for the Wright-Fisher model, but also for many other population models, provided that the variance $\sigma^2$ of the reproductive success between individuals remains finite in the limit of $N \to \infty$ (that is, $\sigma^2/N \to 0$ in the limit of $N \to \infty$) [34]. An example is the Moran model introduced in Section 2.1. For this model it can be shown [34] that the coalescent method is applicable, but with a factor $N/2$ in Eq. (2.9) instead of $N$ as in the Wright-Fisher model.

Although the coalescent process is built upon assumming that the population size remains constant over time, in some cases it can be applied to fluctuating population sizes upon defining a corresponding effective population size $N_\mathrm{e}$ [41, 42, 48]. In Ref. [41] it was shown that the effective population-size approximation is applicable for the cases of both slow and rapid population-size fluctuations (in relation to the coalescent timescale). In the former case, the effective population size is approximately equal to the population size at the present time. In the latter case it is equal to the harmonic mean of temporal population sizes $N_\ell$ [41]

$$N_\mathrm{e} = \left( \lim_{\mathcal{L}\to\infty} \frac{1}{\mathcal{L}} \sum_{\ell=0}^{\mathcal{L}-1} \frac{1}{N_\ell} \right)^{-1} . \tag{2.10}$$

Here, $\mathcal{L}$ is the number of generations back in the past since the present time.

However, when population-size fluctuations are neither slow nor fast in comparison to the coalescent time scale, the result of the standard

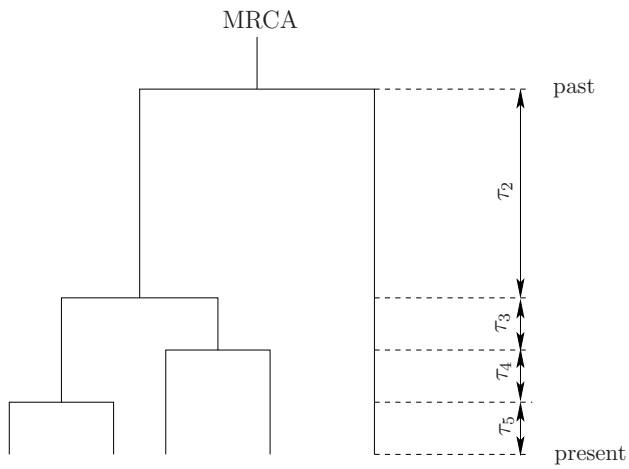


Figure 2.1: Gene genealogy of a sample of size $n = 5$ (illustration). The times during which the gene genealogy has exactly $i = 2, 3, 4, 5$ lines are denoted by $\tau_i$. The most recent common ancestor of the sample is denoted by MRCA. This figure is taken from the Licentiate thesis [3].

coalescent approximation that makes use of Eq. (2.10) may not be appropriate to describe typical gene genealogies [1, 41]. The results in Ref. [1] (their Eq. (20)) allow for computing moments of the total branch length $\langle T_n^k \rangle$ $(k = 1, 2, \ldots)$ of gene genealogies for populations of varying sizes. The approach outlined in Ref. [1] for computing the moments makes use of 'the population-size intensity function' [82], that accounts for temporal changes in the coalescent time scale due to population-size fluctuations. More details on the derivation of this result are given in Refs. [1, 3]. For $k = 1$, Eq. (20) in Ref. [1] agrees with the result in Ref. [83]. The expression for the second moment in Ref. [1] is in agreement with the corresponding result in Ref. [84].

Finally, note that apart from the standard coalescent, there are also other types of coalescents, such as Xi-coalescents [85–88]. Under a Xi-coalescent, multiple ancestral lines are allowed to merge in a single ancestor in a given generation. This type of Xi-coalescents is also known as the Lambda-coalescent [86]. In a more general case, a Xi-coalescent allows for simultaneous multiple mergers in a given generation.

A Xi-coalescent is obtained under models allowing for skewed offspring distribution among individuals in a population [89], in models that account for selective sweeps [88], as well as in models of populations that undergo recurrent bottlenecks in their histories [2, 90].

In the next section, modelling the processes of mutation and recombination are discussed.

## 2.3 Mutation and recombination

Mutations alter the sequence of nucleotides at a given locus, and hence contribute to increasing genetic variation at this locus. The nucleotide sequence can be changed by mutations in several different ways. One possibility is that mutations induce a change of one or more nucleotides in the sequence. Other possibilities include nucleotide rearrangements within sequences, such as inversions, or translocations [21]. Mutations may also shorten or extend the sequences (deletions and insertions). This section discusses commonly used models for neutral mutations. Models including natural selection are covered in Section 2.4.

When modelling neutral mutations, one uses the so-called infinite-alleles model [91]. Under this model, each mutation gives rise to a new type of an allele. This model is appropriate when empirical data provide only the information whether a diploid individual has the same alleles at a given locus (*homozygote*), or it has different alleles at the locus (*heterozygote*). Data of this kind are, for example, amplified fragment length polymorphisms (AFLP) [92].

Another used model is the infinite-sites model [93]. In this model one treats loci as infinitely long sequences of nucleotides (i. e. sites). It is assumed that each mutation occurs at a new site, causing single nucleotide polymorphisms (SNPs). Under this model, thus, exactly two different nucleotides appear at each polymorphic site. This model is appropriate for 'complex' species that have long genomes, e. g. Humans [43].

Yet another used model is the stepwise-mutation model [94–96]. In this model an allele is defined by the number of repeated sequences of base pairs it contains, and it is assumed that a mutation occurring at a given locus may either decrease or increase the number of repeated sequences by one [94] (for example, due to deletions, or insertions). The stepwise-mutation model is commonly used to describe genetic variation at microsatellite loci[3].

In this thesis mutations are modelled according to either the infinite-alleles or the infinite-sites model. It is assumed that mutations accumulate along a given locus with the probability $\mu$ per generation, sequence, individual. For simplicity, the probability $\mu$ is further assumed to be constant over time.

[3]Microsatellite loci contain repeated sequences of two to five base pairs. Alleles at a given microsatellite locus mainly differ by the number of repeated sequences [96].

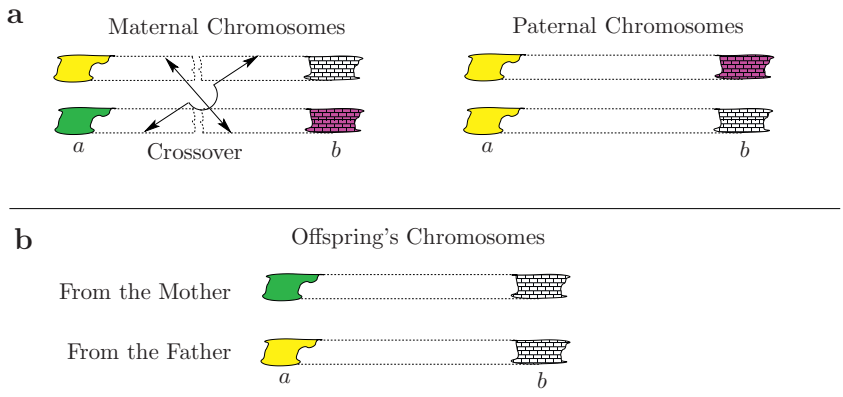


Figure 2.2: Recombination due to crossover of maternal chromosomes (schematically). (**a**) Two maternal and two paternal chromosomes. Left coloured areas of the chromosomes depict allelic types at locus $a$, and right coloured parts depict allelic types at locus $b$. Two maternal chromosomes split in two parts. Each part stemming from one copy of maternal chromosomes attaches to the opposite part of the other copy of maternal chromosomes (depicted by arrows). As a result, the combination of allelic types at the two loci in the offspring is different from that in either of its parents (panel **b**). This figure is a modified version of Fig. 2.1 in the Licentiate thesis [3].

Apart from the process of mutation, in diploid organisms it is further necessary to account for the process of genetic recombination as a source of multi-locus genetic variation. Indeed, recombination re-arranges pairs of maternal or paternal chromosomal sequences (chromosomal crossover, Fig. 2.2), and thus contributes to multi-locus genetic variation. As a consequence, an offspring typically inherits neither a complete ('unbroken') set of chromosomes from the mother, nor a complete set of chromosomes from the father [21, 79]. Apart from crossover, there are other types of recombination such as gene conversion [21]. In this thesis recombination is assumed to occur due to chromosomal crossover.

Empirical data show that the probability that a pair of loci on the same chromosome recombines is larger when the two loci are farther apart [97]. But the recombination rate is known to be inhomogeneous along the Human genome [98]. It is common to express the physical distance between two loci in terms of the probability $r$ that a chromosome recombines between the two loci per generation, chromosome, individual. In this thesis, it is assumed that $r$ is constant over time.

Note that in well mixed populations, assuming random mating and Mendelian inheritance, the association of neutral genetic variation between a pair of loci situated at different chromosomes is random. Such loci are said to be in linkage equilibrium. Otherwise, loci are said to be in linkage disequilibrium [98–102].

Finally, recall that up to now, models of neutral genetic variation were outlined. The next section discusses basic models of natural selection.

## 2.4 Selection

In the previous sections, models and sources of neutral genetic variation were discussed. However, the effect of natural selection inevitably influences the patterns of genome-wide genetic variation. Darwin proposed that the survival (viability) and/or reproductive success (fecundity) of an individual depend on the environment that the individual is exposed to: the individual may be more or less 'fit' [18]. Less fit individuals reproduce less successfully than better fit individuals. In other words, less fit individuals are 'selected' against in a given environment, and their frequency is expected to progressively decrease in the population ('survival of the fittest') [18]. This is the basic idea behind the process of natural selection.

But what determines how individuals perform in a given environment? Depending on the environment, specific biological traits, such as resistance or susceptibility to a certain disease, size, tolerance to high, or low salinity, and similar, may be particularly important for survival, or fecundity of individuals. For example, in the populations of the sea snail *Littorina saxatilis*, individuals that are larger, have thicker shells, and smaller feet survive better in crab-exposed than in wave-exposed environments [36]. The opposite is true for the smaller individuals, with thinner shells, and larger feet: these are better equipped to withstand frequent waves, than crab attacks. The two types of individuals are said to belong to 'divergent ecotypes' of *L. saxatilis*, that are a result of natural selection acting in opposing directions in wave- and crab-exposed environments [36]. Note that the characteristics of biological traits that are relevant for the fitness of an individual are commonly referred to as the phenotype. Consequently, it is commonly stated that natural selection acts on phenotypes.

However, a phenotype under the given environment is determined by the genotype of the individual in question at specific loci on the genome [6–10, 25, 77, 78]. These loci may be exposed to weaker or stronger natural selection. Alternatively, the loci may exhibit 'plasticity' that allows

the individuals with the same genotype to exhibit different phenotypes in response to different environments [103]. The effect of plasticity on the capacity of individuals to adapt is still poorly understood [7], and it is beyond the scope of the present thesis. In what follows, two well-known models of natural selection are outlined.

In a simple model of a population subject to natural selection it is assumed that selection acts on a single locus. The population is further assumed to be diploid, well mixed, and randomly mating. At the locus targeted by selection, the population is assumed to have two alleles, denoted by $A_1$, and $A_2$ below. One of the two alleles (say, $A_2$) is advantageous in comparison to the other ($A_1$). The fitness of the different genotypes, each being determined by a pair of alleles at a given locus, is assumed to be as follows: the fitness of the homozygote $A_1|A_1$ is equal to unity, the fitness of the heterozygote $A_1|A_2$ is equal to $1+s$, and the fitness of the homozygote $A_2|A_2$ is equal to $1+2s$. Here, $s$ is a selection coefficient that determines the selection strength for the beneficial allele $A_2$, and selection is assumed to be additive. The population size is further assumed to be constant over time, and the number of offspring reproduced by a given individual is directly proportional to the ratio of the fitness of the individual over the average fitness of all individuals in the population. Under these assumptions it can be shown under a deterministic approximation that the stable steady state of the system initialised with alleles $A_1$ and $A_2$ corresponds to the fixation of allele $A_2$ (and extinction of allele $A_1$) [25]. However, in finite populations the effect of random genetic drift needs to be taken into account. Due to random fluctuations, the advantageous allele $A_2$ can experience extinction by chance, especially if its initial frequency in the population is low. The fixation probability of the advantageous allele $A_2$ that is introduced in the population of size $N$ at a frequency $p_0$ can be approximated by [77, 78, 104]

$$p_{\text{fix}}(p_0) \approx \frac{1-\mathrm{e}^{-4sp_0N}}{1-\mathrm{e}^{-4sN}} \,. \tag{2.11}$$

Here it is assumed that $N$ is large. When only one advantageous advantageous allele that is weakly selected for is introduced in the population ($p_0 = (2N)^{-1}$), and $Ns \gg 1$, then Eq. (2.11) reduces to $p_{\text{fix}} \approx 2s$. The latter expression was derived in Ref. [105]. The fixation of the advantageous allele at the locus under selection is referred to as *selective sweep* [106]. For populations of large size ($N \gg 1$), and when $Ns \gg 1$, the average number of generations $\tau_{\text{sweep}}$ needed for the sweep to occur when $p_0 = (2N)^{-1}$ can be approximated by [107, 108]

$$\tau_{\text{sweep}} \approx 2\frac{\ln(4Ns)}{s} \,. \tag{2.12}$$

As Eq. (2.12) shows, the duration of the sweep is longer in larger than in smaller populations, all else being the same. Conversely, the duration of the sweep is shorter when selection for the advantageous allele is stronger. Note that genetic variation at a neutral locus closely linked to the locus that experiences selective sweep is expected to be reduced due to the sweep [109]. Namely, an allele at the neutral locus that is associated to the advantageous allele $A_2$ establishes more offspring in relation to other alleles at the neutral locus that are associated to the allele $A_1$. This effect is referred to as *hitchhiking* [109]. The effect of hitchhiking is stronger the closer the neutral locus to the selected one is [107–109].

Another well-known model of natural selection is Fisher's geometric model of adaptation in a well-mixed random mating population subject to a given fixed environment [11]. This model was analysed in great detail in Refs. [66, 67]. In the model individuals are subject to selection that acts on many (say, $\eta$) traits. The traits are visualised as mutually orthogonal axes in an $\eta$-dimensional space. The environment is assumed to be such that it is optimal for a particular point $\boldsymbol{\Theta}$ in this space. Here, $\boldsymbol{\Theta}$ is an $\eta$-dimensional vector, and it is referred to as *optimal phenotype*. The question is: how do the individuals of the population reach this optimal phenotype, and hence *adapt* to the given environment? The individuals that do not have the optimal phenotype may adapt due to, for example, mutations that make changes to individuals' phenotypes [66, 67], or due to recombination that may form new genotypes more or less fit than the genotypes present in the population [7, 10, 17, 75, 110]. Alternatively, 'standing genetic variation' may contribute to adaptation when environmental conditions exhibit temporal changes [111]. Namely, some loci may be neutral under particular environmental conditions, and hence accumulate standing genetic variation. When the environment changes, and these loci become targets of selection, their standing genetic variation can facilitate adaptation by increasing the chance that particularly beneficial alleles are present in the population. These beneficial alleles then may increase in frequency due to selection [111]. This thesis is mostly concerned with the former sources of adaptation. The importance of standing genetic variation for adaptation is briefly discussed in Chapter 6.

In Refs. [66, 67], a mutation is assumed to either increase or decrease the phenotype of the individual in question by an amount that is referred to as the mutation-effect size. The mutation can either move the phenotype towards the optimum or away from it. In the former case, thus, the mutations are deleterious, and they cannot persist and establish in the population due to the effect of natural selection. By contrast, beneficial mutations may establish in the population, but this depends on

the interplay between random sampling effects in populations of finite sizes (genetic drift) and natural selection. This is discussed next for the fitness function used in Ref. [67].

In Ref. [67], the fitness $w_{\mathrm{r}}$ of a resident with the phenotype denoted by $\mathbf{z}_{\mathrm{r}}$ below is assumed to be given by

$$w_{\mathrm{r}} = \mathrm{e}^{-\frac{(\mathbf{z}_{\mathrm{r}}-\boldsymbol{\Theta})^2}{2\sigma^2}} \ . \tag{2.13}$$

Here, $\sigma$ is a parameter that determines the strength of selection towards the optimum. Selection is stronger when $\sigma$ is smaller, and vice versa. To estimate the fixation probability of a beneficial mutation under this model, one can make use of Eq. (2.11). Indeed, assuming that the resident individuals in the population have the phenotype $\mathbf{z}_{\mathrm{r}}$, and that, upon fixation of the mutation, the individuals have the phenotype $\mathbf{z}_{\mathrm{m}}$, the selection strength $2s$ for the mutant individuals can be estimated using [67, 104]

$$2s = \frac{w_{\mathrm{m}}}{w_{\mathrm{r}}} - 1 \ . \tag{2.14}$$

Note that the factor two on the left hand side of Eq. (2.14) appears because here it is assumed that the population is diploid. By contrast, in Refs. [67, 104] the population was assumed to be haploid, and hence the selection strength $s$ in Refs. [67, 104] is two times larger than that given by Eq. (2.14). The results obtained in Ref. [67] are briefly discussed next.

It was found in Ref. [67] that the distribution of the mutation-effect sizes fixed in the course of adaptation is approximately exponential provided $\eta$ is large enough (i. e. $\eta \geq 5$). This conclusion does not depend on the probability distribution from which the mutation-effect sizes are drawn in the course of adaptation [67]. The same is true for alternative fitness functions that differ from that in Eq. (2.13). Furthermore, the conclusion also holds independently of the initial distance of the average phenotype of the individuals from the optimum, and on when during adaptation the statistics of the factors fixed is made [67]. These findings fully describe the effect of selection and genetic drift in a freely-mixing population subject to a given (fixed) environment. However, no such general results are available for geographically structured populations that are subject to different environmental conditions along their distributions, and that are not completely isolated from each other.

In summary, this chapter presented a number of existing theoretical results that concern the effect of different evolutionary processes on genetic variation. They provide a basis for interpreting empirical genome-wide genetic patterns. But because precise life histories of individuals are in general unknown, interpretations of empirical data inevitably depend

on particular assumptions concerning the underlying population-size history, and population structure. As a consequence, different assumptions may give rise to different conclusions concerning the historical evolutionary events [112]. Furthermore, genetic variation in natural populations is inevitably influenced by stochastic fluctuations (random sampling due to finite population sizes, then mutations, recombination, and possibly migration in geographically structured populations occur with a given probability etc.). The effect of different population-size histories, and population structures in the presence of stochastic fluctuations can, however, be tested using models. This thesis first discusses the effect of population-size fluctuations on site frequency spectra of mutations under piecewise constant demographies (Chapter 3, paper [I]). The effect of a population structure in the presence of different levels of multiple paternity is discussed in Chapter 4 (paper [II]). Next, Chapter 5 analyses the effect of a population structure but in the presence of mixed sexual and asexual reproduction. Finally, Chapter 6 (unpublished results) discusses the effect of migration, selection and drift.

# 3

# Frequency spectra of SNPs under varying population sizes

As explained in the introduction, genome-wide patterns of genetic variation are shaped by a joint effect of random genetic drift, demographic history, and natural selection. In the previous chapter it was described how each of these processes individually influences genetic variation. However, the results outlined in the previous chapter are based on the assumption that the demographic history of the population in question, as well as the selection strength are known, whereas this is not true in reality [7, 10, 36, 104]. While the demography is expected to influence the whole genome in a similar manner [27], the effect of selection is likely to be different genome-wide. The latter is because the strength of selection may differ between different regions targeted by selection [28, 84], as well as because the effect of selection on closely linked neutral regions can differ due to genome-wide inhomogeneity of the recombination rate [98]. Thus, interpreting genome-wide empirical data is challenging [112]. A number of past and present advances concerning this task are discussed in this chapter.

In the past, many statistical tests of neutrality of genome regions were proposed. Examples include Tajima's $D$ [32], Fay & Wu's $H$ [20], and others. These tests are based on comparing empirical frequency distributions of SNPs to those expected under the null model that assumes that mutations are neutral, and that the population size is constant. The frequency distributions of SNPs are commonly referred to as *site frequency spectra* (SFS). The site frequency spectra can be either unfolded or folded. To obtain unfolded SFS it is necessary to know the ancestral nucleotide at the site in question. If this information is available, then

the number of individuals ($i$) that carry a mutation at this site ranges from $i = 1$ to $i = n - 1$ where $n$ denotes the sample size. The number of mutations appearing in $i$ individuals is denoted by $\xi_i$. The unfolded SFS is given by the counts $\xi_i$ ($i = 1, \ldots, n - 1$). When empirical data do not provide an information about the ancestral nucleotide at a given site, one uses folded SFS. The counts of the folded SFS (denoted by $\eta_i$ below) satisfy [113]

$$\eta_i = \xi_i + \xi_{n-i}\,, \text{ where } i = 1, \ldots \lceil \frac{n-1}{2} \rceil\,. \tag{3.1}$$

In Eq. (3.1), $\lceil x \rceil$ denotes the smallest integer not less than $x$. The moments of the site frequency spectra of neutral mutations (Appendix A) can be expressed in terms of the branch lengths of gene genealogies of a given sample, and they were derived in Ref. [114].

Tests of neutrality mentioned above make use of the fact that the SFS of a genome region under selection and its neighbourhood differ from the SFS expected for neutral regions (not linked to the selected ones). However the difference between the former and the latter SFS depends on the strength of selection, as well as whether sampling is done during or after a selective sweep [112, 115]. If the locus under selection or a closely linked neutral region are sampled during a selective sweep, it is expected that mutations of intermediate frequency appear in excess [20, 112]. However immediately after a selective sweep, the locus under selection and its closely linked neutral regions show an abundance of mutations of high frequency [20]. Furthermore, the effect of hitchhiking on neutral loci decreases with increasing the time after the selective sweep because new mutations at the neutral locus accumulate after the sweep. Thus signatures of selection can be visible in different parts of the SFS (typically high or intermediate mutation counts). For this reason different tests assign different weights to the spectrum counts $\xi_i$ [116]. For example, Tajima's $D$ is designed to capture an excess (or a deficit) of mutations of intermediate frequency. This test is powerful against relatively recent selective sweeps that are caused by strong selection [115]. By contrast, Fay & Wu's $H$ is sensitive to an excess of mutations of high frequency, and it can be more powerful than Tajima's $D$ to detect very recent hitchhiking events [20].

However recall that the null-model in these tests of neutrality is not only based on the assumption that all mutations are neutral, but also that the past population size was constant over time. As a consequence, deviations from 'neutrality' detected by these tests can be due to selection, as well as due to population-size fluctuations [20, 115–117]. Indeed, the SFS of neutral mutations obtained under population-size fluctuations, as well

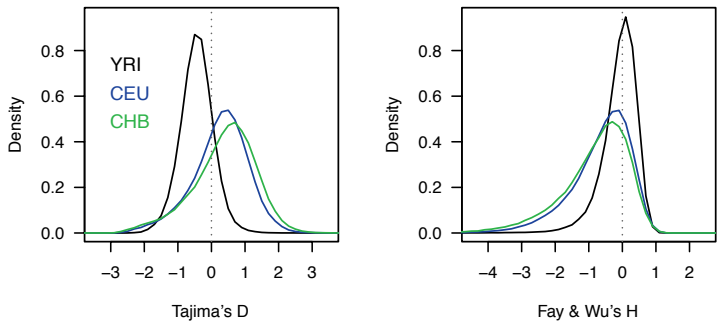


Figure 3.1: Distribution of test values over all sliding windows (window size $10^5$ base pairs, sliding step size $10^4$ base pairs) across the genomes of the Human populations: YRI (black), CEU (blue), CHB (green). Left: Tajima's $D$. Right: Fay & Wu's $H$. The empirical data are taken from Ref. [43]. This figure shows the first two upper panels of Fig. 5 in Ref. [I]. For further details, see Ref. [I].

as the corresponding null-distributions of the tests may differ substantially from those expected for populations of constant size, e. g. due to recent population-size expansions or bottlenecks [28, 113]. However, a demographic history is expected to have an impact on all genome regions, whereas the effect of selection is more or less local on a genome-wide scale. This motivated many researchers to use the quantiles of empirical genome-wide distributions to detect deviations from neutrality [29–31, 118, 119]. As an example, Fig. 3.1 shows empirical genome-wide test distributions obtained by scanning the genomes of three different Human populations [I]. As this figure shows, the empirical test distributions differ substantially between the different populations. The question arises: since the empirical distributions corresponding to different populations have different shapes, how can one compare the extents of selection on candidate regions between the different populations? In order to answer this question, the underlying demographies must be estimated. Next, the estimated demographies need to be integrated into SFS-based tests of neutrality to obtain *demography-adjusted tests* [I]. A method for estimating Human demographies is discussed next. Details concerning how demography-adjusted tests are obtained are given in Ref. [I].

The demographic history of the population in question can be inferred by applying a maximum-likelihood method to empirical spectra of intergenic, physically distant SNPs [28, 120, 121]. The former is required because intergenic SNPs are expected to be neutral [28, 120, 121],

but this expectation has been challenged [122]. The latter requirement serves to simplify the analysis because distant SNPs are approximately uncorrelated. For uncorrelated SNPs, the counts $\xi_1, \xi_2, \ldots$ are multinomially distributed. Apart from these two requirements concerning empirical data, it is further necessary to make assumptions on the underlying model of population-size histories. This facilitates the maximum-likelihood approach. The method applied in Ref. [I] is based on a piecewise constant population-size model in which the population size is assumed to have experienced at most two sudden changes in the past. This model has been suggested [28, 29, 121] as an approximation to the main events of the Human out-of-Africa expansion [37–39, 123]. The maximum-likelihood procedure for inferring the parameters of this demographic model requires the first moment of the site frequency spectrum $\langle \xi_i \rangle$ (or $\langle \eta_i \rangle$) to be computed. This can be done using simulations, but the procedure is facilitated if analytical expressions are available. As shown in Ref. [I], the analytical expression for the first moment of the site frequency spectrum under the demographic model assumed can be obtained by combining the results given in Ref. [1] to those given in Ref. [114] (Appendix A). To avoid possible miss-specifications of the ancestral sequences, the demography estimation in Ref. [I] was based on folded SFS. After obtaining the analytical expression for $\langle \eta_i \rangle$, the maximum-likelihood demography is estimated as follows. A wide range of candidate demographies (with different values of model parameters) are chosen to be tested. For each demography, the expected spectrum counts $\langle \eta_i \rangle$ ($i = 1, \ldots, n-1$) in a sample of size $n$ are computed. Next, for each demography, the probability to observe empirically obtained counts $\eta_i$ under the demography is computed. Finally, the maximum-likelihood demography is estimated by finding the demography with the highest likelihood among the candidate demographies tested.

Note that empirical spectra are inevitably influenced by stochastic fluctuations. The effect of fluctuations is reduced by using a large number of SNPs as an input for demography estimation. To estimate the number of SNPs needed for the demography estimation to be reliable, the maximum-likelihood procedure described above was tested in Ref. [I] against simulated data under two reference demographies. The maximum-likelihood estimation was found to perform well when the estimation was based on at least $10^5$ independent SNPs [I]. By contrast, the estimation performed poorly when the number of SNPs was set to $10^4$ [I]. In Ref. [I], this analysis of the performance of demography estimation served to guide the sampling of empirical data (gathered in the 1000 Genomes Project [43]) that were then used for the estimation of the demographies of ten Human populations.

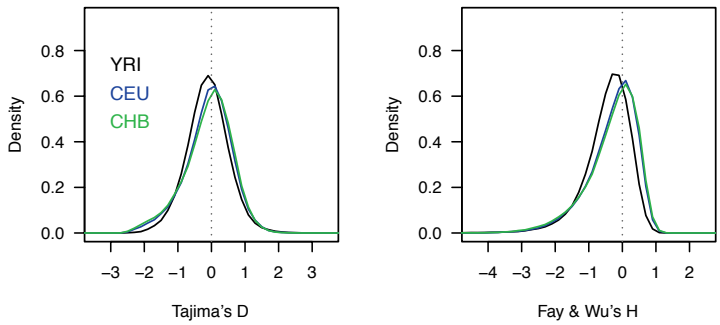


Figure 3.2: Distribution of demography-adjusted test values over all sliding windows (window size $10^5$ base pairs, sliding step size $10^4$ base pairs) across the genomes of the Human populations: YRI (black), CEU (blue), CHB (green). Left: Tajima's $D$. Right: Fay & Wu's $H$. The empirical data are taken from Ref. [43]. This figure shows the first two lower panels of Fig. 5 in Ref. [I]. For further details, see Ref. [I].

The Human demographies estimated in Ref. [I] are consistent with the demographies estimated in Refs. [28, 121]. The site frequency spectra of non-African populations under the model assumed are consistent with a population bottleneck, whereas the maximum-likelihood demography of the African population ASW corresponds to two past population-size expansions. The maximum-likelihood demographies of the two other African populations sampled (YRI, and LWK) correspond to a population-size expansion followed by a very recent population-size decline.

The demographies estimated allow for demography-adjusted tests to be constructed and applied to genome scans [I]. Recall that demography estimation is based only on intergenic, physically distant SNPs, each pair of SNPs being at least $5 \cdot 10^4$ base pairs separated. By contrast, the distributions of demography-adjusted and unadjusted tests are obtained by computing test values along continuous sliding windows containing $10^5$ base pairs (sliding distance $10^4$ base pairs). The sliding-window approach was suggested in Ref. [30]. The results reported in Ref. [I] show that, unlike the distributions of the demography-unadjusted tests, the distributions of demography-adjusted tests are similar between the Human populations sampled (compare Fig. 3.2 to Fig. 3.1). Thus a comparison of the extents of selection at the candidate regions between different populations is facilitated when using demography-adjusted tests of neutrality.

In Ref. [I] it was further found that the unadjusted empirical test

values are, apart from some deviations, roughly linearly related to the adjusted empirical test values. As a consequence, both demography-adjusted and unadjusted tests detect the same candidate regions under selection [I]. A linear relationship was also found between demography-adjusted and unadjusted test values obtained for simulated SFS of neutral mutations under a given null demography [I]. However, in the simulations, as well as in the derivation of both adjusted and unadjusted tests, the effect of recombination on genetic sequences is neglected [20, 30, 32, 115, 116]. By contrast, the effect of recombination cannot be neglected along empirical genome-wide sequences. As suggested in Ref. [32], recombination is expected to shrink the theoretically expected distributions. However, the effect of recombination can differ between adjusted and unadjusted tests, and thus possibly distort the linearity observed in computer simulations. The extent of distortion is likely to depend on the recombination rate at a given genome region. As mentioned above, empirical test distributions revealed that a small amount of regions deviate from a linear relationship between the adjusted and unadjusted test values. These deviations can be caused by a joint effect of selection and recombination along these regions. In order to understand how recombination alters demography-adjusted and unadjusted tests, further computer simulations that incorporate the effect of recombination must be made.

The demography-adjusted tests described above are based on demographies estimated using empirical spectra. In Ref. [124] it was argued that the exact underlying demography cannot be estimated using site frequency spectra, because substantially different demographies can give rise to exactly the same spectra. However when demography estimation is constraint to a simple model, such as the one used here, the parameter values of the maximum-likelihood demography are sufficiently close to the corresponding parameter values of the true underlying demography, provided that the estimation is based on a large number of SNPs.

In summary, both unadjusted and adjusted test distributions detect the same candidate regions under selection. However empirical distributions of demography-adjusted tests facilitate the comparison of the extents of selection on candidate regions between different populations, whereas such a comparison is difficult to make using unadjusted tests. Still, the results outlined above are based on a number of simplifying assumptions. Firstly, intergenic regions may not be neutral [122]. Secondly, candidate demographies were constrained to a simple demographic model. Finally, the effect of recombination is neglected in test definitions. It remains to be understood how these assumptions influence the results presented.

# 4

# Multiple paternity in geographically structured populations

In the previous chapter it was discussed how the patterns of neutral genetic variation in populations of varying size differ from the patterns in populations of constant size. The demographic history of Humans was approximated by varying population sizes that account for past sudden population-size changes. In other words, in the approximation made, the exact geographic structure with migration between the individual populations after their establishment, as well as the process of establishment of a given population after the first founder event were neglected. However natural populations are geographically structured, and many are currently ongoing colonisations of new habitats thanks to the process of migration. Furthermore, natural populations can be subject to frequent extinctions in particular local geographic areas. In this case, the populations that persist in near-by habitats (so-called refuge areas) may give founders to the locally extinct areas, and hence re-establish populations. Whether or not this happens depends on the movement and dispersal capabilities of the population (in relation to the underlying geographic structure of the habitat), as well as on the capacity for population growth starting from the first founders in a given area. For the latter, the amount of genetic variation that the founders bring into new areas can be critical, especially if areas are subject to unstable environmental conditions. If the population colonises the habitat in a stepwise fashion from a large source area, the genetic variation of the founders can decrease as the distance from the source refuge population increases (repeated founder

events) [125]. This is particularly true when considering genetic variation in recently colonised areas, as new genetic variants arising due to typically rare mutations can be neglected on short timescales [125]. But it has been suggested in Ref. [126] that this tendency may be counteracted by a particular mating pattern - multiple paternity [15, 127]. Therefore, it needs to be understood: how does multiple paternity influence genetic variation? Does the effect of multiple paternity differ during and after the establishment of populations in geographically structured habitats? To answer this question, a mainland-island colonisation model was analysed in Ref. [II]. The results obtained in Ref. [II] are presented in this chapter.

An example species for which the model analysed in Ref. [II] is relevant is the marine snail *Littorina saxatilis*. This snail has been (and is) colonising new areas [126], and offspring of single females stem from multiple males [15, 127]. *L. saxatilis* spread from refuge areas to the northern Atlantic after the last ice-age, $12000-15000$ years ago [15]. Due to the retreat of the ice-cap, many islands became available as potential habitats, and new geographic areas over which this species can spread continue to emerge due to the uplift of land [15, 126, II]. In the archipelago along the Swedish west coast, that comprises of *mainlands*, *islands* and *skerries* (ordered by a decreasing size and age, see Fig. 4.1), the mainlands are likely to be the source populations from which younger islands and skerries were populated [126, 128]. Furthermore, since *L. saxatilis* has limited movement capabilities, and its individuals are unlikely to survive in the water, the colonisation of new habitats probably occurred (and is still occurring) by rafting in a stepwise fashion from the mainlands as the oldest source populations [126, 128]. The colonisation of new areas is relatively infrequent [128]. But this species has a high capacity for population growth, as females carry up to a hundred or more offspring beneath their shells. Census population sizes on islands and skerries are found to range between $10^2$ and $10^3$ [128]. This suggests that one founder female can colonise an island or a skerry in a single generation. As noted above, *L. saxatilis* is also known for its extremely high level of multiple paternity. A brood of a single female is typically sired by $15-23$ males [15, 127]. Multiple paternity has also been found in a number of fish and invertebrate species, but in these species the number of sires per female brood is smaller, around six to ten [44–47].

To understand the importance of high levels of multiple paternity during and after colonisation of a geographically structured habitat, a model that allows for both low and high levels of multiple paternity in the population was analysed in Ref. [II]. The results in Ref. [II] concern the spatial and temporal dynamics of genetic variation at a single locus

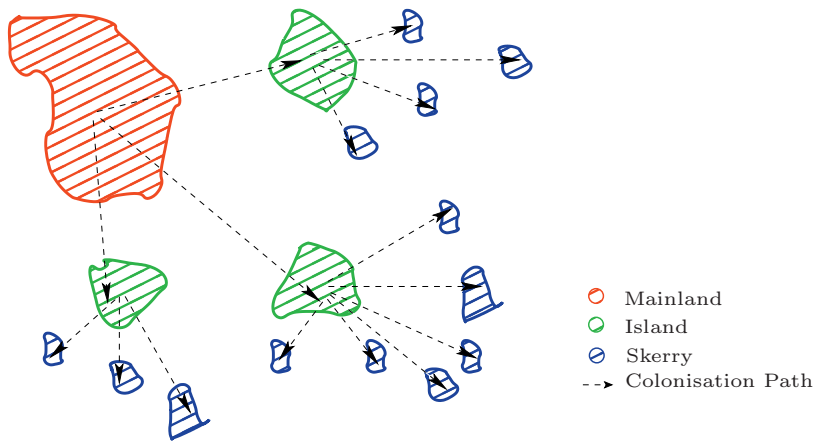


Figure 4.1: Illustration of the colonisation process of an archipelago by the marine snail *L. saxatilis* (schematically) in accordance with the results reported in Refs. [126, 128]. The mainland is coloured red, whereas islands, and skerries are coloured green, and blue, respectively. The size of islands is typically much smaller than the size of the mainland. The same is true when comparing skerries to islands. Furthermore, the mainland is the oldest habitat, whereas islands and skerries appeared subsequently by the uplift of land [126]. Arrows depict a possible colonisation path in agreement with the findings of Refs. [126, 128]. This figure is taken from Ref. [3].

that is assumed to be neutral. The model and the results obtained in Ref. [II] are discussed next.

Motivated by the life history of the snail *L. saxatilis* outlined above, the colonisation of the habitat is modelled as follows. The habitat is assumed to consist of a mainland, and of islands arranged linearly at increasing distances from the mainland. In most of simulations, the number of islands $k$ was set to $k = 10$. At the beginning of the process of colonisation, the mainland is assumed to be occupied, whereas the remaining islands are empty. The lifecycle of individuals (that are assumed to be diploid) occurs in the following order: mating, migration, reproduction, death of adult individuals. The generations are assumed to be discrete and non-overlapping. Since mating is assumed to occur before migration, the migration of males can be neglected. The process of mating occurs locally within the mainland, or a given island, and it is

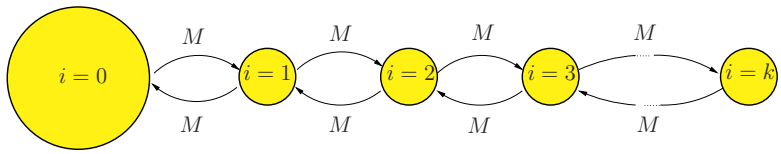


Figure 4.2: Spatial model for the colonisation of the mainland-island habitat. The mainland is denoted by $i = 0$. The islands are assumed to be empty in the beginning of the process of colonisation, whereas the mainland is populated. The population size of the mainland is assumed to be constant over time, and much larger than the population size of any island (after the island becomes populated). An empty island becomes populated upon the arrival of one or more founder females. An occupied island sends in each generation on average $M$ females to its right, and left closest neighbors (except for the last island, that only has the left neighbour). Likewise, the mainland sends on average $M$ females per generation to the first island. An empty island may only receive individuals. This figure is taken from Ref. [3].

discussed in more detail below. Migration is assumed to occur between the neighbouring islands (Fig. 4.2). Thus the model corresponds to a stepping-stone colonisation model [83, 129–131]. In each generation, an occupied island $i = 1, \ldots, k-1$ sends on average $M$ females to each of its two closest neighbours. Island $k$, and the mainland send on average $M$ females to their respective (single) neighbours. Finally, empty islands can only receive immigrants.

An empty island is assumed to be populated up to its carrying capacity (denoted by $2N$ below) immediately after it receives the first founder females. Namely, each female is assumed to carry a large number of offspring, and the founder females at the given island are assumed to produce $2N$ offspring, males and females being equally likely. Once an island is populated, its population size is assumed to remain constant over time. Furthermore, the population size of the mainland is assumed to be much larger than that of a populated island, and also constant in time. To simplify the analysis, the process of mutation is neglected in islands, whereas the mainland is the only source of genetic variation, and its heterozygosity on the locus of interest is assumed to be equal to unity. These assumptions simplify computer simulations, because the process of reproduction, and mutation do not need to be explicitly simulated for the mainland. More importantly, these assumptions do not affect qualitatively the main conclusions, as long as the scaled migration

rate ($M$) is not too low. Indeed, upon assuming that the per-individual per-generation per-allele mutation rate is the same for individuals on the mainland and on any island, the mutation rate scaled by the population size of an island is much smaller that the mutation rate scaled by the population size of the mainland. As a result, when $M$ is not too low, the process of migration from the mainland brings new genetic variants into islands more efficiently that the process of mutation.

Finally, mating (that occurs prior to migration) is modelled as follows. Each female is assumed to receive in total $l$ active sperm packages (thus, those that are not sterile). In the model, the parameter $l$ is equal for all females, and in all generations. The sperm packages are assumed to be of equal quality. Furthermore, each package is assumed to be able to persist for a long time after the female receives it, as well as to fertilise a large number of eggs. These assumptions are in agreement with empirical findings [126]. Assuming that the eggs are fertilised after all $l$ sperm packages are collected, and that each package persists until the end of the mating season of the female, the probability $p$ that two eggs are fertilised by the same sperm package is equal to $p = 1/l$. The model further assumes that during the mating season each female encounters $s$ males that are randomly chosen from the males present in the population. The parameter $s$ is assumed to be the same for all females, and in all generations. Furthermore, one male among the $s$ mating partners of a given female can be dominant over the others, either because he is a more successful mate, or a female exhibits a preference for this male (suggested in Ref. [15]). In the model it is assumed that the female mates with one male (chosen randomly among her $s$ partners) with the probability $\alpha \geq 1/s$. Conversely, any of the remaining $s - 1$ partners experiences the probability $(1 - \alpha)/(s - 1)$ to mate this female. The parameter $\alpha$ is assumed to be the same for all females, and in all generations. When $\alpha = 1/s$, all mating partners of a given female are on average equally successful mates. Otherwise, when $\alpha > 1/s$, one mate is dominant over the others. Under these assumptions, it is possible to determine how the parameters $l$, $s$, and $\alpha$ determine the effective population size of a well-mixed population. Recall that in the populated islands, the population size is assumed to be equal to $2N$, and that males and females are reproduced equally likely. The computation of the effective population size under the model is briefly described next.

Under the model assumptions, and assuming further that females are on average equally successful in producing offspring, it follows that two offspring come from the same female with the probability $1/N$. Furthermore, two offspring come from the same male with the probability $p_{\mathrm{m}}$

that is given by [II]

$$p_{\mathrm{m}} = \frac{1}{N}\kappa + (1 - \frac{1}{N})\frac{1}{N} \ . \tag{4.1}$$

The second term stands for the probability that two offspring do not share a mother (probability $1-1/N$), but they share a father (probability $1/N$). The first term in Eq. (4.1) stands for the probability that two offspring that share a mother (probability $1/N$), also share a father (probability $\kappa$), where

$$\kappa = p + (1-p)\Big(\alpha^2 + \frac{(1-\alpha)^2}{s-1}\Big) \ . \tag{4.2}$$

Here, the first term stands for the probability that two eggs are fertilised by the same sperm package ($p = 1/l$), and the second term stands for the probability that two eggs are fertilised by distinct sperms packages (probability $1-p$), but they both come from the same male. In the latter case, the male is either the dominant one (probability $\alpha^2$), or any mate of the remaining ones (probability $(1-\alpha)^2/(s-1)$). When $\alpha = 1/s$, the probability that two packages come from the same male is equal to $1/s$. The probability $\kappa$ directly determines the level of multiple paternity within the population [II]. The largest level of multiple paternity is achieved when $\kappa = 0$, and this is obtained under the commonly used random mating model (i. e. in the limit of $l \to \infty$, $\alpha = 1/s$, $s = N$, $N \gg 1$ [132]). Single-paternity (monogamous mating) is obtained when $\kappa = 1$. In Ref. [II], Eqs. (4.1)-(4.2) were used to show that the effective population size $N_{\mathrm{e}}$ in a large well-mixed population ($N \gg 1$) is given by

$$N_{\mathrm{e}} = 4\frac{N}{2+\kappa} \ . \tag{4.3}$$

Thus, when $\kappa = 0$ (largest level of multiple paternity), the effective population size is largest and equal to the census population size ($N_{\mathrm{e}} = 2N$). In this case, thus, the effective population size under the model presented reduces to Eq. (2.3). By contrast, when $\kappa > 0$ (thus, for smaller levels of multiple paternity) the effective population size is smaller than the census population size. In some special cases, the mating model presented reduces to other existing mating models (see the comparisons in Ref. [II]).

The mating model was tested against the empirical data on paternity obtained from females mated under controlled conditions in the laboratory (data taken from Ref. [16]), as well as from females sampled from wild populations (data taken from Ref. [15]). The empirical distributions of the number of offspring coming from a single male in a brood of a given

female were used to fit the model parameters $p$, $s$, and $\alpha$. The distributions obtained under the model for the best-fitted parameters resemble the corresponding empirical distributions [II]. Using a chi-square test, it was found that the model cannot be rejected. However, in comparison to the best-fitted theoretical distributions, the empirical distributions revealed an excess of males with single progeny. This may be because all sperm packages in the model are treated as on average equally successful in fertilising eggs, whereas in natural populations females can exhibit cryptic choice of sperm [15].

To understand the consequences of multiple paternity and the geographic structure of the population described above, the population heterozygosity needs to be determined, both on short and long timescales. The expected population heterozygosity $H_{\mathrm{c}}^{(i)}$ in island $i$ in the generation when this island is populated (colonisation-phase heterozygosity) is [II]

$$H_{\mathrm{c}}^{(i)} = P(0|i)H^{(0)} \ . \tag{4.4}$$

Here, $H^{(0)}$ is the mainland heterozygosity (that is set to unity in the simulations), and $P(0|i)$ is the probability that the most recent common ancestor of two alleles sampled randomly in island $i$ (in the generation when this island is populated) stems from an individual that is born on the mainland. In the case of $M \ll 1$ it was found that [II]

$$P(0|i) = \Big(1 - \frac{1}{8}(1+\kappa)\Big)^{i} \Big(\frac{2MN_{\mathrm{e}}}{2MN_{\mathrm{e}+1}}\Big)^{i-1} \ . \tag{4.5}$$

Similarly, the expected steady-state heterozygosity under the model can be computed using a system of recursion relations, as shown in Appendix S3 in Ref. [II].

As expected, both the colonisation-phase, and the steady-state heterozygosity for a given level of multiple paternity decrease as the distance from the mainland increases [II]. Furthermore, both the colonisation-phase, and the steady-state heterozygosity in any given island are promoted by increased levels of multiple paternity. The results further show that the ratio of the colonisation-phase heterozygosity for a given value of multiple paternity over the heterozygosity for a single paternity increases substantially as the distance from the mainland increases. The same is true for the steady-state heterozygosity. However, the ratio is much larger in the former than in the latter case. The increase of the steady-state heterozygosity due to multiple paternity arises due to the corresponding increase of the effective population size. By contrast, the increase in the initial colonisation phase is much larger and it does not arise due to the increase in the effective population size alone. Indeed,

from Eq. (4.5) it follows that the colonisation-phase heterozygosity in islands cannot be expressed in terms of $M$ (migration is the only source of genetic variation) and the effective population size $N_e$. Instead, there is an additional nontrivial dependence on $\kappa$ (the inverse of the degree of multiple paternity).

In addition to the above, simulations further show that temporal fluctuations of the heterozygosity during the steady state are substantial, and they are larger at larger distances from the mainland (Fig. 4.3**A**). In the island farthest from the mainland, the distribution of the heterozygosity is bimodal, as expected when the rate of income of new genetic material is small [133]. The population in this island experiences long periods of low heterozygosity values (lower than 0.1), and short periods of relatively high heterozygosity values (higher than 0.4). The analysis shows that with increasing the level of multiple paternity, the duration of the low-heterozygosity phase on average decreases, whereas the opposite is true for the high-heterozygosity phase (Fig. 4.3**B**). The latter effect is more pronounced than the former effect.

In summary, the results discussed above show that in a geographically structured population arranged in a stepping-stone fashion with a mainland as a source population, genetic variation decreases as the distance from the mainland increases. This is true for short timescales (colonisation phase) as well as for long timescales (steady state). By contrast, in stepping-stone models that do not include a mainland, the steady-state heterozygosity is the same for all islands [83, 129–131]. The latter is also true for mainland-island models where the distance between island populations is neglected (either all islands receive migrants from the mainland simultaneously, or each island is equally likely to receive migrants from the mainland) [134]. Furthermore, bursts of high genetic variation coming from the mainland are promoted by higher levels of multiple paternity. This effect is much more pronounced in the initial colonisation phase, than on the long run. The finding that temporal fluctuations of the heterozygosity at a single neutral locus are substantial under the model analysed suggests that similar fluctuations should be observed when analysing empirical genome-wide patterns of neutral genetic variation at a given time. These results suggest that multiple paternity may be an important survival strategy of species that spread to newly available geographic areas and have limited dispersal capabilities, especially if the new areas are subject to environments that differ from those in refuge areas. In this case, genetic variation that is neutral in refuge areas may be exposed to selection in new areas. Thus, genetic variation carried by the founders (the colonisation-phase heterozygosity) can be understood as standing genetic variation [111] (see Section 2.4).

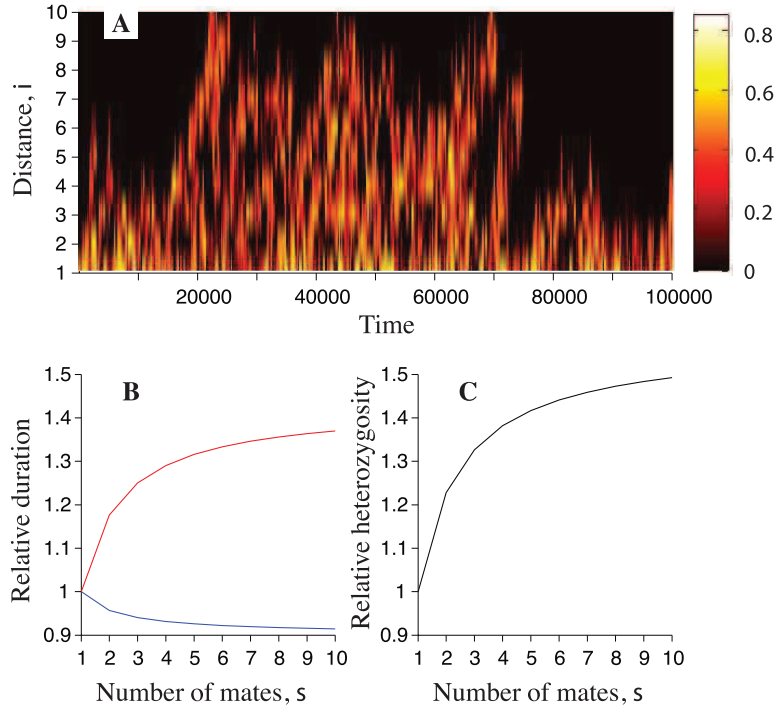


Figure 4.3: Temporal fluctuations of heterozygosity. **(A)** Heterozygosity as a function of distance from the mainland and of time (single realisation of the model described). Mainland is not shown. The data correspond to $10^5$ generations after the initial $7 \cdot 10^6$ generations. The number of available mates is $s = 10$. **(B)** Analytically computed durations of low-and high-heterozygosity phases (blue, and red) relative to their corresponding values for $s = 1$. **(C)** Analytically computed steady-state heterozygosity (black) relative to its corresponding value for $s = 1$. Remaining parameters: all available males are on average equally successful, $\alpha = 1/s$, the mainland heterozygosity is set to unity, the scaled female migration rate is $M = 0.05$, the number of females in each populated island is $N = 100$, the probability that two eggs are fertilised by the same sperm package is $p = 0.1$. This figure is taken from Ref. [II].

While the results outlined above suggest that multiple paternity may be an important strategy for species colonising new areas, such as *L.*

*saxatilis*, it must be emphasised that mating is also costly [15]. The empirical distributions of male-family sizes of wild populations showed that the best fit with the mating model was obtained upon assuming that no two eggs are fertilised by the same sperm package. This could be because each sperm package is able to ferilise only one egg (this constraint was not accounted for in the model). However, this could also be because females experience extremely large number of matings ($l \to \infty$). If this is the case, than multiple paternity could have also developed in this species because the cost of rejecting the mating could be higher than the cost of accepting it (as suggested earlier in Ref. [15]). The mating model presented could not be rejected, but the best-fitted model distributions largely disagreed with the empirical distributions in the region of small male-family sizes. The model showed that the probability that a father sires a single offspring is smaller than that observed in the laboratory, but also in the wild. This difference could arise because females exhibit cryptic choice of sperm. However, this difference could also arise because the approximation that the population is well-mixed is not necessarily supported in natural populations, especially with limited movement and dispersal capabilities. Indeed, a few females can be surrounded by many males, or vice versa, and the density of the population can differ between different parts of the habitat. Chapter 5 discusses a model that includes the exact spatial structure of sexes in populations with limited dispersal capabilities.

# 5

# Limited dispersal in populations with sexual and asexual reproduction

Chapters 3-4 summarised a number of advances in understanding genetic variation in natural populations under varying population sizes (Chapter 3), or during and after colonisation of geographically structured habitats. Recall that the results outlined in the previous two chapters are based on the following two assumptions: 1) the population reproduces sexually, and 2) individuals are freely mixing (in Chapter 4 the latter is applied to each habitat locally). However many biological species reproduce both clonally and sexually. Examples include macroalgae, a number of terrestrial and aquatic plants, as well as a number of animals [49, 135–137]. Furthermore, dispersal (or movement) capabilities in most species are limited, suggesting that individuals cannot mix freely [49, 128]. Therefore, it needs to be understood: how is genetic variation influenced upon relaxing one, or both assumptions given above? This chapter discusses a number of empirical and theoretical advances concerning this question. The discussion and results presented below were given in Ref. [III]. A closely related model was analysed in collaboration with Johan Fries (Master thesis [138]).

In species reproducing both asexually and sexually, one commonly finds that during invasions, or in recently inhabited areas, or at the edges of species' distributions, clones are more frequent than genetically unique individuals [50–56]. An example species supporting this finding is a recently established seaweed, *Fucus radicans* (see Fig. 5.1). To explain this dominance of clonal recruits over sexual recruits, many authors invoked

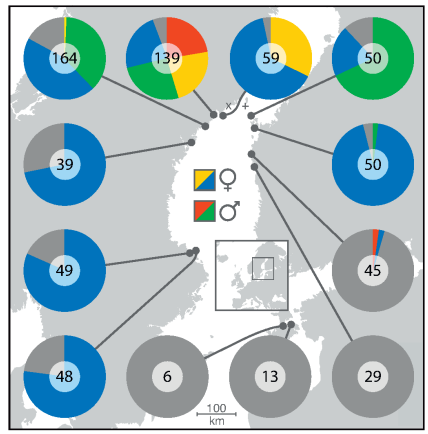


Figure 5.1: Distribution of *Fucus radicans* in the Baltic Sea, northern Europe. Grey shows unique individuals or local clones. Remaining colours denote clones that were found at more than one site. Each site is additionally labelled by the number of individuals sampled in it. The symbol "x" indicates a location where a free-floating individual was found. The symbol "+" represents a location where one attached individual of *Fucus radicans* was found (reported in Ref. [139]). The remaining data are taken from Ref. [140]. *Fucus radicans* is not found in established populations north of the sampled populations shown in this figure. This figure is taken from Ref. [III].

selection [53, 57–61]. The selection-based hypotheses suggested in these studies advocate the importance of clonal reproduction for conservation of particularly beneficial genotypes in the population [57–59], or of superior genotypes that originated from hybiridsation with closely related species [53, 60, 61]. However, the existing selection-based hypothesis lack empirical support [62, 63, 141]. Indeed, empirical studies of *F. radicans* designed to find signatures that the genotype of the highly dominant female (coloured blue in Fig. 5.1) is superior over the others were without success [62, 63]. This raises the question: can an alternative selection-free hypothesis explain the observed patterns in this and similar species?

An important difference between sexual, and asexual reproduction, that is typically neglected in studies of species with both modes of reproduction, is that sexual reproduction is possible only in the presence of both sperms and eggs. Therefore, in species with limited dispersal capabilities, solely inhomogeneous spatial sexual distribution of individuals

may result in facultative uniparental reproduction [64]. Consequently the spatial sexual distribution of a given species can have an important effect on the local (but also global) capacity of the species to reproduce sexually. It has been suggested [54, 55] that this effect may be particularly important during expansions of species over new geographic areas. However, this suggestion has not been tested, and hence empirical as well as theoretical studies commonly neglect it. In order to understand the importance of facultative uniparental reproduction, a model that includes a neutral, selection-free interplay between sexual and asexual recruits in species with limited dispersal capabilities must be studied.

A neutral spatially-explicit model with sexual and asexual reproduction was studied in Ref. [III]. The model, and the results given in Ref. [III] are discussed next.

The model is constructed to allow for analysing the process of colonisation of a habitat by a species that reproduces both clonally and sexually. In the model, the dispersal capability of propagules (gametes, and asexual fragments) are limited, and spatial distribution of sexes is explicitly accounted for, as described below. All individuals in the model are assumed to have the same characteristics, whether they are reproduced sexually or asexually. Firstly, the rate of production of "successful" sexual propagules is assumed to be the same for all individuals and in all generations. The same is true for "successful" asexual propagules. Here, "successful" propagules account for fertile gametes (that is, viable asexual propagules) that are able to successfully reach another patch. Secondly, the dispersal capability is assumed to be the same for all propagules, whether they are sexual or asexual. Thirdly, the survival probability to adulthood is the same for sexual and asexual recruits. Fourthly, the average lifetime of all individuals is the same. The details of the model are described below.

In the model, the habitat is assumed to consist of $N$ patches that are arranged linearly in a circle. Each patch is assumed to be either empty, or populated with one individual. Each individual is either a male or a female with a given genotype (ID). All individuals are assumed to produce and disperse locally both asexual and sexual propagules. Local dispersal of propagules occurs through the process of short-range dispersal. The short-range dispersal displacement for a given propagule is sampled randomly from a distribution with a parameter $\alpha \ll N$. The short-range dispersal distance is roughly proportional to $\alpha$. The distribution of propagules in a given patch determines the birth rates of sexual and asexual recruits in this patch (given in Ref. [III]). But at most one of the recruits in each patch can survive to adulthood (see below). The model further assumed that adult individuals die (after reproduction) at

a per-generation per-individual death rate $d$. Thus the generations are either overlapping (for $d < 1$), or non-overlapping (for $d = 1$). Next, in each empty patch random sampling is performed between the sexual and asexual recruits reproduced in this patch. The recruit sampled occupies the patch. By contrast, all recruits reproduced in a patch that is occupied with an adult die. After this step, the individuals in occupied patches are treated as adults constituting the next generation.

In addition to the above, the model further allows for occasional long-range dispersal of whole adult individuals (relocation). This process has been suggested to have important consequences on sexual and genetic patterns of populations [142]. Long-range dispersal is assumed to occur at a per-individual per-generation rate $l_{\mathrm{r}}$. This process is modelled using a symmetric power-law distribution with a parameter $\beta$. For the value of $\beta$ used in Ref. [III] ($\beta = 1.25$), the average long-range dispersal distance is roughly proportional to the size of the habitat. While long-range dispersal of adults is possible for aquatic plants (water can transport large body masses), for terrestrial plants it is more appropriate to assume instead that fertilised eggs (seeds) can be transported by large distances [143]. To asses the consequences of this difference between terrestrial and aquatic plants, additional simulations with long-range dispersal of seeds (instead of adults) were performed in Ref. [III].

Note that the sexes, and IDs of individuals allow for tracking the dynamics of sexual and genetic structure of the population. The sex and ID are copied from a parent to its asexual recruit. Similarly, the sex and ID are copied in the process of long-range dispersal of adults. By contrast, a sexual recruit has equal chances to be a male or a female, and each sexual recruit receives a new ID.

The model described above was simulated in Ref. [III] for a wide range of parameter values. In all simulations, the per-individual per-generation rate of production of successful sexual propagules was set to two, whereas the per-individual per-generation rate of production of successful asexual propagules (hereafter *clonal birth rate*) was set to a constant $c \leq 1$. Note that a well-mixed population with equal number of males and females has the same capacity for sexual and asexual reproduction when $c = 1$. Conversely, when $c < 1$, the former capacity is larger than the latter. The habitat size was set to a large value ($N = 2000$ in most simulations), and the population was initialised with 100 individuals occupying neighbouring patches around a single origin of population expansion. The sexes were initially homogeneously distributed, and individuals were assigned unique IDs. For other parameter values, refer to Ref. [III].

The population under the model described above has a fixed maximum size. Thus, due to random fluctuations, the population must even-

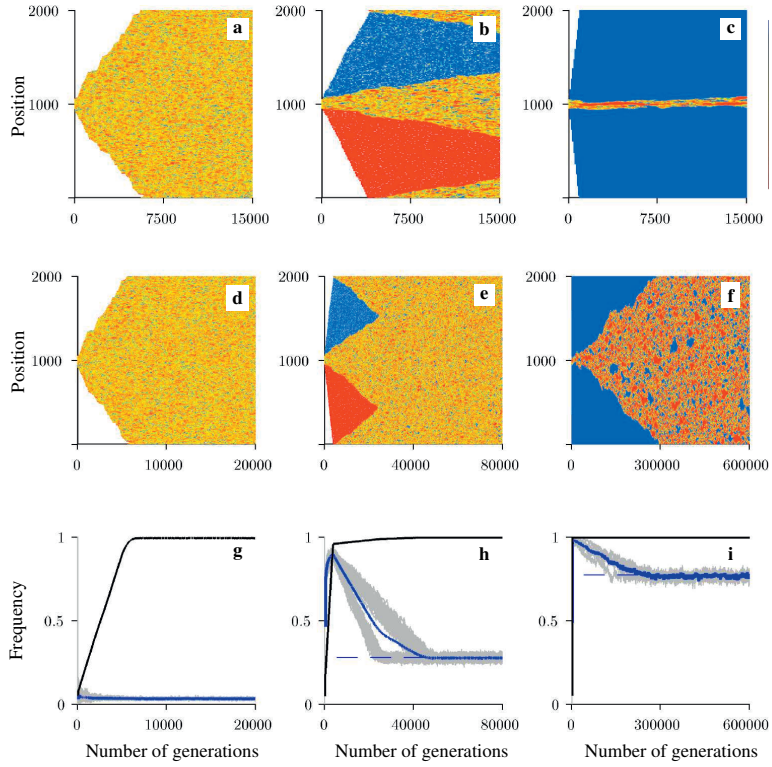


Figure 5.2: The importance of the rate of clonal reproduction $c$. Panels **a**-**c**: space-time patterns of local sex ratios (pools of 8 neighbouring patches) obtained from single stochastic realisations of the model with $c = 0.02$ in **a**, $c = 0.2$ in **b**, and $c = 1$ in **c**. Empty patches are coloured white. Blue depicts all males, and red all females. Yellow denotes sex ratio $1 : 1$. Panels **d**-**f**: same as in **a**-**c**, respectively, but during longer times. Panels **g**-**i**: time dependence of the average frequency of occupied patches (black line), and of the average frequency of asexuals (solid blue line) for the parameter values in **d**-**f**, respectively. Grey lines depict the results of independent simulations (100 runs in **g**, **h**, and 5 runs in **i**). Dashed blue lines indicate the frequency of asexuals in the quasi-steady state. Initial configuration: 100 neighboring patches occupied, alternating sexes. The per-individual rate of production of sexual propagules is fixed to two. Remaining parameters: $N = 2000$, $\beta = 1.25$, $\alpha = 3$, $d = 10^{-2}$, $l_r = 0$. This figure and the caption are taken from Ref. [III].

tually experience extinction [144]. The time to extinction depends on the population size, as well as on the overall population death and birth rate. For most of the parameter values tested, the population initialised with 100 individuals increases in size until the habitat is filled. Thereafter the population size remains roughly constant during the time simulated (up to $10^6$ generations, [III]). However, this is not true when the death rate is high ($d = 1$), and the clonal birth rate is small ($c \leq 0.2$, recall also that the short-range dispersal parameter is set to $\alpha = 3$, or 10 in Ref. [III]). In these cases, the population does not manage to colonise the habitat, and instead experiences rapid global extinction.

Apart from global extinction, the population is expected to eventually experience global fixation of one sex (unless extinction occurs prior to global fixation). However, global fixation was not observed in the simulations carried out in Ref. [III] (excluding the parameter values for which rapid global extinction occurred). In what follows, global extinction and fixation are disregarded.

The results show that in a moderately to a highly asexual species that colonises new areas, clones establish the front of the colonisation ("an asexual wave"). This is true even when the rate of clonal reproduction is low in comparison to the rate of sexual reproduction (but not too low, Fig. 5.2). Note that sexual reproduction is locally impossible in the single-sex areas established during the colonisation of the habitat (Fig. 5.2**b**-**c**). The IDs of the individuals show that these areas are single-clone colonies.

The front of the colonistion in the form of the single-clone colonies establishes because sexual reproduction is hindered in areas that are yet to be colonised due to the lack of mates. Thus, clones are more frequent than sexual recruits during colonisation. However, sexual reproduction is preserved around the origin of population expansion. The region where sexual reproduction is possible progressively spreads over the clonal colonies established, leaving behind essentially homogeneously distributed sexes. In the model analysed in the Master thesis [138] (closely related to the present model), it was shown that the region of essentially homogeneous sex ratio spreads in the form of a traveling wave with a speed proportional to $\alpha$. The speed was found to be larger for smaller values of the clonal birth rate $c$ [138].

When the colonisation of the habitat is completed, and there are no new areas for clones to expand over, asexual recruitment decreases due to the expansion of the region that carries sexual recruits. This process eventually completely erodes the clonal colonies that were formed during the colonisation of the habitat. Finally, the population reaches a quasi-steady state. In this state, the sexes are essentially homogeneously distributed. Due to local random fluctuations, however, small single-sex

colonies appear temporarily. Unlike the single-sex colonies that form during the colonisation, the colonies in the quasi-steady state consist of multiple genotypes of the same sex. The overall frequency of asexuals in this state fluctuates around a constant value that increases with increasing the rate of clonal reproduction, all else being the same. The population, however, must eventually exit this state to experience global extinction, prior to which it can also experience global fixation (see above).

The colonisation of the habitat in the form of a clonal expansion, as well as the overall persistence time of the clonal colonies formed during the colonisation are supported by longer lifetimes of individuals [III]. This is expected because sexual recruits spread over already colonised areas after the individuals in the colonised areas die (the latter occurs with the probability $d$). Thus, the waiting time for the expansion of the region with sexual recruits is longer when $d$ is smaller (all else being the same). By contrast, occasional long-range dispersals of adults support asexuals during the colonisation, but the persistence time of the clonal colonies is shorter when the long-range dispersal rate is larger [III]. This is because when the habitat is empty, patches located far from the origin of colonisation can be colonised thanks to long-range dispersal. These colonisers, being unable to find local mates, reproduce strictly clonally (their sexual propagules remain unused). But during later stages, long-range dispersal promotes the introduction of the individuals of the opposite sex into the clonal colonies established, and hence facilitates the erosion of the clonal colonies. These effects are important for the underlying genetic structure of the population in the presence of long-range dispersal [III]. With solely short-range dispersal, the population establishes two dominant single-clone colonies. By contrast, when occasional long-range dispersal is allowed for, the population typically establishes more than two dominant clonal colonies. Among these, multiple widespread colonies may belong to single clones. Finally, with increasing the size of the habitat, the chances that the population establishes large clonal colonies and to maintain them over extended periods of time increase, because the mean long-range dispersal distance increases [III].

The results obtained under the model are in agreement with empirical observations concerning geographic distributions of clonal and sexual recruits [50–56]. Indeed, the model confirms that during colonistion of new areas, a species that reproduces both sexually and asexually is expected to have a higher frequency of asexuals than of sexuals, with a few clones being dominant. Thus, the dominant female observed in *F. radicans* is not necessarily dominant because her genotype is superior. She may be a "lucky" female that established the front of the colonisation.

By comparing the results of the model with relocation of seeds to

those of the model with relocation of adults, it is found that widespread multiple (distant) clonal colonies of a single clone are much less likely in the former than in the latter case [III]. In terms of natural populations, the former model is more appropriate for terrestrial than for aquatic plants [143]. The results obtained under the two models are in agreement with empirical findings concerning terrestrial plants [145–147] in relation to aquatic ones [49–51, 54].

Finally, note that due to limited dispersal, the overall genotypic diversity in the population on the long run is larger than that expected in well-mixed populations with mixed sexual and asexual reproduction (results not shown). Therefore it may be very difficult to distinguish between clonal and sexual recruits from empirical genetic sequences. Inferring the rate of clonal reproduction is difficult also in well-mixed populations [65, 148]. But, the problem is more difficult in the presence of limited dispersal capability, as the genotypic diversity in this case is larger than in well-mixed populations.

The model outlined neglected the fact that sperms, eggs, and asexual propagules have different biomasses, and that they may have different dispersal capabilities. Furthermore, in the model it was assumed that each individual produces on average the same amount of sexual (and of asexual) propagules per generation. In other words, the 'reproductive-strategy trait' was fixed. The question is: how does this trait evolve during colonisation? This depends on the genetic content of the first founders, but also on dispersal capabilities of propagules. Because eggs are typically heavier than sperms, and hence eggs are expected to disperse by shorter distances than sperms, it is unclear whether a reproductive strategy of males can differ from that of females [138]. In order to answer this question, further modelling work is required.

This chapter concludes the discussion in this thesis concerning neutral genetic variation. The following chapter analyses the effect of selection, migration and drift on the dynamics of local adaptation.

# 6

# Adaptation in small partly isolated subpopulations

Chapters 3-5 described how varying population sizes, geographic structure, multiple paternity, and mixed sexual and asexual reproduction influence the patterns of neutral genetic variation. However, some loci are inevitably influenced by natural selection (see Section 2.4). This chapter discusses how selection in geographically structured populations influences genetic variation at loci targeted by selection. The results presented below were obtained in collaboration with Anna Emanulesson (Master thesis [149]), Fengchong Wang (Master thesis [150]), Roger Butlin, Kerstin Johannesson, and Bernhard Mehlig.

As explained in Section 2.4, an individual may be more or less fit in a given environment, depending on how close its phenotype is to the optimal phenotype in the environment. When environmental conditions differ across the geographic distribution of a given population, different parts of the habitat have different optimal phenotypes, and the population is said to be exposed to divergent selection. Whether or not individuals in the different subpopulations are able to adapt to the respective optima under divergent selection, depends on the selection strength, but also on the geographic structure of the subpopulations, i. e. whether they are isolated from each other or not. Theoretical studies suggest that adaptation is more difficult in the latter than in the former case [68]. This is because migration opposes the effect of selection. While selection tends to remove deleterious alleles from, and to increase the frequency of beneficial alleles in each subpopulation, migration between subpopulations exposed to divergent selection has an opposite effect: migration repeatedly removes beneficial alleles from, and introduces dele-

terious alleles in each subpopulation. This opposing effect of migration and divergent selection results in an establishment of migration-selection balance that inevitably limits, or prevents the capability of the individuals in the subpopulations to adapt to their respective environments [68, 69, 110, 151, 152]. Still, empirical studies show that it is common to find in nature examples of adaptive divergence when subpopulations are not isolated from each other, and indications are that adaptation in the presence of migration may occur rapidly [6, 7, 9, 22–24]. An example is the sea snail *Littorina saxatilis* with two divergent ecotypes formed [36], as already mentioned in Section 2.4. The two divergent ecotypes of *L. saxatilis* are capable of producing viable and fertile offspring, and hence they are not treated as separate species. However, mating between them (though possible at the present time) is difficult due to the difference in their sizes. Thus, the loci giving rise to the size differences between the two ecotypes can be understood as potential candidates for establishing primary reproductive barriers between the ecotypes [6, 8]. On the long run, the two ecotypes might establish complete reproductive barriers, and hence evolve to separate species [6–8, 10]. By studying the genomes of the two ecotypes, the mechanisms responsible for establishing primary reproductive barriers can be revealed [7, 10]. In order to be able to interpret empirical genome data, and detect loci that are responsible for the establishment of primary reproductive barriers between divergent ecotypes, it is necessary to understand the dynamics of local adaptation.

An example of local adaptation with two divergent partly isolated subpopulations is shown in Fig. 6.1. In the subpopulation with the positive value of the optimal phenotype, individuals with positive phenotypes are more frequent than those with negative phenotypes (Fig. 6.1). The opposite is true for the other subpopulation. Recall that the phenotype of an individual is determined by the genotype at the locus (or loci) targeted by selection. In diploid populations, a genotype is determined by pairs of alleles at loci subject to selection. Usually, each allele is characterised by its effect size, and the phenotype is determined by the sum of allele-effect sizes at the loci in question. An allele-effect size can be altered by a mutation. Similarly to alleles, mutations are also characterised by effect sizes. Upon experiencing a mutation, an allele attains the effect size that is equal to the sum of its size prior to the mutation, and the mutation-effect size. Adaptive divergence in Fig. 6.1 occurs thanks to mutations that manage to spread and maintain in the population. The question arises: under which conditions can a mutation increase in frequency despite being beneficial in one and deleterious in the other subpopulation?

As explained in the previous paragraph, local adaptation in one sub-

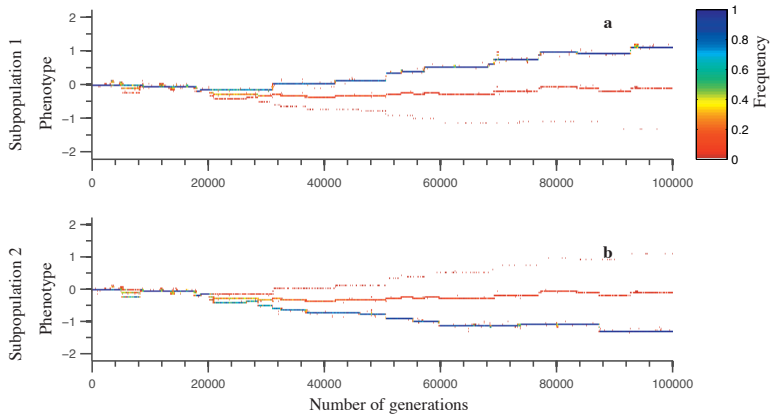


Figure 6.1: Local adaptation in two partly isolated subpopulations subject to divergent selection acting upon a single locus. The results shown are obtained using a stochastic simulation of adaptation. Panels **a**, and **b** show the time dependence of the phenotypes of individuals in the subpopulation with a positive value of the optimal phenotype ($\theta^{(1)} = 2$, **a**), and in the subpopulation with a negative value of the optimal phenotype (**b**, $\theta^{(2)} = -2$). The frequencies of individual phenotypes are colour coded (see the colour bar). The subpopulations are initially monomorphic and contain only alleles of zero effect size. Mutation-effect sizes are drawn from a normal distribution with mean zero, and a standard deviation of $\sigma_\mu = 0.05$. Remaining parameters used: selection parameter $\sigma = 2.5$, mutation rate $\mu = 10^{-4}$, migration rate $m = 0.01$, population size $N = 100$. The simulations were performed by Anna Emanuelsson (Master thesis [149]).

population occurs through the accumulation of alleles of positive effect sizes. Conversely, local adaptation in the other subpopulation occurs through the accumulation of alleles of negative effect sizes. Therefore, in order for adaptive divergence to occur, the population as a whole must be polymorphic at one or more loci subject to selection. The frequency of deleterious alleles in each subpopulation is determined by a migration-selection balance. Many authors have analysed the conditions under which a population subject to divergent selection establishes a stable polymorphism. In Refs. [68–74] the stability properties of polymorphic states were analysed under the assumption that the subpopulations were of infinite sizes, thus neglecting the effect of random genetic drift. The

main finding in these studies is that when migration is too strong in comparison to selection, polymorphic states are unstable, and hence divergent adaption to the opposing environments is prevented. However, even if a given polymorphic state is found to be stable, this does not guarantee that the population establishes it due to the effect of random genetic drift.

The effect of random genetic drift was accounted for in, e. g. Refs. [17, 75, 76, 153]. In Ref. [76], selection was assumed to act on one locus. Two to five loci under selection were analysed in Ref. [153]. Conversely, complex models of genetic architectures during adaptation were analysed in Refs. [17, 75]. The analysis in Ref. [76] concerns the conditions for which an initially monomorphic population establishes a dimorphism upon introducing a mutant allele beneficial in one, and deleterious in the other subpopulation. Due to a migration-selection balance, there is a critical migration rate above which the initially monomorphic population cannot establish and maintain a dimorphism [76]. The critical migration rate increases with increasing the population size, showing a decreasing importance of the effect of random genetic drift. The same is true when the selection strength for a mutant allele in the subpopulation where it is beneficial is increased, and the selection strength against the mutant allele in the subpopulation where it is deleterious is decreased [76]. A similar analysis can be found in Ref. [153], but in the model in Ref. [153], the population is initialised with one monomorphic locus, and a few other dimorphic loci. This study found that the probability that a dimorphism is established at an initially monomorphic locus is typically not influenced by divergence at other loci, but by the selection strength for a mutant allele in the subpopulation where it is beneficial, and the selection strength against the mutant allele in the opposite subpopulation [153]. However, the results in Refs. [76, 153] are relevant for understanding the first step of adaptation at a given locus - a transition from a monomorphic to a dimorphic state. Therefore, it needs to be further understood how the population manages to escape one dimorphism and establishes another polymorphic state. In order to understand this, the following questions need to be answered. Firstly, how does the (deterministic) critical migration rate above which a mutation of a given effect size cannot establish in the population depend on the extent of divergence between the subpopulations? Secondly, how does the probability that a mutation of a given effect size spreads in the population depend on how far the subpopulations have diverged prior to the mutation? Specifically, does an increased level of divergence reduce the probability that a mutation of a given effect size establishes in the population (as is the case for freely-mixing populations exposed to a fixed environment [67])? Thirdly, do the answers to

the first and second question posed above depend on whether selection acts upon one or multiple loci? The answers to these questions are necessary in order to determine the conditions under which local adaptation in the two subpopulations is possible to progress towards the divergent local optima, despite migration and random genetic drift.

In order to answer these questions, this chapter analyses the following model for local adaptation. The population consists of two diploid subpopulations of equal size ($N$ individuals). Note that an experimental design with two divergent subpopulations has been suggested to be crucial for increasing the statistical power of data in empirical studies of adaptation [152]. In the model generations are discrete and non-overlapping. In the beginning of each generation adult individuals migrate to the neighbouring subpopulation with the probability $m$ per generation, individual, subpopulation. After migration, adults in each subpopulation mate randomly, whereas individuals residing in different subpopulations after migration do not mate with each other. In order to keep the population size constant in time, $2N$ gametes are sampled from parental individuals in each subpopulation. The number of gametes that a parent contributes to the next generation is assumed to be proportional to the fitness of the parent relative to the average fitness of all parental individuals in a given subpopulation. This corresponds to soft fecundity selection, as implemented in Ref. [75]. The fitness of a parent is determined by its phenotype, which is in turn assumed to be determined by the diploid genotype at a given locus under selection. Alleles at the locus are given allele-effect sizes, and the phenotype is assumed to be equal to the sum of the corresponding pair of allele-effect sizes at the locus. The fitness $w_i^{(j)}$ of the parent with phenotype $z_i$ in subpopulation $j = 1, 2$ is given by

$$w_i^{(j)} = \mathrm{e}^{-\frac{(z_i-\theta^{(j)})^2}{2\sigma^2}} \ . \tag{6.1}$$

Here $\sigma$ is a selection parameter (assumed to be the same in both subpopulations), and it is such that selection is effectively weaker when $\sigma$ is larger, all else being the same. Furthermore, $\theta^{(j)}$ denotes the optimal phenotype in subpopulation $j$. It is assumed that $\theta^{(1)}$, and $\theta^{(2)}$ are symmetric around zero, so that $\theta^{(1)} = -\theta^{(2)} = 2$. The fitness function (6.1) corresponds to the fitness function (2.13) introduced in Chapter 2, but here the 'trait space' is assumed to be one-dimensional. The $2N$ gametes sampled as described above pair randomly (without replacement) to form the next generation of $N$ diploid individuals. In the example of local adaptation shown in Fig. 6.1, each gamete is assumed to experience mutations with the probability $\mu$ per generation, gamete, subpopulation. Mutation-effect sizes are drawn from a Gaussian distribution with mean

zero, and a standard deviation $\sigma_\mu$. However, to understand whether or not a mutation of a given effect size (beneficial in one, and deleterious in the other subpopulation) can spread in the population, the main analysis in this chapter is performed as follows.

First, the population is assumed to consist of allele-effect sizes $X$, and $-X$ (where $0 \leq X \leq 1$). Thus, an individual can have a phenotype equal to $2X$ (a homozygote with both alleles having effect sizes equal to $X$), or $-2X$ (a homozygote with both alleles having effect sizes equal to $-X$), or 0 (a heterozygote). A deterministic approximation of the model shows that the population initialised with allele-effect sizes $X$, and $-X$ establishes a stable dimorphism (hereafter denoted by $\{X, -X\}$). This is true for any value of the migration rate, as also shown in Ref. [76]. The frequency of deleterious alleles in each subpopulation is governed by a migration-selection balance, and it decreases with increasing $X$ (that is, with increasing the extent of divergence between the two subpopulations). Second, starting from the stable dimorphism $\{X, -X\}$, a mutation of effect size $\epsilon$ ($0 \leq \epsilon \leq 1 - X$) is assumed to alter one allele of size $X$ in the subpopulation with the positive optimal phenotype (where it is beneficial). The entire population (with allele-effect sizes $X$, $-X$, and $X + \epsilon$) is afterwards assumed to evolve until the new (or the old) steady state is achieved, whereas further mutations are neglected.

Additional analysis is performed to account for two loci under selection. In this case, it is assumed that with the probability $r$ per generation, gamete, subpopulation, a gamete is a product of recombination between the chromosomes of its parent. The effect of a mutation in the two-locus model is studied in a similar manner as in the one-locus case. But here the two loci are allowed to differ in the extents of their divergence. This is because the migration-selection balance established at one locus may differ from that established at the other locus if the two loci differ in the extents of their divergence. To be able to understand the consequences of this effect, the population is initialised with allele-effect sizes symmetric around zero at both loci, but with $Y$, and $-Y$ at one locus, and $Y + \alpha$, and $-Y - \alpha$ at the other (where $0 \leq Y \leq 0.5$, and $0 \leq \alpha$). When $r = 0$, the two-locus model reduces to the one-locus model, in which case it holds that $X = 2Y + \alpha$. The parameter $\alpha$ serves to distinguish between the extents of divergence at the loci. When $\alpha = 0$, the loci do not differ by the extent of their divergence, whereas when $\alpha > 0$, the second locus exhibits a higher divergence than the first one. To emphasize this, in the analysis of the two-locus model below, the first, and the second locus are referred to as the locus of smaller, and larger effect size, respectively. The effect of a mutation is studied upon introducing it to the locus of either larger, or smaller effect size.

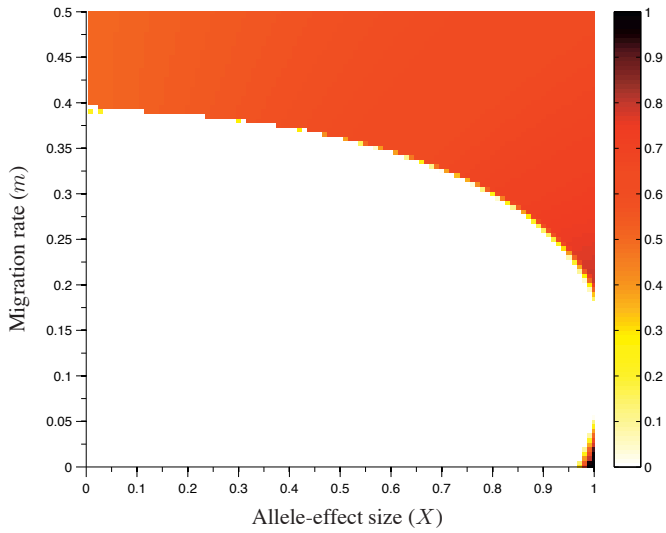


Figure 6.2: The stable-state frequency of allele-effect size $X$ in the subpopulation with the positive optimal phenotype in dependence of $X$, and migration rate $m$ (one-locus model). The population is initialised in vicinity of the steady dimorphism $\{X, -X\}$ in such a way that a small fraction of alleles of effect size $X$ is substituted by the same amount of alleles of effect size $X + \epsilon$. The frequencies are colour coded (see the colour bar). The values of migration rate $m$ tested are: $0 \leq m \leq 0.5$, and of allele-effect size $X$ are: $0.01 \leq X \leq 1$ (mesh size 0.01 in both cases). The final steady state is determined by iterating the dynamics of the system (deterministically) until a predetermined stopping condition for the steady-state is met. The stopping condition used: none of the frequencies of allele-effect sizes changes by more that $10^{-5}$ within 1000 consecutive generations. The maximum number of generations used to find the steady state: $10^6$. Remaining parameters: mutation-effect size $\epsilon = 0.05$, selection parameter $\sigma = 2.5$. The simulations were performed by Anna Emanuelsson (Master thesis [149]).

To answer the first question posed above, a deterministic approximation is employed. By tracing the (deterministic) dynamics of the frequencies of allele-effect sizes $X$, $-X$, and $X+\epsilon$ (initialised in vicinity of the steady dimorphism $\{X,-X\}$ as explained above) one can arrive at the stable steady state of the population (Fig. 6.2, see also Figs. B.1-B.2). The stability of the steady state found (for given $m$, and $X$) is checked by numerically computing the eigenvalues of the stability matrix of the system (Appendix B). In the white region shown in Fig. 6.2 the frequency of $X$ is approximately zero within the numerical precision used. The same region is white in the opposite subpopulation as well (not shown). Thus, in this region the deterministic approximation predicts that alleles of effect size $X$ experience extinction. In conclusion, when the migration rate is not too large, the population establishes a stable dimorphism $\{X+\epsilon,-X\}$ (hereafter, the mutant allele *replaces* the resident one). However, when migration is too frequent, the population does not manage to escape the initial dimorphism $\{X,-X\}$. The critical migration rate above which the mutant allele of size $X+\epsilon$ cannot replace the resident one of size $X$ decreases with increasing the extent of divergence between the subpopulations ($X$). This effect is more pronounced for weak levels of selection (large $\sigma$), and it becomes negligible for stronger selection (Fig. B.3). The dependence of the critical migration rate on the extent of divergence between the two subpopulations is similar for mutation-effect sizes ($\epsilon$) smaller than that set in Fig. 6.2. But the critical migration rate in late stages of divergence is somewhat higher for smaller than for larger values of $\epsilon$ (results not shown).

Note that the algorithm used did not manage to estimate the steady-state frequencies for several combinations of the parameters $m$, and $X$ in a narrow region depicted in blue in Fig. B.1. The reason for this is that the steady states in this region are marginally stable.

Since natural populations are of finite sizes, the results obtained using a deterministic approximation need to be adjusted to account for the effect of random genetic drift. Indeed, while under the deterministic approximation $X+\epsilon$ replaces $X$ when the migration rate is below the critical value (Fig. 6.2), genetic drift may limit this tendency in small populations. As expected, the simulation results show that the probability that $X+\epsilon$ replaces $X$ decreases with increasing the migration rate $m$ (Fig. 6.3). Furthermore, the replacement probability is typically larger when the mutation-effect size ($\epsilon$) is larger (compare red and blue symbols in Fig. 6.3). However, the difference between the two replacement probabilities decreases with increasing the migration rate. For sufficiently frequent migration (Fig. 6.3**c**), the replacement probability obtained for the smaller mutation-effect size is approximately equal to that obtained

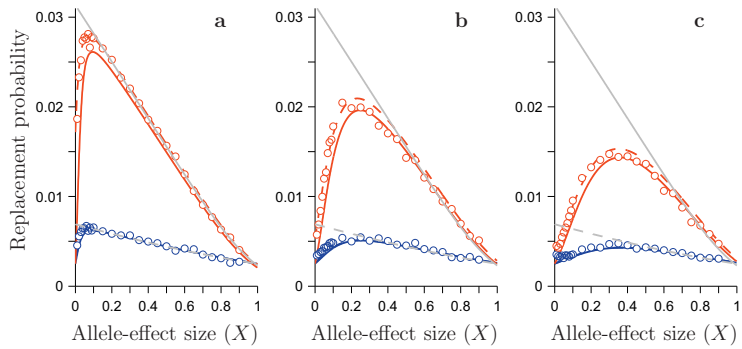


Figure 6.3: Replacement probability in dependence of allele-effect size $X$ (one-locus model). The population is initialised in the (deterministically determined) steady dimorphism $\{X, -X\}$, where one allele of effect size $X$ in the subpopulation with the positive optimal phenotype experiences a mutation of size $\epsilon$. The simulations are advanced until either allele-effect size $X$, or $X + \epsilon$ experiences extinction. In the former case, the replacement event is noted. The mutation-effect size $\epsilon$ is set to $\epsilon = 0.01$ (blue), or $\epsilon = 0.05$ (red). The migration rate is $m = 0.01$ in **a**, $m = 0.05$ in **b**, and $m = 0.1$ in **c**. Grey lines show the analytically computed fixation probabilities (Eq. (2.11) in Section 2.4 [77, 78]) in an isolated subpopulation ($m = 0$) for $\epsilon = 0.01$ (dashed), and $\epsilon = 0.05$ (solid). The simulation results are shown by symbols. Blue and red lines show Eq. (2.11) but modified as follows: the selection strength in Eq. (2.11) is substituted by a diversification coefficient evaluated at the stable dimorphism $\{X + \epsilon, -X\}$ (solid), or at the unstable dimorphism $\{X, -X\}$ (dashed). Remaining parameters: selection parameter $\sigma = 2.5$, population size $N = 200$, $2 \cdot 10^5$ stochastic simulations. The simulations were performed by Anna Emanuelsson (Master thesis [149]).

for the larger mutation-effect size during the initial stages of adaptation. Lastly, recall that in the case of a freely-mixing population (exposed to a fixed environment), the fixation probability of a mutation of a given size strictly decreases as the individuals in the population approach the optimum, i. e. as $X$ increases [67]. This is depicted by the grey lines in Fig. 6.3. By contrast, when divergent subpopulations exchange migrants, the replacement probability does not strictly decrease as the individuals in the subpopulations approach the respective divergent optima. Rather, when $X$ is small, the replacement probability increases with increasing $X$. This occurs until $X$ reaches a certain threshold value at which point the replacement probability starts decreasing with increasing $X$. The opposite trend is observed for the average time needed for the replacement to occur (results not shown).

What causes this complex dependence? Whether or not the replacement occurs depends on how advantageous the mutant homozygote is over the resident ones, but also on the strength of the counteracting effect of migration. In the subpopulation where the mutant allele is introduced, the selection strength for the mutant homozygote over the resident beneficial homozygote strictly decreases with increasing $X$, all else being the same (see grey lines in Fig. 6.3). But the effective migration rate between the two subpopulations also decreases with increasing $X$. This can be seen in Fig. B.2: for a given migration rate, the frequency of $X+\epsilon$ changes from yellow to dark red as $X$ increases, suggesting a reduction of the frequency of $-X$ in this subpopulation. The effective migration rate decreases much more rapidly in the beginning of adaptation than towards its end (Fig. B.2). Thus, the results in Fig. 6.3 suggests that, with increasing $X$ in the beginning of adaptation, the (negative) effect of migration decreases more rapidly than the (positive) effect of selection, thus facilitating the replacement. But when the effective migration is reduced to a large extent, the negative effect of migration decreases much more slowly with increasing $X$ than does the positive effect of selection. As a result, in this region one finds a decreasing trend of the replacement probability. The slope of the decreasing trend in this region is approximately the same for different migration rates, including the case of isolated subpopulations.

Finally, the results described above are not only valid when selection acts on a single locus, but also in the two-locus model (Fig. 6.4): in the beginning of adaptation, the replacement probability increases as the individuals in the two subpopulations approach their respective optima, while for intermediate to high levels of divergence between the subpopulations the replacement probability decreases. When allele-effect sizes at the two loci differ ($\alpha > 0$), the locus with the larger allele-effect

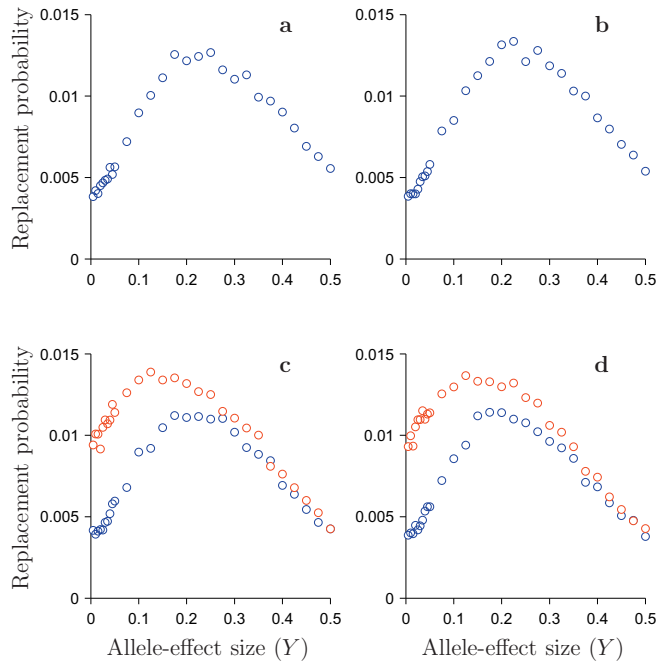


Figure 6.4: Replacement probability in dependence of allele-effect size $Y$ (two-locus model). The population is initialised in the (deterministically determined) steady state with allele-effect sizes $Y$, and $-Y$ at one locus, and $Y+\alpha$, and $-Y-\alpha$ at the other. Thereafter, one allele of effect size $Y$ in the subpopulation with the positive optimal phenotype experiences a mutation of size $\epsilon$. The simulations are advanced until either allele-effect size $Y$, or $Y+\epsilon$ experiences extinction. A separate set of simulations is performed for the case that a mutation lands on the second locus. The parameter $\alpha$ is set to $\alpha = 0$ (**a**-**b**), or $\alpha = 0.1$ (**c**-**d**). In **c**-**d**, the replacement at the locus of smaller effect size ($Y$) is depicted in blue, whereas at the locus of larger effect size ($Y+\alpha$) is depicted in red. The recombination rate $r$ is set to $r = 0.05$ (**a**, **c**), or $r = 0.5$ (**b**, **d**). Remaining parameters: selection parameter $\sigma = 2.5$, mutation-effect size $\epsilon = 0.05$, migration rate $m = 0.1$, population size $N = 200$, $2 \cdot 10^5$ stochastic simulations. The simulations were performed by Anna Emanuelsson (Master thesis [149]).

size experiences a larger replacement probability than the locus with the smaller allele-effect size (compare red and blue symbols in Fig. 6.4**c-d**). This suggests that one locus tends to diverge faster than the other in the course of adaptation, and hence supports the finding of 'concentrated genetic architectures' in Ref. [17]. The difference between the replacement probabilities at the two loci, however, decreases as the individuals in the subpopulations approach their respective optima. In the cases shown, the replacement probability is approximately the same for different values of the recombination rate (Fig. 6.4, compare also to Fig. 6.3**c** recalling that the one-locus model is obtained upon setting $r = 0$ in the two-locus model, and it holds that $X = 2Y + \alpha$).

In summary, the findings outlined above show a number of features of the dynamics of divergent local adaptation. First, the deterministic critical migration rate above which a mutation of a given effect size cannot spread in the population typically decreases as the individuals in the subpopulations approach the corresponding optima. This effect is more pronounced for weak than for strong selection. Second, the probability that a given mutation-effect size spreads in the population of a finite size depends in a complex manner on how close the individuals in the subpopulations are to their respective optima. In the beginning of adaptation, the probability increases as the individuals in the subpopulations approach their optima. But at a certain divergence level, the probability starts decreasing with increasing the extent of divergence. In this region the decreasing trend of the replacement probability is approximately the same for different migration rates (including the case of isolated subpopulations). This suggests that signatures of migration should be more difficult to detect at later than at early stages of adaptation. However, the results also show that in the beginning of adaptation, the difference between the replacement probability under a mutation of smaller effect size and that under a mutation of larger effect size decreases with increasing the migration rate. This suggests that under relatively frequent migration between the two subpopulations, initial stages of adaptation may be carried out by small mutation-effect sizes. These may be difficult to detect in genome-wide scans [8, 20, 112, 115].

The findings under the one-locus model hold also for the two-locus model. But the advantage of the two-locus model is that it allows to understand qualitatively how the genetic architecture is expected to evolve in the course of divergent adaptation. The two-locus model shows that a mutation of a given effect size experiences a larger chance of being established in the population if it lands on a more than on a less diverged locus. This is because the migration-selection balance established at one diverged locus differs from that at the other locus. The effective migra-

tion rate at the locus with the higher extent of divergence is smaller than at the other locus, and hence the mutation is more likely to spread when it lands on the former than on the latter locus. This can be understood as a mechanism for establishing 'concentrated genetic architectures' [17]. According to the results shown here, it is expected that the tendency of the population to establish 'concentrated genetic architectures' is most pronounced for small levels of divergence between the subpopulations, and it decreases as the individuals in the subpopulations approach their respective optima. The latter suggests that 'concentrated genetic architectures' should be less expected in populations that are polymorphic prior to divergence, where the polymorphism consists of intermediate to large allele-effect sizes (for example, these can come from standing genetic variation [111], refer also to Chapter 4).

The results presented here stress the complexity of the interplay between migration, selection, and random genetic drift in the course of adaptation in partly isolated subpopulations. They can serve as a starting step towards estimating the distribution of mutation-effect sizes established in the population as adaptation progresses. Clearly, the results suggest that the corresponding distributions of the established mutation-effect sizes during adaptation in partly isolated subpopulations should differ from that in completely isolated subpopulations. But it should also be expected that the differences between the distributions change as the individuals in the subpopulations approach their respective optima.

The dependence of the replacement probabilities on the extent of divergence between the subpopulations is a necessary step towards understanding under which conditions the divergence is possible, and how close the individuals can approach the local optima. The next step is to estimate the persistence time of the different extents of divergence between the subpopulations. This result, and the results presented here on the replacement probability can be used to estimate the typical magnitude of the mutation-effect sizes, and their frequency of appearance that are necessary for divergent adaptation to progress towards the divergent optima, despite migration and random genetic drift.

# 7

# Summary and conclusions

It is well-know that nature is rich in species [154], as well as that new species continue to emerge from the existing ones through the process of speciation [6, 7, 9, 22–24]. But species are also subject to the risk of extinction. A confounding factor for the survival of a given species is its genetic variation, especially at genome regions (loci) subject to selection. Many authors have raised the question: which loci are subject to selection (see, for example, Refs. [27, 29–32, 119])? While today it is possible to obtain empirical genome-wide patterns of genetic variation, it is still difficult to interpret them. Indeed, there may be a bias in the interpretation due to the (usually unknown) underlying demographic history of the population in question, or its geographic structure. Furthermore, genetic variation is inevitably subject to stochastic fluctuations. To understand the importance of stochastic fluctuations, more or less complex demographies, and geographic structures for the patterns of genetic variation, it is necessary to use models. This thesis presented and analysed a number of models relevant for biological species.

Chapter 3 discussed the effect of varying population sizes on commonly used tests of neutrality based on site frequency spectra (SFS) of SNPs, e. g. Tajima's $D$ [32], Fay & Wu's $H$ [20], and others. These tests are built upon assuming that the population size is constant. Empirical genome-wide distributions of Tajima's $D$ computed along the genomes of different Human populations differ substantially between the different populations [I]. This raises the question: since the distributions have very different shapes, how can one compare the extents of selection at candidate regions between the different populations? In order to answer this question, it was proposed in Ref. [I] to redefine the SFS-based tests by integrating the underlying demographies of the populations into the

tests. The tests obtained in such a way are termed demography-adjusted. To obtain demography-adjusted tests, the underlying demography of the population in question must be estimated. This can be done by applying a maximum-likelihood approach on empirical data sampled from intergenic (presumably neutral), physically distant SNPs [28, 121]. In Ref. [I] candidate demographies were constraint to a piecewise constant demographic model with at most two population-size changes. This model has been argued [28, 29, 121] to capture the main events of the human out-of-Africa expansion [37–39, 123]. The maximum-likelihood procedure was tested agains simulated data. The procedure constrained to the demographic model assumed performs well if the estimation is based on at least $10^5$ SNPs. The same procedure was executed on empirical data from ten human populations. This allowed to construct the corresponding demography-adjusted tests. The empirical genome-wide distributions of the demography-adjusted tests were found to be similar between the different populations. Therefore demography-adjusted test distributions facilitate the comparison between the extents of selection of candidate regions between different populations. However the values of demography-adjusted tests were found to be roughly linearly related to the values of unadjusted tests. Thus demography-adjusted and unadjusted tests yield the same candidate regions under selection. But, are intergenic regions neutral [122]? Recall that a linear relationship observed between the empirical values of demography-adjusted and unadjusted tests was confirmed by computer simulations of neutral genetic variation at loci (not subject to recombination) under a given null demography. But if intergenic regions are under selection, how does this bias demography estimation and the resulting demography-adjusted test distributions? Furthermore, the effect of recombination cannot be neglected along wide windows of the genomes of Humans, whereas it is neglected in test definitions [20, 32, 116, I] as well as in computer simulations [I]. Therefore, it remains unclear: why is a linear relationship observed between empirical values of adjusted- and unadjusted tests? The effect of selection on intergenic regions and of genome-wide recombination rate on demography-adjusted tests of neutrality is yet to be understood.

Chapter 4 analysed genetic variation under the model that was constructed in Ref. [II] to mimic the colonisation history of the snail *Littorina saxatilis* [126, 128, II]. This species inhabits large refuge areas (mainlands) from which it spreads in a stepwise fashion over smaller islands that have been, and are emerging due to the uplift of land [126]. It was suggested in Ref. [126] that a reduction of genetic variation that is expected due to repeated founder events [125] can be counteracted by multiple paternity observed in this species. But, does the effect of multi-

ple paternity differ during colonisations of new areas and on the long run? In order to answer this question, mating model that allows for both high and low levels of multiple paternity was constructed, and single-locus heterozygosity during and after the colonisation of a mainland-island habitat was analysed in Ref. [II]. The mating model was tested against empirical data gathered from paternal analyses [15, 16]. The model could not be rejected, but empirical distributions showed that the number of males siring one or two offspring was larger than that obtained under the model. As expected [125], it was found that genetic variation decreases with increasing the distance from the mainland, both during the colonisation phase, and in the steady state. The decreasing trend of genetic variation is more severe when the level of multiple paternity is lower, than when it is higher. Thus multiple paternity enhances the heterozygosity. However, the results showed that this effect of multiple paternity is much larger during colonisation than on the long run. This effect is particularly pronounced at islands that are far from the mainland. Here, genetic variation fluctuates significantly with periods of complete loss of genetic variation. Multiple paternity decreases the duration of phases of low genetic variation, and increases the duration of phases of high genetic variation. These results suggest that multiple paternity may be a strategy of this snail to increase genetic variation that the founders introduce to new emerging islands, and thus to increase chances to successfully invade new areas. While multiple paternity in this respect may facilitate the colonisation of new habitats, this may not be the main reason for why this mating pattern evolved. Other reasons, such as avoiding the cost of rejecting an intercourse [15], could also be relevant.

Chapter 5 discussed the sexual and genetic patterns in species that reproduce both sexually and asexually, and have limited dispersal capabilities. Many empirical studies reported that in young or marginal areas species with both modes of reproduction are commonly characterised by widely distributed clones, whereas genotypically unique individuals expand over small areas [49–51, 53, 55, 62]. To explain this pattern, many selection-based hypotheses have been proposed [53, 57–61]. However, empirical tests to support the selection-based hypotheses were not successful [62, 63]. This raises the question: does the dominance of clones in young areas or at the edges of species distributions necessarily imply that selection disfavours recombination of genotypes through sexual reproduction? In order to answer this question, a selection-free model was analysed in Ref. [III]. The model is constructed to explicitly account for the spatial distribution of sexes, and limited (short-range) dispersal capabilities of propagules. The model further allows for occasional long-range dispersal of whole adult individuals. In the model, all individ-

uals are treated equally, independently of whether they are reproduced sexually or asexually. While long-range dispersal of adult individuals is possible for aquatic plants, in terrestrial plants only seeds can be transported by long distances [143]. To be able to assess the consequences of this difference between terrestrial and aquatic plants, an additional model with long-range relocation of seeds was analysed in Ref. [III]. Using simulations, the dynamics of sexual and genetic patterns was tracked during the colonisation of the habitat, and long thereafter. The model results show that moderately to highly asexual populations that manage to colonise the habitat are expected to be dominated by widespread single-sex colonies during the colonisation phase. Each of these large single-sex colonies consists of a single clone (clonal colonies). The clonal-colonies establish the front of the colonisation even when the rate of production of asexual propagules is low in comparison to the rate of production of sexual propagules (but not too low). This is the consequence of limited dispersal capability of propagules: sexual reproduction is locally possible only if both sperms and eggs are present, whereas no such constraint can be made for asexual reproduction. However sexual reproduction persists around the origin of colonisation. The region where sexual reproduction is possible expands over time leaving behind essentially homogenous distribution of sexes. When the colonisation of the habitat is completed, asexual recruitment starts decreasing. The region that carries sexual recruits eventually depletes the clonal colonies formed during the colonisation phase, and hence establishes essentially homogeneous distribution of sexes along the habitat. This distribution of sexes persists for a long time (quasi-steady state) until the population eventually experiences global extinction (alternatively, global fixation of one sex may occur prior to extinction). However, the time until extinction after establishing the quasi-steady state is expected to scale exponentially with the population size [144]. The same is true for the time to fixation. The formation of clonal colonies during colonisation, and their persistence time thereafter are supported by longer lifetimes of individuals. The former is also true for the process of long-range dispersal. The persistence time, by contrast, reduces with increasing the rate of long-range relocation, as this process promotes the introduction of individuals of the opposite sex in the clonal colonies established during colonisation. Finally, with long-range dispersal of adults, the population typically establishes multiple widespread colonies belonging to single clones. This is, by contrast, not a typical outcome under the model with relocation of seeds. In conclusion, the patterns obtained under the selection-free model described are in agreement with observations reported in empirical studies of young colonisers of new areas that reproduce both sexually

and asexually. Thus, this model can serve as a basis for selection-based hypotheses to be tested against.

Finally, Chapter 6 discussed the dynamics of two partly isolated subpopulations subject to divergent selection at a given locus. Individuals are assumed to be diploid, and the phenotype of an individual is determined by the sum of its pair of allele-effect sizes at the locus. It is well recognised that a condition for both subpopulations to experience adaptation towards their respective optima is that the population (as a whole) is polymorphic at the given locus. However, the tendency of the population to establish a polymorphism may be counteracted by migration. Indeed, migration repeatedly removes beneficial alleles from, and introduces deleterious alleles in each subpopulation. The contrasting interplay between migration and selection is termed migration-selection balance. Many studies analysed the conditions under which a population that is initially monomorphic at the locus in question establishes a dimorphism [76, 109, 110]. However, it must be understood: under which conditions the diverged subpopulations can diverge further? Under which conditions a mutation of a given effect size may spread in the population that consists of two already diverged subpopulations? These question were discussed in Chapter 6 . A deterministic approximation shows that when migration is not too frequent, the initially dimorphic population establishes a new dimorphism after a mutation of a given effect size is introduced: the mutant allele spreads in the population, whereas its non mutated version experiences extinction. In Chapter 6, this was termed replacement. The migration rate above which the replacement is not expected to occur decreases with increasing the extent of divergence between the two subpopulations. To assess how random genetic drift influences the tendency of the mutant allele to replace its non mutated version, computer simulations were used. The results show that, for a mutation of a given effect size, the replacement probability depends in a complex manner on the extent of divergence between the subpopulations. By increasing the extent of divergence up to a certain critical value, the replacement probability increases. But after this critical value, the replacement probability decreases with increasing the extent of divergence. Note that in the absence of migration, the replacement probability strictly decreases as the individuals in the subpopulation (where the beneficial mutation is introduced) approach the optimal phenotype [67]. Thus, in the presence of migration the following can be deduced. In the initial stages of adaptation (small extents of divergence) the negative effect of migration decreases with increasing the extent of divergence. The same is true for the positive effect of selection. But the former occurs much more rapidly than the latter. Therefore, in the initial stages

of adaptation the replacement probability increases. In later stages of adaptation, by contrast, the effective migration is so small that it does not have any significant effect on further adaptation. By comparing the results obtained for a mutation of a larger effect size to those for a mutation of a smaller effect size, it was found that the replacement probability in the former case is higher than in the latter case. By increasing the migration rate, the difference between the two replacement probabilities decreases. When migration is relatively frequent, the results show that in early stages of adaptation the replacement probability for the larger mutation-effect size is close to that for the smaller mutation-effect size. Assuming that mutations of small effect sizes appear more frequently than mutations of large effect sizes [17, 75], it should be expected that early stages of adaptation of the subpopulations that exchange migrants relatively frequently are led by mutations of small effect sizes. At later stages, however, mutations of larger effect sizes are more likely to spread in the population, as is also the case in the absence of migration. The results obtained in the one-locus model were found to hold for the two-locus model as well. Here, the replacement probability at the locus with the larger allele-effect size was found to be larger than that at the locus with the smaller allele-effect size. This suggests that one locus may tend to drive local adaptation, as was also suggested in Ref. [17] ('concentrated genetic architecture'). However, the difference between the replacement probabilities at the two diverged loci decreases as adaptation progresses, and vanishes in late stages of adaptation. This suggests that concentrated genetic architectures may not be obtained if adaption is initialised from standing genetic variation.

This thesis provides a number of advances in understanding how different demographies, mating patterns, and selection contribute to shaping the patterns of genetic variation in natural populations. As discussed above, the amount of genetic variation can be critical for the establishment and maintenance of species in environments they are exposed to, or in new environments they colonise. If environmental conditions are different between the different habitats of a given species, its populations may exhibit adaptive divergence, the process that may eventually lead to a complete reproductive isolation between the different populations, that is, speciation. An example species that is currently ongoing speciation is the snail *Littorina saxatilis* that has two divergent ecotypes formed [6, 10]. This makes *Littorina saxatilis* a suitable species for studying speciation at work and hence gaining a better understanding of the mechanisms that allow for speciation to occur [6, 7, 9, 10, 155]. For example, does speciation occur thanks to standing genetic variation, or due to new mutations? In the latter case, what are the typical mutation

effect-sizes, and at which frequency should mutations appear in order for divergence of ecotypes to be possible? Does assortative mating (mate preference) facilitate divergence? *L. saxatilis* is likely in an initial stage of speciation, and genome-wide sequences sampled from its divergent ecotypes should provide a valuable insight into the process of speciation [8, 10, 155]. But in order to be able to interpret empirical data, further theoretical work is required. To this end, it is necessary to build a model that accounts for genome-wide patterns of genetic variation with some regions assumed to be subject to selection, and others being neutral. Related modelling work that includes both neutral, and regions under selection can be found in Ref. [110]. The model should further incorporate earlier findings concerning the life history of *L. saxatilis*, such as the colonisation of new areas from a large source population [126, 128], multiple paternity [15, 127] with a possibility for mate-preferences based on mates' sizes [156], but also different selection pressures in different areas over the species' geographic distribution (wave-exposed, crab-exposed, hybrid zones) [6]. Using such a model, it will be possible to contrast the patterns expected when adaptation is initialised with or without standing genetic variation. In the former case, multiple paternity is expected to promote the extent of standing genetic variation that founders introduce to new areas [II]. But the effect of divergent selection on genome-wide regions in the presence of multiple paternity or mate preferences is yet to be understood. In the near future, a spatially-explicit genome-wide model described above will be built and analysed in collaboration with the Linneaus Centre for Marine Evolutionary Biology (CeMEB).

Further work will also be done to understand better the interplay between sexual and asexual reproduction in species that have both modes of reproduction. As mentioned in this thesis, limited dispersal capabilities of propagules contribute to increasing genotypic variation on the long run, which makes it difficult to distinguish clones from sexual recruits using genetic sequences. The problem of inferring the rate of clonal reproduction is difficult even in well-mixed populations [65, 148]. Therefore, further theoretical efforts are needed to make it possible to estimate the rate of clonality in a given species. The future work will also aim to understand the consequences on the evolution of species' reproductive strategies (clonal in relation to sexual reproduction) due to possible differences in dispersal capabilities of sperms, eggs, and asexual propagules. The starting analysis will consider spatially and temporally homogeneous environments. This analysis will be extended to include spatially heterogeneous environments with different selection pressures over species' distributions. The modelling will be guided by empirical data obtained within CeMEB in experimental studies of species that reproduce both

sexually and asexually [157]. The modelling results will then be used to guide future genome-wide sampling of species with both modes of reproduction. The aim on the long term is to improve the existing methods for interpreting empirical data, and to understand better causes and consequences of the evolutionary processes.

# A

# Moments of frequency spectra of SNPs

The frequency spectrum of SNPs is an important set of observables because it provides an information about the underlying population-size history [28, 113]. The expressions for the first two moments are necessary for constructing tests of neutrality based on site frequency spectra, such as Tajima's $D$ [32].

The unfolded frequency spectrum of SNPs is determined by the number of mutations $\xi_i$ carried by $i$ individuals in a sample of size $n$ ($i = 1, \ldots, n-1$). It is common to refer to $\xi_i$ as the number of counts $i$. The folded frequency spectrum is related to the unfolded one according to Eq. (3.1) in Chapter 3. The total number of SNPs ($S_n$) is given by $S_n = \sum_{i=1}^{n-1} \xi_i$. It was shown in Ref. [114] how $\langle \xi_i \rangle$ and $\langle \xi_i \xi_j \rangle$ can be computed in terms of the moments of branch lengths $\tau_i$ (scaled by the effective population size) of gene genealogies. The corresponding expressions are [114]:

$$\langle \xi_i \rangle = \frac{\theta}{2} \sum_{k=2}^{n} k p(k,i) \langle \tau_k \rangle \ , \tag{A.1}$$

$$\begin{aligned} \langle \xi_i \xi_j \rangle = {} & \delta_{i,j} \sum_{k=2}^{n} k p(k,i) \left( \frac{\theta}{2} \langle \tau_k \rangle + \frac{\theta^2}{4} \langle \tau_k^2 \rangle \right) \\ & + \frac{\theta^2}{4} \left( \sum_{k=2}^{n} k(k-1) p(k,i;k,j) \langle \tau_k^2 \rangle \right. \\ & \left. + \sum_{k<m}^{n} km \left( \, p\,(k,i;m,j) + p\,(k,j;m,i) \, \right) \langle \tau_k \tau_m \rangle \right), \end{aligned} \tag{A.2}$$

where $\delta_{ij} = 1$ for $i = j$, and $\delta_{ij} = 0$ otherwise. The terms $p(k, i)$, $p(k, i; k, j)$, $p(k, i; m, j)$ appearing in Eqs. (A.1)-(A.2) come from Ref. [114]. Furthermore, $\theta = 2\mu N$ is the scaled mutation rate, $N$ being the (effective) population size, and $\mu$ the mutation rate per sequence, individual, generation. Note that in a population of constant size, one has $\langle \xi_i \rangle = \theta / i$ [114]. The expression for the second moment under a constant population size is also known [114]. By contrast, the expressions for the first two moments of the site frequency spectrum under varying population sizes are in general unknown. But they can be obtained using the results of Ref. [1]. The expressions for the first two moments under piecewise constant demographies with at most two population-size changes are given in Ref. [I]. The expression for the first moment derived in Ref. [I] agrees with the corresponding result reported in Ref. [28].

In Ref. [I], the expression for the first moment was used to infer the parameters of the underlying Human demographies based on empirical data gathered in the 1000 Genomes Project [43]. The expressions for the first two moments of the folded and unfolded site frequency spectra under piecewise constant demographies were used in Ref. [I] to define demography-adjusted tests. This is further discussed in Chapter 3.

# B

# Deterministic approximation for a model of adaptation

In this appendix, a deterministic approximation of the model introduced in Chapter 6 is analysed. The deterministic approximation, valid in the limit of infinite population sizes, serves to qualitatively understand the interplay between migration and divergent selection. Furthermore, it simplifies the analysis of the model. The effect of random genetic drift is discussed in Chapter 6 in the main text.

The model is constructed as follows. The population is assumed to consist of two diploid subpopulations of equal size. The subpopulations are subject to opposing environments, and hence experience divergent selection. The generations are assumed to be discrete and non-overlapping. The lifecycle of individuals is modelled in the following order: migration, mating, selection. Individuals migrate to the neighbouring subpopulation with the probability $m$ per generation, individual, subpopulation. Mating occurs locally, within each subpopulation, and it is assumed to be random. In the model, the number of gametes that an individual contributes to the next generation in a given subpopulation is proportional to the fitness of the individual relative to the average fitness of all individuals in the subpopulation (see below). The optimal phenotypes in the subpopulations (denoted by $\theta^{(1)}$, $\theta^{(2)}$) are assumed to be symmetric around zero ($\theta^{(1)} = -\theta^{(2)}$). For simplicity, in the analyses below, as well as in the main text $\theta^{(1)}$ is set to $\theta^{(1)} = 2$. The phenotype of an individual is assumed to depend on its genotype at one (or two) loci targeted by selection. Each locus is characterised by a pair of allele-effect sizes, and the phenotype is equal to the sum of the pair of allele-effect sizes at the locus (or loci) targeted by selection. In the two-locus model, a

gamete is a product of recombination between the parental chromosomes with the probability $r$ per generation, gamete, individual, subpopulation. When $r = 0$, the two-locus model reduces to the one-locus model. In the analyses outlined below, a mutation of a given effect size is assumed to alter one allele in the population, and further mutations are neglected (but a model that includes recurrent mutations was also simulated, see Fig. 6.1).

This appendix is organised as follows. In Section B.1, the one-locus model is analysed. The two-locus model is covered in Section B.2.

## B.1 One-locus model

To understand how a population under the model presented above progresses from one to another polymorphism, the analysis is performed as follows. The population is assumed to be initialised with allele-effect sizes $X$, and $-X$ (where $0 \leq X \leq 1$). Under this initial condition, and assuming that the subpopulations are of infinite size, it can be shown that the population inevitably establishes a stable dimorphism, denoted by $\{X, -X\}$ below (see also [76]). When the dimorphism $\{X, -X\}$ is established, it is assumed that one allele of size $X$ in the subpopulation with the positive optimal phenotype experiences a mutation of effect size $\epsilon$ (where $0 < \epsilon < 1 - X$; this condition assures that the mutation is beneficial in this subpopulation). In terms of a deterministic approximation, a small proportion of alleles of effect size $X$ is substituted by (the same) proportion of alleles of effect size $X + \epsilon$, so that the population stays in vicinity of the initial steady dimorphism $\{X, -X\}$. Further mutations are neglected. The question is: how does the mutation influence the dynamics of the population? To which stable state does the population relax?

Under a deterministic approximation of the model, the frequencies $p_{i;\tau}^{(k)}$ $(i = 1, 2, 3)$ of allele effect sizes $x_i$ (where $x_1 = X$, $x_2 = -X$, and $x_3 = X + \epsilon$) in subpopulation $k = 1, 2$ are expected to evolve from generation $\tau$ to generation $\tau + 1$ according to:

$$p_{i;\tau+1}^{(1)} = \frac{1}{w_\tau^{(1)}} \Bigg\{ \Big[ (1-m) \Big( p_{i;\tau}^{(1)} \Big)^2 + m \Big( p_{i;\tau}^{(2)} \Big)^2 \Big] w_{i|i}^{(1)} + \sum_{\substack{j=1 \\ j \neq i}}^{3} \Big[ (1-m) p_{i;\tau}^{(1)} \, p_{j;\tau}^{(1)} + m p_{i,\tau}^{(2)} \, p_{j;\tau}^{(2)} \Big] w_{i|j}^{(1)} \Bigg\} , \tag{B.1}$$

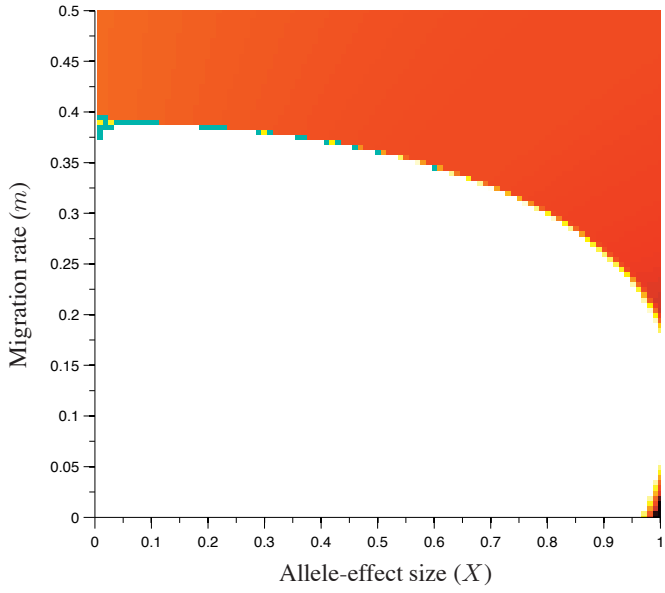


Figure B.1: Same as in Fig. 6.2, but here light blue denotes the region where no steady state was found within the stopping condition used. Numerical calculations were done by Anna Emanuelsson (Master thesis [149]).

$$p^{(2)}_{i;\tau+1} = \frac{1}{W^{(2)}_{\tau}} \Big\{ \Big[ (1-m) \Big( p^{(2)}_{i;\tau} \Big)^2 + m \Big( p^{(1)}_{i;\tau} \Big)^2 \Big] w^{(2)}_{i|i} + \sum_{\substack{j=1 \\ j \neq i}}^{3} \Big[ (1-m) p^{(2)}_{i;\tau} p^{(2)}_{j;\tau} + m p^{(1)}_{i;\tau} p^{(1)}_{j;\tau} \Big] w^{(2)}_{i|j} \Big\} \ . \tag{B.2}$$

Here $w^{(k)}_{i|j}$ denotes the fitness of the parental single-locus genotype $i|j$ that consists of a pair of allele-effect sizes $x_i$, and $x_j$, in subpopulation $k$

$$w^{(k)}_{i|j} = \mathrm{e}^{-\frac{(x_i + x_j - \theta^{(k)})^2}{2\sigma^2}} \ , \text{ for } i,j = 1,2,3 \ . \tag{B.3}$$

In Eq. (B.3), $\sigma$ is a selection parameter that is assumed to be the same in the subpopulations. As $\sigma$ increases, selection strength decreases, and vice versa. Furthermore,

$$W^{(1)}_{\tau} = \sum_{i=1}^{3} \sum_{j=1}^{3} \Big[ (1-m) p^{(1)}_{i;\tau} p^{(1)}_{j;\tau} + m p^{(2)}_{i;\tau} p^{(2)}_{j;\tau} \Big] w^{(1)}_{i|j} \tag{B.4}$$

stands for the average parental phenotype in subpopulation $k = 1$. The corresponding expression for the subpopulation $k = 2$ is obtained by exchanging superscripts $^{(1)}$ and $^{(2)}$ in Eq. (B.4).

In order to find the steady-state frequencies of the allele-effect sizes $X$, $-X$, and $X + \epsilon$ in the two subpopulations it is necessary to solve the system of Eqs. (B.1)-(B.2). Because it holds that $\sum_{i=1}^{3} p^{(k)}_{i;\tau} = 1$, it follows that the dynamics is fully determined by a set of four recursive equations. The stability of the steady states found can be deduced upon evaluating the leading eigenvalue of the stability matrix $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} \frac{\partial p^{(1)}_{1;\tau+1}}{\partial p^{(1)}_{1;\tau}} & \frac{\partial p^{(1)}_{1;\tau+1}}{\partial p^{(2)}_{1;\tau}} & \frac{\partial p^{(1)}_{1;\tau+1}}{\partial p^{(1)}_{2;\tau}} & \frac{\partial p^{(1)}_{1;\tau+1}}{\partial p^{(2)}_{2;\tau}} \\ \frac{\partial p^{(2)}_{1;\tau+1}}{\partial p^{(1)}_{1;\tau}} & \frac{\partial p^{(2)}_{1;\tau+1}}{\partial p^{(2)}_{1;\tau}} & \frac{\partial p^{(2)}_{1;\tau+1}}{\partial p^{(1)}_{2;\tau}} & \frac{\partial p^{(2)}_{1;\tau+1}}{\partial p^{(2)}_{2;\tau}} \\ \frac{\partial p^{(1)}_{2;\tau+1}}{\partial p^{(1)}_{1;\tau}} & \frac{\partial p^{(1)}_{2;\tau+1}}{\partial p^{(2)}_{1;\tau}} & \frac{\partial p^{(1)}_{2;\tau+1}}{\partial p^{(1)}_{2;\tau}} & \frac{\partial p^{(1)}_{2;\tau+1}}{\partial p^{(2)}_{2;\tau}} \\ \frac{\partial p^{(2)}_{2;\tau+1}}{\partial p^{(1)}_{1;\tau}} & \frac{\partial p^{(2)}_{2;\tau+1}}{\partial p^{(2)}_{1;\tau}} & \frac{\partial p^{(2)}_{2;\tau+1}}{\partial p^{(1)}_{2;\tau}} & \frac{\partial p^{(2)}_{2;\tau+1}}{\partial p^{(2)}_{2;\tau}} \end{bmatrix} . \tag{B.5}$$

The steady states can be found by iterating the dynamics of the system (Eqs. (B.1)-(B.2)) until the steady-state condition within a predetermined precision is met (see Chapter 6 in the main text). This was done for a range of values of the migration rate $m$, and allele-effect size $X$. However, for a small number of parameter values tested, no steady state was found within the precision used (see light blue regions

depicted in Fig. B.1). When the steady state was found, its stability was checked upon computing numerically the eigenvalues of the stability matrix. The eigenvalues showed that the steady states found using the iterative method were stable. However, the leading eigenvalues for the steady states found in close vicinity of the light-blue regions depicted in Fig. B.1 were very close to unity. This suggests that the steady states in the light-blue regions (Fig. B.1) are only marginally stable, and hence a long time is needed for the system to achieve them (the iterative method was allowed to search for the steady state up to a maximum of $10^6$ generations).

Fig. B.1 further shows that when migration is not too frequent, a mutation of size $\epsilon$ manages to spread in the population. In fact, it causes the population to escape the initial dimorphism $\{X, -X\}$, and relax into a new stable dimorphism $\{X + \epsilon, -X\}$. This is referred to in the main text as the replacement of $X$ by $X + \epsilon$. When $\epsilon$ is beneficial in the subpopulation where it is introduced, i. e. when $X \leq 1 - \epsilon$, the dimorphism $\{X + \epsilon, -X\}$ is established below a particular critical value of migration rate. For larger values of $X$, the mutation $\epsilon$ 'overshoots' the optimum; but under a weak selection, and provided that migration is neither too rare, nor too frequent, the mutation can spread in the population despite 'overshooting' (compare Fig. B.2 to Fig. B.3). The effect of 'overshooting' is not further discussed here.

Finally, recall that the analysis above assumes that the subpopulations are of infinite sizes. For finite population sizes the effect of random genetic drift needs to be taken into account. In order to estimate the effect of random genetic drift, one can use the results of the stability analysis of the steady states found (in analogy to the approximation suggested in Ref. [76]). Namely, in the region where $X + \epsilon$ (deterministically) replaces $X$, one can compute an 'unstable-state diversification coefficient' by subtracting unity from the leading eigenvalue of the stability matrix evaluated at the unstable dimorphism $\{X, -X\}$ (in analogy to the 'diversification coefficient' defined in Ref. [76]). Similarly, one can estimate a 'stable-state diversification coefficient' by subtracting from unity the leading eigenvalue of the stability matrix evaluated at the stable dimorphism $\{X + \epsilon, -X\}$. Following the approach in Ref. [76], the diversification coefficients (corresponding to the stable, and to the unstable dimorphism) can be used to estimate the probability that $X + \epsilon$ replaces $X$. For the results presented in the main text, two approximations were made by using one, or the other diversification coefficient in place of the selection coefficient in Kimura's result (2.11) [77, 78] (Section 2.4). The two approximations agree well with the simulation results (Fig. 6.3). But in the beginning of adaptation, the agreement is bet-

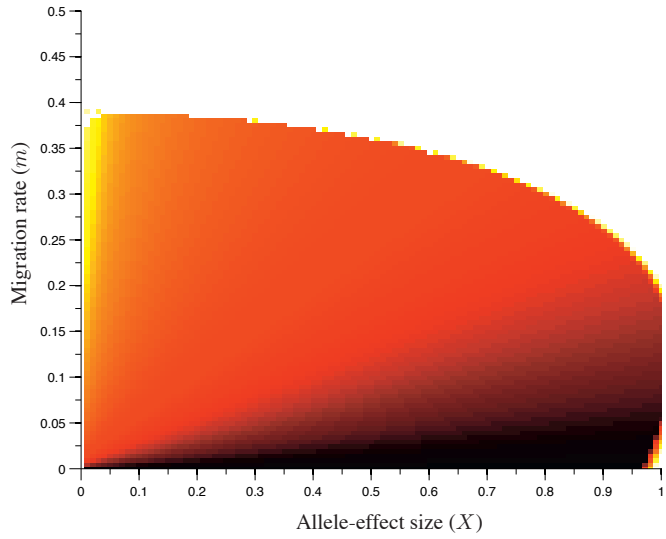


Figure B.2: Same as in Fig. 6.2, but for the allele-effect size $X + \epsilon$. Numerical calculations were done by Anna Emanuelsson (Master thesis [149])

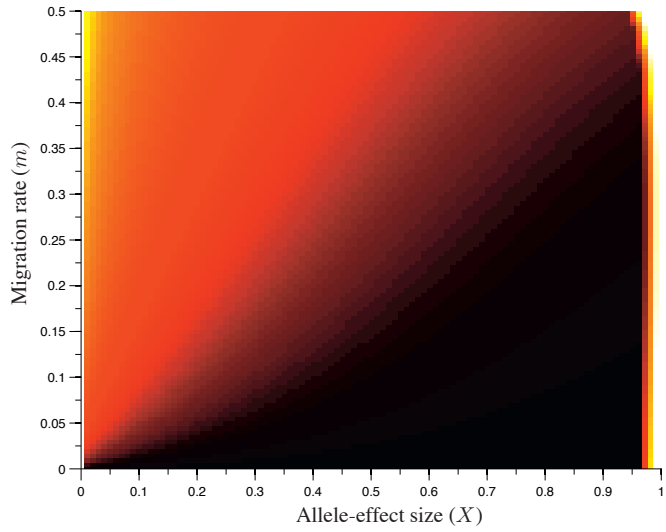


Figure B.3: Same as in Fig. B.2, but for a stronger selection: selection parameter $\sigma = 1.5$. Numerical calculations were done by Anna Emanuelsson (Master thesis [149]).

ter with the approximation made using the unstable-state diversification coefficient, whereas the opposite is true for latter stages of adaptation.

The results concerning the one-locus model are discussed in more detail in Chapter 6 in the main text.

## B.2 Two-locus model

In the two locus model, the population is initialised with allele-effect sizes $Y$, and $-Y$ at one locus, and with $Y + \alpha$, and $-Y - \alpha$ at the other locus. Here it is assumed that $0 \leq Y \leq 0.5$, and $0 \leq \alpha$. The first, and the second locus are referred to as the locus of smaller, and larger effect size, respectively. Similarly to the approach used in the one-locus model, the population is (before introducing a mutation) allowed to relax into a stable steady state. For the parameter values tested (Fig. 6.4), the population in the stable steady state is dimorphic at both loci (results not shown). Thereafter, the effect of a mutation of size $\epsilon$ is analysed in two cases: the mutation is assumed to land either on the locus of smaller allele-effect size ($Y$), or on the locus of larger allele-effect size ($Y + \alpha$).

In both cases, the dynamics of the system is tracked until the population relaxes into a new (or the old) stable steady state.

The dynamics in the two-locus model is more complex than in the one-locus model because here more possible genotypes appear in the population. The genotype with allele-effect sizes $y_i$, and $y_a$ at one chromosome, and $y_j$ and $y_b$ at the other is labeled by $i, a|, j, b$. Here $i, a$, and $j, b$ represent the corresponding two-locus haplotypes. The indexes $i$, and $j$ serve to distinguish between allele-effect sizes $y_i$ and $y_j$ at the first locus, whereas $a$, and $b$ serve to distinguish between allele-effect sixes $y_a$, and $y_b$ at the second locus. In the case a mutation lands on the first locus (that has alleles of smaller affect size), the indexes $i$, and $j$ take values $i, j = 1, 2, 3$, where $y_1 = Y$, $y_2 = -Y$, and $y_3 = Y + \epsilon$. Similarly, the indexes $a$, and $b$ take values $a, b = 4, 5$, where $y_4 = Y + \alpha$, and $y_5 = -Y - \alpha$. A similar indexing (and the analysis described below) can be made for the case when a mutation lands on the allele of larger effect size. Therefore, this case is not explicitly described in this appendix.

Under a deterministic approximation, the frequency $p^{(1)}_{1,4;\tau+1}$ of a haplotype with allele-effect sizes $y_1 = Y$ at one locus, and $y_4 = Y + \alpha$ at the other is expected to depend on the genotype frequencies in generation $\tau$ according to

$$
\begin{aligned}
p^{(1)}_{1,4;\tau+1} = \frac{1}{W^{(1)}_\tau}\Big\{ & \Big[(1-m)\,(p^{(1)}_{1,4;\tau})^2 + m\,(p^{(2)}_{1,4;\tau})^2\Big]\omega^{(1)}_{1,4|1,4} + \\
& \Big[(1-m)\,p^{(1)}_{1,4;\tau}\,p^{(1)}_{1,5;\tau} + m\,p^{(2)}_{1,4;\tau}\,p^{(2)}_{1,5;\tau}\Big]\omega^{(1)}_{1,4|1,5} + \\
& \Big[(1-m)\,p^{(1)}_{1,4;\tau}\,p^{(1)}_{2,4;\tau} + m\,p^{(2)}_{1,4;\tau}\,p^{(2)}_{2,4;\tau}\Big]\omega^{(1)}_{1,4|2,4} + \\
& \Big[(1-m)\,p^{(1)}_{1,4;\tau}\,p^{(1)}_{3,4;\tau} + m\,p^{(2)}_{1,4;\tau}\,p^{(2)}_{3,4;\tau}\Big]\omega^{(1)}_{1,4|3,4} + \\
& +(1-r)\Big[(1-m)\,p^{(1)}_{1,4;\tau}\,p^{(1)}_{3,5;\tau} + m\,p^{(2)}_{1,4;\tau}\,p^{(2)}_{3,5;\tau}\Big]\omega^{(1)}_{1,4|3,5} + \\
& (1-r)\Big[(1-m)\,p^{(1)}_{1,4;\tau}\,p^{(1)}_{2,5;\tau} + m\,p^{(2)}_{1,4;\tau}\,p^{(2)}_{2,5;\tau}\Big]\omega^{(1)}_{1,4|2,5} + \\
& r\Big[(1-m)\,p^{(1)}_{1,5;\tau}\,p^{(1)}_{2,4\tau} + m\,p^{(2)}_{1,5;\tau}\,p^{(2)}_{2,4\tau}\Big]\omega^{(1)}_{1,5|2,4} + \\
& r\Big[(1-m)\,p^{(1)}_{1,5;\tau}\,p^{(1)}_{3,4;\tau} + m\,p^{(2)}_{1,5;\tau}\,p^{(2)}_{3,3;\tau}\Big]\omega^{(1)}_{1,5|3,4}\Big\}\,. \qquad \text{(B.6)}
\end{aligned}
$$

The frequency of this haplotype in the opposite subpopulation is obtained by exchanging superscripts $^{(1)}$ and $^{(2)}$ in Eq. (B.6). Note that in each subpopulation there are six possible haplotypes. The expressions for the evolution of the remaining haplotypes are similar to that given by Eq. (B.6), but with different indexes. For simplicity, these expressions are not shown here. In Eq. (B.6), $W^{(1)}_\tau$ and $W^{(2)}_\tau$ are the average fitnesses

of parental genotypes in the two subpopulations in generation $\tau$

$$W_{\tau}^{(1)} = \sum_{i,j=1}^{3} \sum_{a,b=4}^{5} \Big[ (1-m) p_{i,a;\tau}^{(1)} p_{j,b;\tau}^{(1)} + m p_{i,a;\tau}^{(2)} p_{j,b;\tau}^{(2)} \Big] \omega_{i,a|j,b}^{(1)} \ ,$$

$$W_{\tau}^{(2)} = \sum_{i,j=1}^{3} \sum_{a,b=4}^{5} \Big[ (1-m) p_{i,a;\tau}^{(2)} p_{j,b;\tau}^{(2)} + m p_{i,a;\tau}^{(1)} p_{j,b;\tau}^{(1)} \Big] \omega_{i,a|j,b}^{(2)} \ .$$

Here, the fitnesses of the two-locus diploid genotypes in subpopulation $k = 1, 2$ are computed according to

$$\omega_{i,a|j,b}^{(k)} = \mathrm{e}^{-\frac{\left( y_i + y_j + y_a + y_b - \theta^{(k)} \right)^2}{2\sigma^2}} \ .$$

The steady states of the system can be found iteratively (similarly to the approach used in the one-locus case). But here the dynamics is fully determined by a set of ten equations. This is because in each subpopulation, the frequency of one haplotype can be expressed in terms of the remaining five due to the constraint that the sum of the frequencies of all possible haplotypes is equal to unity. The stability of the steady states found can be determined using a stability matrix. The stability matrix is defined similarly as before (see Eq. (B.5)), but now it is a $10x10$ matrix and hence it is not written explicitly.

The stability analysis shows that for the parameters set in Fig. 6.4, a mutation of size $\epsilon$ is expected to spread in the population. Thus the population is expected to exit the initial dimorphism at the first locus, and relax into a new stable state. The new stable state is dimorphic and it consists of allele-effect sizes $Y + \epsilon$, and $-Y$. Thus, similarly to the one-locus case, the mutant allele is expected to replace its non mutated version.

In the main text it is shown and discussed how genetic drift influences the tendency of the mutation to spread in the population (Fig. 6.4). Furthermore, the differences between the effect of a mutation that lands on the locus of smaller effect size ($Y$), and the effect of this mutation upon landing on the locus of larger effect size ($Y + \alpha$, $\alpha > 0$) are discussed in Chapter 6 in the main text.

# Bibliography


[1] A. Eriksson, B. Mehlig, M. Rafajlović and S. Sagitov, *The total branch length of sample genealogies in populations of variable size*, Genetics **186**, 601–611 (2010).

[2] E. Schaper, A. Eriksson, M. Rafajlović, S. Sagitov and B. Mehlig, *Linkage disequilibrium under recurrent bottlenecks*, Genetics **190**, 217–229 (2012).

[3] M. Rafajlović, *Genetic variation in structured populations*. Licentiate thesis, Department of Physics, University of Gothenburg, Gothenburg, Sweden, 2012.

[4] M. Pagel, *Inferring the historical patterns of biological evolution*, Nature **401**, 877–884 (1999).

[5] U. Kutschera and K. J. Niklas, *The modern theory of biological evolution: an expanded synthesis*, Naturwissenschaften **91**, 255–276 (2004).

[6] K. Johannesson, *Parallel speciation: a key to sympatric divergence*, Trends in Ecology & Evolution **16**, 148–153 (2001).

[7] R. Butlin, A. Debelle, C. Kerth, R. R. Snook, L. W. Beukeboom, R. F. Castillo Cajas, W. Diao, M. E. Maan, S. Paolucci, F. J. Weissing, van de Zande. L. *et al.*, *What do we need to know about speciation?*, Trends Ecol Evol **27**, 27–39 (2012).

[8] R. K. Butlin, *Population genomics and speciation*, Genetica **138**, 409–418 (2008).

[9] R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin *et al.*, *Hybridization and speciation*, Journal of Evolutionary Biology **26**, 229–246 (2013).

[10] O. Seehausen, R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. A. Hohenlohe, C. L. Peichel, G.-P. Saetre, C. Bank, Å. Brännström *et al.*, *Genomics and the origin of species*, Nature Reviews Genetics **15**, 176–192 (2014).

[11] R. A. Fisher, *The genetical theory of natural selection*. Clarendon, Oxford, 1930.

[12] H. Muller, *Some genetic aspects of sex*, The American Naturalist **66**, 118–138 (1932).

[13] J. Arjan, G. M. de Visser and S. F. Elena, *The evolution of sex: empirical insights into the roles of epistasis and drift*, Nature Reviews Genetics **8**, 139–149 (2007).

[14] L. Ross, N. B. Hardy, A. Okusu and B. B. Normark, *Large population size predicts the distribution of asexuality in scale insects*, Evolution **67**, 196–206 (2013).

[15] M. Panova, J. Boström, T. Hofving, T. Areskoug, A. Eriksson, B. Mehlig, T. Mäkinen, C. André and K. Johannesson, *Extreme female promiscuity in a non-social invertebrate species*, PLoS ONE **5**, e9640 (2010).

[16] S. Hintz Saltin, *Mate choice and its evolutionary consequences in intertidal snails (Littorina spp.)*. PhD thesis, Department of Biological and Environemntal Sciences - Tjärnö, University of Gothenburg, Strömstad, Sweden, 2013.

[17] S. Yeaman and M. J. Whitlock, *The genetic architecture of adapttation under migration-selection balance*, Evolution **65**, 1897–1911 (2011).

[18] C. Darwin and A. Wallace, *On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection*, J. Proc. Linn. Soc. Lond. **3**, 45–63 (1858).

[19] J. M. Smith and J. Haigh, *The hitch-hiking effect of a favorable gene*, Genet, Res., Camb. **23**, 23–35 (1974).

[20] J. C. Fay and C.-I. Wu, *Hitchhiking under positive Darwinian selection*, Genetics **155**, 1405–1413 (2000).

[21] D. L. Hartl and A. G. Clark, *Principles of Population Genetics*. Sinauer Associates, 1998.

[22] K. Johannesson, *Evolution in Littorina: ecology matters*, Journal of Sea Research **49**, 107–117 (2003).

[23] P. Nosil, *Speciation with gene flow could be common*, Molecular Ecology **17**, 2103–2106 (2008).

[24] K. Johannesson, M. Panova, P. Kemppainen, C. André, E. Rolán-Alvarez and R. K. Butlin, *Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation*, Philosophical Transactions of the Royal Society B: Biological Sciences **365**, 1735–1747 (2010).

[25] S. Wright, *Evolution in Mendelian populations*, Genetics **16**, 97–159 (1931).

[26] P. Moran, *Random processes in genetics.*, Proc. Cambridge Philos. Soc. **54**, 60–71 (1958).

[27] J. M. Akey, M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson and L. Kruglyak, *Population history and natural selection shape patterns of genetic variation in 132 genes*, PLoS Biology **2**, e286 (2004).

[28] G. T. Marth, E. Czabarka, J. Murvai and S. T. Sherry, *The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations*, Genetics **166**, 351–372 (2004).

[29] J. E. Stajich and M. W. Hahn, *Disentangling the effects of demography and selection in human history*, Mol. Biol. Evol. **22**, 63–73 (2005).

[30] C. S. Carlson, D. J. Thomas, M. A. Eberle, J. E. Swanson, R. J. Livingston, M. J. Rieder and D. A. Nickerson, *Genomic regions exhibiting positive selection identified from dense genotype data*, Genome Research **15**, 1553–1565 (2005).

[31] S. R. Grossman, K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, M. Cabili *et al.*, *Identifying recent adaptations in large-scale genomic data*, Cell **152**, 703 – 713 (2013).

[32] F. Tajima, *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism*, Genetics **123**, 585–595 (1989).

[33] J. Kingman, *The coalescent*, Stoch. Proc. Appl. **13**, 235–248 (1982).

[34] J. F. C. Kingman, *Essays in statistical science*, Journal of Applied Probability **19**, 27–43 (1982).

[35] R. R. Hudson, *Gene genealogies and the coalescent process*, Oxford Surveys in Evolutionary Biology **7**, 1–44 (1990).

[36] K. Janson, *Selection and migration in two distinct phenotypes of Littorina saxatilis in Sweden*, Oecologia **59**, 58–61 (1983).

[37] S. Ramachandran, O. Deshpande, C. Roseman, N. Rosenberg, M. Feldman and L. Cavalli-Sforza, *Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa*, Proceedings of the National Academy of Sciences of the United States of America **102**, 15942–15947 (2005).

[38] H. Liu, F. Prugnolle, A. Manica and F. Balloux, *A geographically explicit genetic model of worldwide human-settlement history*, Am. J. Hum. Genet. **79**, 230–237 (2006).

[39] K. Tanabe, T. Mita, T. Jombart, A. Eriksson, S. Horibe, N. Palacpac, L. Ranford-Cartwright, H. Sawai, N. Sakihama, H. Ohmae, M. Nakamura *et al.*, *Plasmodium falciparum accompanied the human expansion out of Africa*, Curr Biol **20**, 1283–9 (2010).

[40] J. Pujolar, S. Vicenzi, L. Zane, D. Jesensek, G. De Leo and A. Crivelli, *The effect of recurrent floods on genetic composition of marble trout populations*, PLoS ONE **6**, e23822 (2011).

[41] P. Sjödin, I. Kaj, S. Krone, M. Lascoux and M. Nordborg, *On the meaning and existence of an effective population size*, Genetics **169**, 1061–1070 (2005).

[42] J. Wakeley and O. Sargsyan, *Extensions of the coalescent effective population size*, Genetics **181**, 341–345 (2009).

[43] McVean *et al.*, *An integrated map of genetic variation from 1092 human genomes*, Nature **491**, 56–65 (2012).

[44] I. G. Paterson, V. Partridge and J. Buckland-Nicks, *Multiple paternity in Littorina obtusata (Gastropoda, Littorinidae) revealed by microsatellite analyses*, The Biological Bulletin **200**, 261–267 (2001).

[45] K. Trontti, N. Thurin, L. Sundström and S. Aron, *Mating for convenience or genetic diversity? Mating patterns in the polygynous ant Plagiolepis pygmaea*, Behavioral Ecology **18**, 298–303 (2007).

[46] A. Brante, M. Fernandez and F. Viard, *Microsatellite evidence for sperm storage and multiple paternity in the marine gastropod Crepiduda coquimbensis*, J. Exp. Mar. Biol. Ecol. **396**, 83–88 (2011).

[47] S. W. Coleman and A. G. Jones, *Patterns of multiple paternity and maternity in fishes*, Biological Journal of the Linnean Society **103**, 735–760 (2011).

[48] W. Ewens, *On the concept of the effective population size*, Theoretical Population Biology **21**, 373–378 (1982).

[49] A. Tatarenkov, L. Bergström, R. B. Jönsson, E. A. Serraõ, L. Kautsky and K. Johannesson, *Intriguing asexual life in marginal populations of the brown seaweed Fucus vesiculosus*, Molecular Ecology **14**, 647–651 (2005).

[50] C. Eckert, K. Lui, K. Bronson, P. Corradini and A. Bruneau, *Population genetic consequences of extreme variation in sexual and clonal reproduction in an aquatic plant*, Molecular Ecology **12**, 331–344 (2003).

[51] C. G. Eckert, *The loss of sex in clonal plants*, Evolutionary Ecology **15**, 501–520 (2002).

[52] M. Kearney, *Why is sex so unpopular in the Australian desert?*, Trends in Ecology & Evolution **18**, 605–607 (2003).

[53] M. Kearney, *Hybridization, glaciation and geographical parthenogenesis*, Trends in Ecology & Evolution **20**, 459–502 (2005).

[54] A. Kliber and C. Eckert, *Interaction between founder effect and selection during biological invasion in an aquatic plant*, Evolution **59**, 1900–1913 (2005).

[55] T. Kawecki, *Adaptation to marginal habitats*, Annual Review of Ecology, Evolution, and Systematics **39**, 321–342 (2008).

[56] R. C. Vrijenhoek and D. E. J. Parker, *Geographical parthenogenesis: General purpose genotypes and frozen niche variation*, in *Lost sex*, I. Schön, K. Marens and P. Van Dijk, eds., pp. 99–131. Springer Science, 2009.

[57] E. D. J. Parker, R. K. Selander, R. O. Hudson and L. J. Lester, *Genetic diversity in colonizing parthenogenetic cockroaches*, Evolution **31**, 836–842 (1977).

[58] R. Vrijenhoek, *Ecological differentiation among clones: The frozen niche variation model*, in *Population Biology and Evolution*, K. Wöhrmann and V. Loeschcke, eds., Proceedings in Life Sciences, pp. 217–231. Springer Berlin Heidelberg, 1984.

[59] Y. Peck, J. Yearsley and D. Waxman, *Explaining the geographic distributions of sexual and asexual populations*, Nature **391**, 889–892 (1998).

[60] P. Stenberg, M. Lundmark, S. Knutelski and A. Saura, *Evolution of clonality and polyploidy in a weevil system*, Molecular Biology and Evolution **20**, 1626–1632 (2003).

[61] E. Hörandl, *Geographical parthenogenesis: Opportunities for asexuality*, in *Lost Sex*, I. Schön, K. Martens and P. van Dijk, eds., pp. 161–186. Springer Netherlands, 2009.

[62] K. Johannesson, H. Forslund, N. A. Capetillo, L. Kautsky, D. Johansson, R. T. Pereyra *et al.*, *Phenotypic variation in sexually and asexually recruited individuals of the Baltic Sea endemic macroalga Fucus radicans: in the field and after growth in a common-garden*, BMC Ecology **12**, 1–8 (2012).

[63] D. Johansson, *Evolution of the brown algae Fucus radicans and F. vesiculosus in the Baltic Sea*. PhD thesis, Department of Biological and Environmental Sciences - Tjärnö, University of Gothenburg, Strömstad, Sweden, 2013.

[64] H. Baker, *Self-compatibility and establishment after 'long-distance' dispersal*, Evolution **9**, 347–349 (1955).

[65] B. O. Bengtsson, *Genetic variation in organisms with sexual and asexual reproduction*, J. Evol. Biol. **16**, 189–199 (2003).

[66] M. Kimura, *The neutral theory of molecular evolution*. Cambridge University Press, 1985.

[67] H. A. Orr, *The population genetics of adaptation: the distribution of factors fixed during adaptive evolution*, Evolution **52**, 935–949 (1998).

[68] J. Felsenstein, *The theoretical population genetics of variable selection and migration*, Annu. Rev. Genet. **10**, 253–280 (1976).

[69] T. Lenormand, *Gene flow and the limits to natural selection*, Trends in Ecology and Evolution **17**, 183–189 (2002).

[70] T. Naglyaki and Y. Lou, *Evolution under multi allelic migration-selection models*, Theoretical Population Biology **72**, 21–40 (2007).

[71] B. Star, R. J. Stoffels and H. G. Spencer, *Single-locus polymorphism in a heterogeneous two-deme model*, Genetics **176**, 1625–1633 (2007).

[72] T. Nagylaki and Y. Lou, *The dynamics of migration-selection models*, in *Tutorials in Mathematical Biosciences IV*, A. Friedman, ed., vol. 1922 of *Lecture Notes in Mathematics*, pp. 117–170. Springer Berlin Heidelberg, 2008.

[73] A. Akerman and R. Bürger, *The consequences of gene flow for local adaptation and differentiation: a two-locus two-deme model*, J. Math. Biol **68**, 1135–1198 (2014).

[74] L. Geroldinger and R. Bürger, *A two-locus model of spatially varying stabilizing or directional selection on a quantitative trait*, Theoretical Population Biology **94**, 10–41 (2014).

[75] C. K. Griswold, *Gene flow's effect on the genetic architecture of a local adaptation and its consequences for QTL analyses*, Heredity **96**, 45–453 (2006).

[76] S. Yeaman and S. P. Otto, *Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift*, Evolution **65**, 2123–2129 (2011).

[77] M. Kimura, *Some problems of stochastic processes in genetics*, Ann. Math. Statist. **28**, 882–901 (1957).

[78] M. Kimura, *On the probability of fixation of mutant genes in a population*, Genetics **47**, 713–719 (1962).

[79] W. J. Ewens, *Mathematical population genetics*. Springer, Berlin, 1979.

[80] G. H. Hardy, *Mendelian proportions in a mixed population*, Science **28**, 49–50 (1908).

[81] W. Weinberg, *Über den nachweis der vererbung beim menschen*. 1908.

[82] R. Griffiths and S. Tavaré, *Sampling theory for neutral alleles in a varying environment*, Phil. Trans. Roy. Soc. Lon. B **344**, 403–410 (1994).

[83] F. Austerlitz, B. JungMuller, B. Godelle and P. Gouyon, *Evolution of coalescence times, genetic diversity and structure during colonization*, Theoretical Population Biology **51**, 148–164 (1997).

[84] D. Zivkovic and T. Wiehe, *Second-order moments of segregating sites under variable population size*, Genetics **180**, 341–357 (2008).

[85] J. Pitman, *Coalescents with multiple collisions*, The Annals of Probability **27**, pp. 1870–1902 (1999).

[86] S. Sagitov, *The general coalescent with asynchronous mergers of ancestral lines*, Journal of Applied Probability **36**, pp. 1116–1125 (1999).

[87] M. Möhle and S. Sagitov, *A classification of coalescent processes for haploid exchangeable population models*, The Annals of Probability **29**, pp. 1547–1562 (2001).

[88] J. Schweinsberg, *Coalescents with simultaneous multiple collisions*, Electron. J. Probab. **5**, 1–55 (2000).

[89] O. Sargsyan and J. Wakeley, *A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms*, Theoretical Population Biology **74**, 104 – 114 (2008).

[90] M. Birkner, J. Blath, M. Möhle, M. Steinrücken and J. Tams, *A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks*, ALEA **6**, 25–61 (2009).

[91] M. Kimura and J. F. Crow, *The number of alleles that can be maintained in a finite population*, Genetics **49**, 725–738 (1964).

[92] A. Bonin, D. Ehrich and S. Manel, *Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists*, Molecular Ecology **16**, 3737–3758 (2007).

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

[94] M. Kimura and T. Ohta, *Mutation and evolution at the molecular level.*, Genetics **73** (1973).

[95] T. Ohta and M. Kimura, *A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population.*, Genet Res **22**, 201–204 (1973).

[96] A. M. Valdes, M. Slatkin and N. B. Freimer, *Allele frequencies at microsatellite loci: The stepwise mutation model revisited*, Genetics **133**, 737–749 (1993).

[97] M. S. McPeek and T. P. Speed, *Modelling interference in genetic recombination*, Genetics **139**, 1031–1044 (1995).

[98] A. Eriksson and B. Mehlig, *On the effect of fluctuating recombination rates on the decorrelation of gene histories in the human genome*, Genetics **169**, 1175–1178 (2005).

[99] T. Ohta and M. Kimura, *Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population.*, Genetics **68**, 571–580 (1971).

[100] R. C. Griffiths, *Neutral 2-locus multiple allele models with recombination.*, Theoretical Population Biology **19**, 169–186 (1981).

[101] R. R. Hudson, *Properties of a neutral allele model with intragenetic recombination*, Theoretical Population Biology **23**, 183–201 (1983).

[102] G. McVean, *A genealogical interpretation of linkage disequilibrium*, Genetics **162**, 987 – 991 (2002).

[103] T. D. Price, A. Qvarnström and D. E. Irwin, *The role of phenotypic plasticity in driving genetic evolution*, Proc. R. Soc. Lond. B **270**, 1433–1440 (2003).

[104] S.-C. Park, D. Simon and J. Krug, *The speed of evolution in large asexual populations*, J. Stat. Phys. **138**, 381–410 (2010).

[105] J. B. S. Haldane, *A mathematical theory of natural and artificial selection, part V: Selection and mutation*, Mathematical Proceedings of the Cambridge Philosophical Society **23**, 838–844 (1927).

[106] A. J. Berry, J. W. Ajioka and M. Kreitman, *Lack of polymorphism on the Drosophila fourth chromosome resulting from selection*, Genetics **129**, 1111–1117 (1991).

[107] N. H. Barton, *The effect of hitch-hiking on neutral genealogies*, Genetics Research **72**, 123–133 (2000).

[108] A. Eriksson, P. Fernström and B. Mehlig, *An accurate model of genetic hitchhiking*, Genetics **178**, 439–451 (2008).

[109] J. Maynard Smith, *Sympatric speciation*, American Naturalist 637–650 (1966).

[110] J. L. Feder, S. P. Egan and P. Nosil, *The genomics of speciation-with-gene-flow*, Trends in Genetics **28**, 342–350 (2012).

[111] R. D. H. Barrett and D. Schluter, *Adaptation from standing genetic variation*, TRENDS in Ecology and Evolution **23**, 38–44 (2008).

[112] R. Nielsen, I. Hellmann, M. Hubisz, C. Bustamante and A. G. Clark, *Recent and ongoing selection in the human genome*, Nat. Rev. Genet. **8**, 857–868 (2007).

[113] N. Nawa and F. Tajima, *Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human*, Genes & Genetic Systems **83**, 353–360 (2008).

[114] Y. X. Fu, *Statistical properties of segregating sites*, Theoretical Population Biology **48**, 172 – 197 (1995).

[115] K. Simonsen, G. Churchill and C. Aquadro, *Properties of Statistical Tests of Neutrality for DNA Polymorphism Data*, Genetics **141**, 413–29 (1995).

[116] G. Achaz, *Frequency spectrum neutrality tests: One for all and all for one*, Genetics **183**, 249–258 (2009).

[117] F. Tajima, *The effect of change in population size on DNA polymorphism*, Genetics **123**, 597–601 (1989b) [`http://www.genetics.org/content/123/3/597.full.pdf+html`].

[118] R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark and C. Bustamante, *Genomic scans for selective sweeps using SNP data*, Genome Research **15**, 1566–1575 (2005).

[119] B. F. Voight, S. Kudaravalli, X. Wen and J. K. Pritchard, *A map of recent positive selection in the human genome*, PLoS Biol **4**, e72 (2006).

[120] R. Nielsen, *Estimation of population parameters and recombination rates from single nucleotide polymorphisms*, Genetics **154**, 931–942 (2000).

[121] A. A. Adams and R. R. Hudson, *Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms*, Genetics **168**, 1699–1712 (2004).

[122] D. Enard, P. W. Messer and D. A. Petrov, *Genome-wide signals of positive selection in human evolution*, Genome Res. **24**, 885–895 (2014).

[123] L. L. Cavalli-Sforza and M. W. Feldman, *The application of molecular genetic approaches to the study of human evolution*, Nature Genetics Supplement **33**, 266–275 (2003).

[124] S. Myers, C. Fefferman and N. Patterson, *Can one learn history from the allelic spectrum?*, Theoretical Population Biology **73**, 342–348 (2008).

[125] E. Mayr, *Animal species and evolution*. Harvard University Press, Cambridge, 1963.

[126] K. Janson, *Genetic drift in small and recently founded populations of the marine snail Littorina saxatilis*, Heredity **58**, 31–37 (1987).

[127] A. Eriksson, B. Mehlig, M. Panova, C. André and K. Johannesson, *Multiple paternity: determining the minimum number of sires of a large brood*, Molecular Ecology Resources **10**, 282–291 (2010).

[128] K. Johannesson and B. Johannesson, *Dispersal and population expansion in a direct developing marine snail (Littorina saxatilis) following a severe population bottleneck*, Hydrobiologia **309**, 173–180 (1995).

[129] M. Kimura and G. H. Weiss, *The stepping stone model of population structure and the decrease of genetic correlation with distance*, Genetics **49**, 561–576 (1964).

[130] M. Notohara, *The coalescent and the genealogical process in geographically structured population*, Journal of Mathematical Biology **29**, 59–75 (1990).

[131] N. Takahata and M. Slatkin, *Genealogy of neutral genes in two partially isolated populations*, Theoretical Population Biology **38**, 331 – 350 (1990).

[132] F. Balloux and L. Lehmann, *Random mating with a finite number of matings*, Genetics **165**, 2313–2315 (2003).

[133] A. Eriksson, B. Haubold and B. Mehlig, *Statistics of selectively neutral genetic variation*, Phys. Rev. E **65**, 040901 (2002).

[134] O. E. Gaggiotti, *Population genetic models of source-sink metapopulations*, Theoretical Population Biology **50**, 178–208 (1996).

[135] R. Cousens, C. Dytham and R. Law, *Dispersal in plants: a population perspective*. Oxford University Press, USA, 2008.

[136] I. Schön, K. Martens and P. Van Dijk, *Lost sex: The evolutionary biology of parthenogenesis*. Springer Verlag, 2009.

[137] M. Vallejo-Marin, M. Dorken and S. Barrett, *The ecological and evolutionary consequences of clonality for plant mating*, Annual Review of Ecology, Evolution, and Systematics **41**, 193–213 (2010).

[138] J. Fries, *Dynamics of sex ratio and genetics in populations with mixed sexual and asexual reproduction*, 2014.

[139] J. Albertsson and L. Bergström, *Undervattensvegetation i Holmöarnas naturreservat*. Länsstyrelsens Tryckeri, 2008.

[140] A. Ardehed *et al.* (in preparation).

[141] S. A. Krueger-Hadfield, J. E. Kübler and S. R. Dudgeon, *Reproductive effort of Mastocarpus papillatus (Rhodophyta) along the California coast*, Journal of Phycology **49**, 271–281 (2013).

[142] K. Johannesson, D. Johansson, K. H. Larsson, C. Huenchuñir, J. Perus, H. Forslund *et al.*, *Frequent clonality in Fucoids (Fucus radicans and Fucus vesiculosus; Fucales, Phaeophyceae) in the Baltic Sea*, Journal of Phycology **47**, 990–998 (2011).

[143] O. Eriksson, *Dynamics of genets in clonal plants*, Trends in Ecology and Evolution **8**, 313–316 (1993).

[144] A. Eriksson, F. Elías-Wolff and B. Mehlig, *Metapopulation dynamics on the brink of extinction*, Theoretical Population Biology **83**, 101–122 (2013).

[145] D. A. Carino and C. C. Daehler, *Genetic variation in an apomictic grass, Heteropogon contortus, in Hawaiian Islands*, Molecular Ecology **8**, 2127–2132 (1999).

[146] O. Paun, J. Greilhuber, E. M. Temsch and E. Hörandl, *Patterns, sources and ecological implications of clonal diversity in apomictic Ranunculus carpaticola (Ranunculus auricomus complex, Ranunculaceae)*, Molecular Ecology **15**, 897–910 (2006).

[147] J. A. Wilk, A. T. Kramer and M. V. Ashley, *High variation in clonal vs. sexual reproduction in populations of the wild strawberry, Fragaria virginiana (Rosaceae)*, Annals of Botany **104**, 1413–1419 (2009).

[148] F. Balloux, L. Lehmann and T. Meeûs, *The population genetics of clonal and partially clonal diploids*, Genetics **164**, 1635–1644 (2003).

[149] A. Emanuelsson, *A model for the evolution of local adaptation of a subdivided population*, 2014.

[150] F. Wang, *Analysis of models for the evolution of haplotypes under selection*, 2013.

[151] P. Nosil and J. L. Feder, *Genomic divergence during speciation: causes and consequences*, Phil. Trans. R. Soc. B **367**, 332–342 (2012).

[152] O. Savolainen, M. Lascoux and J. Merilä, *Ecological genomics of local adaptation*, Nature Reviews: Genetics **14**, 807–820 (2013).

[153] J. L. Feder, R. Gejji, S. Yeaman and P. Nosil, *Establishment of new mutations under divergence and genome hitchhiking*, Phil. Trans. R. Soc. B. **367**, 461–474 (2012).

[154] D. L. Theobald, *A formal test of the theory of universal common ancestry*, Nature **465**, 219–222 (2010).

[155] A. Westram, J. Galindo, M. A. Rosenblad, J. Grahame, M. Panova and R. Butlin, *Do the same genes underlie parallel phenotypic divergence in different Littorina saxatilis populations?*, Molecular Ecology n/a–n/a (2014).

[156] S. H. Saltin, H. Schade and K. Johannesson, *Preference of males for large females causes a partial mating barrier between a large and a small ecotype of Littorina fabalis (W. Turton, 1825)*, Journal of Molluscan Studies eyt003 (2013).

[157] K. Härnström, M. Ellegaard, T. J. Andersen and A. Godhe, *Hundred years of genetic structure in a sediment revived diatom population*, PNAS **108**, 4252–4257 (2011).