Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften an der Fakultät für Biologie der Ludwig-Maximilians-Universität München

Models of adaptation and speciation

Pleuni Simone Pennings
aus
Castricum, Niederlande
1975

Ich versichere hiermit ehrenwörtlich, dass die Dissertation von mir selbständig, ohne unerlaubte Beihilfe angefertigt ist.

Hiermit erkläre ich, dass ich mich anderweitig einer Doktorprüfung ohne Erfolg nicht unterzogen habe.

1. Gutachter: PD Dr. Joachim Hermisson
2. Gutachter: Prof. Dr. Wolfgang Stephan

Dissertation eingereicht am: 04.08.2006
Mündliche Prüfung am: 19.01.2007

Note

In this thesis I present the results from my doctoral research, which I have done between June 2003 and August 2006. Most of the work was done under the supervision of Joachim Hermisson at the Ludwig-Maximilians-Universität in Munich, Germany. Part of the work for chapter 4 was carried out under the supervision of Ulf Dieckmann at the International Institute for Applied Systems Analysis in Laxenburg, Austria.

Chapters 1, 2 and 3 of this thesis are closely related to each other and the result of an intense collaboration between Joachim Hermisson and myself. For chapter 1, I did parts of the conceptual work and model building, I did all of the simulations and contributed to the manuscript preparation. The analytical work and most of the manuscript preparation were done by Joachim.

For chapter 2, Joachim and I shared the conceptual work and the writing. I did the simulations and Joachim did the analytical work.

For chapter 3, the simulations are based on a program which was kindly provided by Yuseob Kim. I made changes to the program and added new parts. The analytical work was done by Joachim and myself. I did most of the writing.

Work for chapter 4 started in the summer of 2005, when I was working with Ulf Dieckmann at the IIASA. While in Laxenburg, I designed the model and derived the main results. Later, Joachim and Michael Kopp joined the project and contributed much to the conceptual and analytical work. The simulations were done by me. To write the code for the simulations, I used Ulf Dieckmanns code from his 1999 paper for reference. The writing of the manuscript was done by Michael and me.

The work in this thesis was supported by an Emmy Noether Grant to Joachim Hermisson and a grant from the Dutch Science Foundation to visit the IIASA for 3 months, to Pleuni Pennings.

Contents

1 Note 5

0 General introduction 9

0.1 About this introduction and the thesis 9

0.2 What is evolutionary biology? 9

0.3 What is theoretical evolutionary biology? 13

0.4 About chapter 1 (Soft sweeps 1) 14

0.5 About chapter 2 (Soft sweeps 2) 26

0.6 About chapter 3 (Soft sweeps 3) 30

0.7 About chapter 4 (Sympatric speciation) 33

1 Soft Sweeps I– Molecular population genetics of adaptation from standing genetic variation 39

1.1 Introduction 42

1.2 Model and Methods 44

1.3 Results 45

1.4 Dicussion 64

1.5 Acknowledgements 70

1.6 Appendix 71

2 Soft Sweeps II – Molecular population genetics of adaptation from recurrent mutation or migration 77

2.1 Introduction 80

2.2 Model and Methods 81

2.3 Results 82

2.4 Discussion 95

2.5 Acknowledgments 99

2.6 Supplemetary material 99

3 Soft Sweeps III – The signature of positive selection from recurrent mutation 105

3.1 Introduction . . . 108

3.2 Model and Methods . . . 110

3.3 Results . . . 113

3.4 Discussion . . . 131

3.5 Acknowledgments . . . 135

3.6 Appendix . . . 136

4 A one-locus model for sympatric speciation 141

4.1 Introduction . . . 144

4.2 The model . . . 145

4.3 Results . . . 151

4.4 Discussion . . . 162

4.5 Appendices . . . 167

5 Summary 171

6 Bibliography 175

7 CV 187

8 List of publications 189

9 Acknowledgments 191

Introduction

0.1 About this introduction and the thesis

It is probably impossible to write an introduction that is interesting and instructive for everyone who will have a look at this thesis. So please, don't be annoyed if parts of this introduction are abracadabra to you and on the other hand, please don't feel offended if it is much too easy. Each of the four chapters of this thesis is a paper (chapter 1 and 2 published, chapter 3 submitted and chapter 4 in preparation) and has a formal introduction. If you are a population geneticist, you may want to skip this introduction and jump to chapters 1, 2, and 3 immediately. If you are interested in competitive sympatric speciation you could start with chapter 4. In this general introduction I have tried to explain the topics of this thesis in such a way that also people outside my field can understand what the questions are that I worked on. I first spend two sections on evolutionary biology and theoretical evolutionary biology followed by four sections to explain the main questions and results of each of the chapters.

I hope you will enjoy reading this introduction.

0.2 What is evolutionary biology?

Two observations are central to evolutionary biology.
1. All species on earth are descended from a common ancestor and
2. Species tend to become adapted to their environment.
The fact that there are different species, that are all descended from one ancestor is because species sometimes *speciate*. Species are adapted to their environment because they evolve through the following mechanism: mutation creates variation, some variants produce more offspring than other variants and the

Speciation: the splitting of a species into two different species.

result is that the genetic composition of the species changes. Evolutionary biology is the science that tries to find the rules that govern both speciation and adaptation. The two main branches of evolutionary biology are often called macro-evolution (explaining the evolutionary relations between species) and micro-evolution (explaining evolution within a species, including adaptation). Knowledge of the rules of evolutionary biology can help us to understand the world as we observe it (Why are there so many species of beetles? Why does HIV evolve so fast?), and it can help us to make predictions to base decisions on (How long will it take before malaria is resistant against the new drug and what can we do to prevent that?).

There are many unsolved questions in evolutionary biology, which is not surprising given the complexity of of the subject and the fact that it is still a relatively young field of science. Why is the subject so complex? To see this, compare the following. A law of physics states that how much an object will speed up or slow down depends on its weight and the forces that work on it. Using this rule, I could calculate the movements of hanging objects in moving trains when I was in secondary school. The law may not always be exact, but it gives a good approximation for almost every moving object on earth. A rule in evolutionary biology states that a population will change in such a direction that its mean *fitness* will be increased (at least if the environment doesn't change), making it better adapted to its environment. Even though there are exceptions to this rule, many biologists believe it is correct most of the time. However, we can still hardly ever use this rule to make predictions about how populations will change. One problem is that it is hard to determine the direction and magnitude of the *selective forces* that work on a population. A second difficulty is that even if we know the forces, the reaction of a population to those forces depends largely on stochastic processes such as mutation. A stochastic process is a process in which it is impossible to predict what will happen next, at best one can know the probability of the next step being a certain event. Mutation is such a stochastic process; it is impossible to predict when or where mutations happen. It is therefore hard to predict how fast a population will change and in which of the possible directions. Even if a beneficial mutation occurs, it may get lost again (more about this in section 0.4).

Fitness: the average number of offspring of an individual.

Selective force: the word force may be misleading here. Selection favors a certain variant (mutant) when such a variant occurs. Mutation creates the variants completely independent of the selective force.

Evolutionary ecology is a subdiscipline of evolutionary biology and deals with the problem of determining the direction of the selective forces. The next example shows why it is so hard to determine this direction. A botanist may observe that taller plants produce more seeds than shorter plants, and expect that the taller plants are fitter. The tall plants should therefore increase in

	fertility (# seeds)	survival (prob. to survive to adulthood)	total fitness (expected # adult offspring)
tall plants	100	0.01	1
short plants	50	0.02	1

Figure 1: Total fitness is determined by different fitness components. In this example there are only two: fertility and survival. The tall plants produce more seeds but the seeds have low probability to survive. When only one fitness component is measured it is not possible to draw conclusions about total fitness. In chapter 4 of this thesis, the hermaphrodite individuals have three fitness components: male fertility, female fertility and survival.

frequency and the mean length of the population should increase. However, to be sure that selection actually favors tall plants we need to check first whether the seeds from the taller plants are not for some reason worse survivors than the seeds of the shorter plants, to check whether the total fitness of the tall plants is indeed higher (see Figure 1). It is also possible that the tall plants do not always produce more seeds. The tall plants may do better this year, but maybe not next year when there is less rain and the shorter plants have an advantage, for example, because they have longer roots. Maybe not in another time, but at another place, the shorter plants are better off. Imagine that our focus population would be on a small island and that there is also a large mainland population where short plants are fitter than tall plants (as in Figure 2). Seeds from short plants from the mainland population will continuously enter the island population by migration, so that the island population will always have short plants. In that case the selective force cannot win over migration. It is also hard to know whether selection actually favors the tallness of the plants. If there is a gene that affects both tallness and another characteristic that determines fitness, the correlation between “many seeds” and “being tall” can be a genetic coincidence (see Figure 3).

The things mentioned in the last paragraph are just some of many reasons why the botanist cannot conclude that selection favors tallness both now and in other times, both here and in other places. For selection to have an effect on a population it does not need to be there for ever, just sufficiently long or often. Also, it has to be strong enough. Selection can also prevail over migration, as long as there are not too many migrants. It needs enough beneficial mutations and not too many deleterious ones. And all this requires a substantial amount of luck as well (see section 0.4).

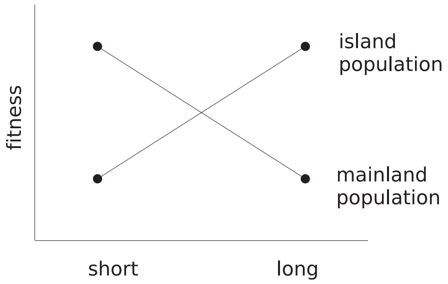


Figure 2: The so-called fitness landscape is different in the island population than in the mainland population. Tall plants are fitter on the island, whereas short plants are fitter on the mainland.

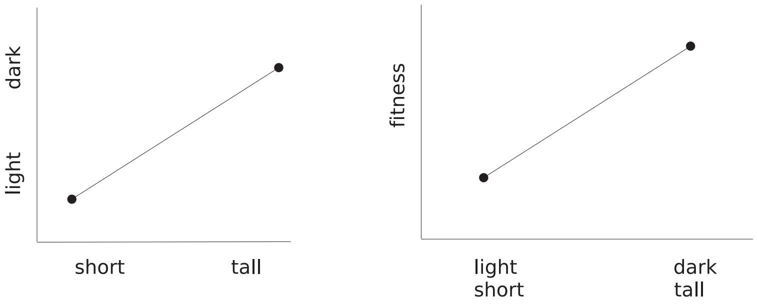


Figure 3: It is a genetic coincidence that taller individuals are fitter than shorter individuals. Tall individuals are dark and dark individuals are fitter. Therefore, automatically, taller individuals are also fitter.

In the first three chapters of this thesis we[1] deal with adaptation. Throughout these chapters we assume that we know the exact direction and magnitude of the selective force. We also assume that the population has only one solution to deal with this force. This one solution is a mutation from one *allele* to another allele. If this beneficial allele reaches a frequency of 100%, we consider evolution completed as far as this locus is concerned. We have made all these assumptions so that we can focus on the stochastic nature of mutation and reproduction. In chapter 4 we look at the splitting of one species into two. We try to understand the role of selective forces in such a speciation process - combining macro- and micro-evolutionary biology.

Allele: an allele is a variant of a gene. A gene is a stretch of DNA that has a certain function. Often more than one variant of a gene exists - each variant may produce a slightly different protein - and these variants are called alleles.

[1] From here on *we* means my collaborators and me.

0.3 What is theoretical evolutionary biology?

A theoretical biologist specializes in the use of models. Models come in many different forms and have various functions.

1. A model can explain observations. The model of natural selection, as Darwin suggested it, can explain why populations are adapted to their environments. Such a model, which gives the answer to a "Why?" question, has certain components, things that we know or believe to be true. After these have been determined, the modeller uses logic or mathematics to determine the outcome of the model. The ingredients of the model of natural selection are 1. heritable variation for a trait and 2. differences in fitness between the variants. The logic of the model is: if some variants are better adapted to the environment than others, they will have higher fitness (i. e. have more surviving offspring in the next generation), and as a result their frequency will increase in the population. If this process is repeated generation after generation, slowly, the population will get better adapted to its environment.

2. A model can also be used to make predictions. Predictions and explanations are closely related. Darwin's theory of natural selection explains *why* populations are adapted to their environment, but it also predicts *that* populations are adapted to their environments. Such predictions do not necessarily deal with the future. They can be independent of time and, for example, state that "after every ice-age (in past or future) animals migrate back to the north". In some cases predictions are there before any observations are made. The model in chapter 1 of this thesis is mostly a predictive model. Our results are therefore formulated in "if - then" constructions such as "*if* the mutation rate is high, *then* a population will adapt from the standing genetic variation" (for explanation see section 0.4).

3. Some processes can not be observed directly. In those cases, we can use models to predict the patterns that will be left behind by a certain process. The pattern can then be used to infer the processes that have happened. A meteor that hits the earth leaves a crater. Even though the process ("meteor hits earth") is not often observed directly, we can infer from the pattern ("crater") that the process has happened. In chapter 3 we determine what pattern would be left in the DNA of a population by a "soft sweep" (for explanation see section 0.6) and we also determine how likely it is that we detect this pattern.

4. A model can be used to estimate a parameter that we can not measure directly. For example, the model of neutral evolution can be used to infer from DNA data how long ago two species speciated. The parameter that we would like to estimate is the divergence time between the two species. The model

of neutral evolution tells us that the number of differences found in the DNA between two species is (roughly) equal to the time that has passed since they split multiplied with the mutation rate per year. Therefore

$$\text{number of differences} = \text{div. time} \times \text{mut. rate}$$

So if we know the number of differences and the mutation rate per year, we can estimate the divergence time.

Models have a long tradition in evolutionary biology, so the models in this thesis have been built on many previous models. There are different reasons why models are much more often used in evolutionary biology than in other biological disciplines. First, evolutionary biology (at least micro-evolution) often deals with populations of individuals. Processes at the individual level (such as mutation and reproduction) are known pretty well and the models are used to predict the effect of these processes at the individual level on processes at the population level. The modeling tradition in evolutionary biology started also because it is more difficult (but not impossible) to carry out experiments in evolutionary biology than it is in the other biological disciplines. Models in evolutionary biology tend to be complicated. Often mathematics, statistics and computer simulations are needed to get to the explanations, predictions and estimates that people are looking for. In the next sections I will try to explain the models that I have worked on, (almost) without using mathematics or much jargon.

0.4 About chapter 1 (Soft sweeps 1)

When a new malaria medication is introduced, *Plasmodium falciparum* (the parasite that causes the most severe form of malaria) is faced with a problem, simply because when a malaria infected person uses the medication, the local parasite population inside this person will die out. *P. falciparum*, however, can become resistant to such a new medication. This sometimes happens quickly, even within the year the medication is introduced, or sometimes slowly or not at all (Talisuna *et al.* 2004). The known resistance mutations in *P. falciparum* are usually very simple. Partial resistance can be caused by a single nucleotide change such as a point mutation at codon 108 in the *dhfr* gene from agc to aac (see Figure 4). This change in the DNA results in an amino acid change in the protein from Serine (S) to Asparagine (N), and this altered protein confers resistance to the malaria drug pyrimethamine (Talisuna *et al.* 2004).

codon number	1	...	105	106	107	**108**	109	110	111	...	163
Ancestral / b allele	caa	...	acc	tgg	gaa	**agc**	att	cca	aaa	...	gga
Protein sequence	Q	...	T	W	E	**S**	I	P	K	...	G
Resistant / B allele	caa	...	acc	tgg	gaa	**aac**	att	cca	aaa	...	gga
Protein sequence	Q	...	T	W	E	**N**	I	P	K	...	G

Figure 4: The DNA sequence of part of the dhfr gene in *Plasmodium falciparum.* The figure shows two alleles, the non-resistant ancestral allele and a resistant mutant allele. Genes are translated into proteins. For this translation three DNA letters correspond to one amino acid. There are 20 different amino acids, and each capital letter stands for an amino acid. The important mutation in the 108th codon changes the 108th amino acid in the protein from Serine (S) to Asparagine (N).

It is of interest how exactly the resistance in the population of parasites evolves – if we knew this we could use this knowledge to try and prevent resistance from evolving. There are very different ideas about how traits such as resistance evolve. Some biologists assume that there is always genetic variation in a population and selection will just change the frequencies of the alleles that are present in the population. This view is typical for quantitative geneticists and breeders. The “breeders equation”, for example, tells us how fast a trait in a population can change depending on the selection pressure and the amount of available genetic variation. Other biologists, e.g., from the field of molecular evolution, tend to think that populations often lack genetic variation for important traits. When selection acts on such a population, new mutations need to happen first before a trait can change. It is unclear whether most adaptation occurs from standing genetic variation as the breeders expect or from new mutations as population geneticists expect. Certainly, populations contain variation for some traits but not for others - so it may be that either scenario is possible. In the first chapter of this thesis we compare these two possibilities.

For now I will focus on just one of the resistance mutations. I will call the old state of the gene the b allele[2] and the new state is called the B allele. I assume that before the new medication was introduced the population con-

[2] Unfortunately, we change our notation between the first paper (chapter 1) and the second (chapter 2). In this introduction I will call the alleles b and B, but in chapter 1 they are called a and A.

sisted mainly or only of individuals carrying the b allele. Substitution of the b allele with the B allele can happen in the two ways described above. The first possibility is adaptation from the standing genetic variation: the population is polymorphic at the b locus, i. e. both small b and big B are present in the population at the time that the new medication is introduced. As soon as the medication is used, the B allele will increase in frequency until it substitutes the b allele in the population. The second possibility is adaptation from new mutations: the population is not polymorphic at the b locus and the population has to wait for mutation to create new B alleles. Once there is a new B allele, it can spread through the population until it reaches fixation.

adaptation from the standing genetic variation

adaptation from new mutations

Probability of fixation from standing genetic variation To be able to compare the probabilities of the two scenarios, we first need a good understanding of each of the scenarios. The standing variation scenario was not often studied before, so we first look at that. We analyse the probability that a B allele from the standing genetic variation ultimately becomes *fixed* in the population. For this, we need to take into account the following things. We first calculate the probability that the B allele was present in the population at the time of the introduction and we calculate the probability that it had a certain frequency in the population at that time. These probabilities depend on the mutation rate, the population size and the disadvantage (if any) of the B allele before the introduction. Given that it was present at a certain frequency, we can calculate the probability that it will actually go to fixation in the population. This fixation probability, in turn, depends on the initial frequency and on the advantage it confers. The total probability that the population adapts from the standing genetic variation is given by equation (8) of chapter 1.

Fixation: an allele is said to go to fixation when it outcompetes alternative alleles and reaches a frequency of 100%.

In chapter 1 we show that the probability of adaptation from the standing genetic variation depends on both the advantage of the B allele over the b allele in the new environment (upon introduction of the new medication) and the disadvantage of B over b in the old environment (without the medication). The important quantity is the ratio between the advantage and the disadvantage. The larger this ratio, the higher the fixation probability in the new environment (see Figure 5 left). This is not hard to explain. If the allele has a large disadvantage, it has low frequency before the environmental change if it is present at all, and only if it has a large advantage after the environmental change it still has a reasonable chance to go to fixation. On the other hand, if the allele has a small disadvantage before, it will have a higher frequency in the population, and even a small advantage after the environmental change suffices to make it go to fixation.

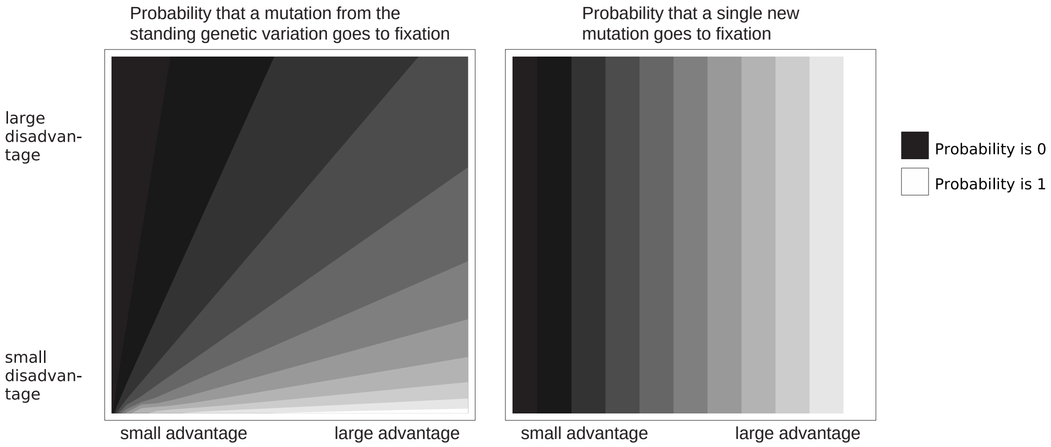


Figure 5: Left: The probability that a B allele from the standing genetic variation becomes fixed, depending on the selective advantage in the new environment and the selective disadvantage in the old environment, for a given value of Θ (0.4). White means the probability is close to 1, black means it is close to 0. Right: same but for a new mutation.

The result is that alleles with a small advantage and a small disadvantage have the same probability to reach fixation as alleles with a big advantage and a big disadvantage. Imagine now that the same number of small and large mutations would occur in a population. And assume that the advantage of an allele would always be strictly proportional to its disadvantage before, so that the advantage-disadvantage ratio is the same for all alleles. In this case, the probability that a mutation with small advantage becomes fixed is the same as the probability that a mutation with large advantage becomes fixed. This means that the population can make a small step towards adaptation or large step with equal probability. If mutations with small advantage are more common than mutations with large advantage (which is generally assumed), then the population would more often take small steps towards adaptation than large steps. This scenario, in which small and large steps are equally likely, contrasts with the situation without standing genetic variation. If a population must wait for new mutations, then mutations with large advantage have a much bigger chance to go to fixation than mutations with a small effect and adaptation will usually procede in large steps (see Figure 5 right and also Figure 1 in chapter 1).

Relative importance of standing genetic variation. Now that we know the probability of adaptation from standing variation, we can go back to the original question: what is the relative importance of standing variation and new mutations? Let's say we observe the population G generations after the introduction of the medication, and we see that the B allele has reached fixation in the population. Now the relative importance of standing variation can be defined as the probability that this B allele originated before the introduction of the new medication. To determine this probability, we need the results from the last paragraph (fixation probability from the standing genetic variation) and the fixation probability for new mutations. For new mutations the calculation is easier (it was already done by HALDANE 1927). To calculate the probability that a new mutation will arise and go to fixation we need the number of mutations per generation and the advantage of the B allele (see elsewhere in this section). The number of mutants that occur in the population per generation is determined by the mutation rate and the population size. The product of mutation rate and population size is usually called Θ[3].

Θ: the product of mutation rate and population size. Θ can be interpreted as the number of mutants in the population per generation, but see later in this section.

Our results for the relative importance of the standing genetic variation are different for mutations with a large advantage and mutations with a small advantage. For mutations with a large advantage, the importance of the standing genetic variation depends mainly on Θ. When Θ is low, the standing variation is not very important, if Θ is high, the standing variation is very important (see Figure 6 left). This is not hard to understand. If the advantage is large, the mutant will certainly go to fixation if it is present in the population. And whether or not it is present mainly depends on the number of mutants per generation, which is determined by Θ. For mutations with a small advantage, the picture looks different. In this case the importance of the standing genetic variation depends mainly on the disadvantage of the mutation in the old environment. If the disadvantage is small, the standing variation is very important, if the disadvantage is large, the standing genetic variation is not important (Figure 6 right). This is because mutations with a small advantage have a low probability to go to fixation, unless they have a high frequency at the time of the environmental change, and this only happens if the disadvantage is small. These results are also shown in a slightly different way in Figure 3 of chapter 1.

Soft sweep from the standing genetic variation In chapter 1, we introduce the term *soft sweep.* When more than one copy of a later beneficial mutation is present in the population before the environmental change, then

[3] In fact, Θ is twice the population size times the mutation rate and the quantity that I talk about in this introduction is $\frac{\Theta}{2}$, however, for readability, I stick to $\Theta = \mu \cdot N$, where μ is the mutation rate and N is the population size.

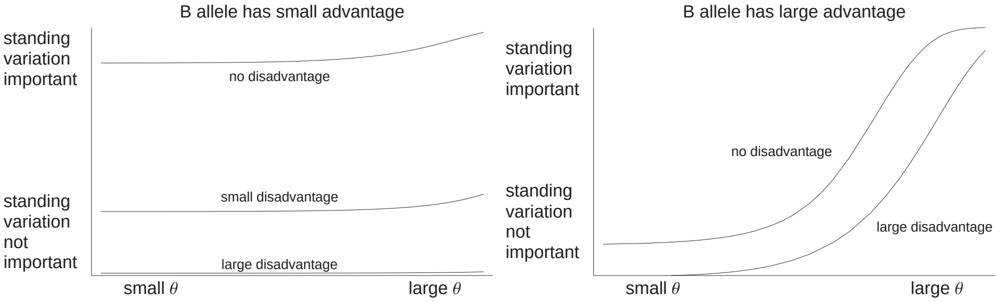


Figure 6: The relative importance of standing genetic variation for adaptation, for alleles with a small advantage (left) and alleles with a large advantage (right). In each plot, there are lines for different levels of disadvantage. On the x-axis is the product of population size and mutation rate Θ.

it is possible that more than one copy contributes to fixation[4]. This is shown in Figure 7. Each line in the figure represents a little fragment of DNA from an individual. In the middle of the fragment is the nucleotide that determines whether an allele is a b or a B allele. The b alleles carry a g at the b locus and the B alleles a t. The t is increasing in frequency and becomes fixed. If more than one copy from the standing genetic variation contributes to fixation of the B allele, we call it a *soft sweep from the standing genetic variation*. This has happened in panel 3: individuals 1-4 are descendents of individual 2 in panel 1, whereas individuals 5 and 6 are descendents of individual 3 in panel 1. If there is only one copy that outcompetes all others, we call it a *hard sweep* and this can be seen in panel 2. Here all individuals are descendents from individual 2 in panel 1. The probability of a soft sweep from the standing genetic variation is shown in Figure 5 of chapter 1. It is important to note that a soft sweep and a hard sweep lead to different patterns in the DNA. More about this in section 0.6. Figure 13 of this introduction shows a second type of soft sweep that is the focus of chapter 2 and 3.

[4] Later in this text I use the word fixation not only for true fixation (when an allele reaches a frequency of 100%), but also for cases where an allele only contributes to fixation. So *the probability to go to fixation* should be usually read as *the probability to contribute to fixation*.

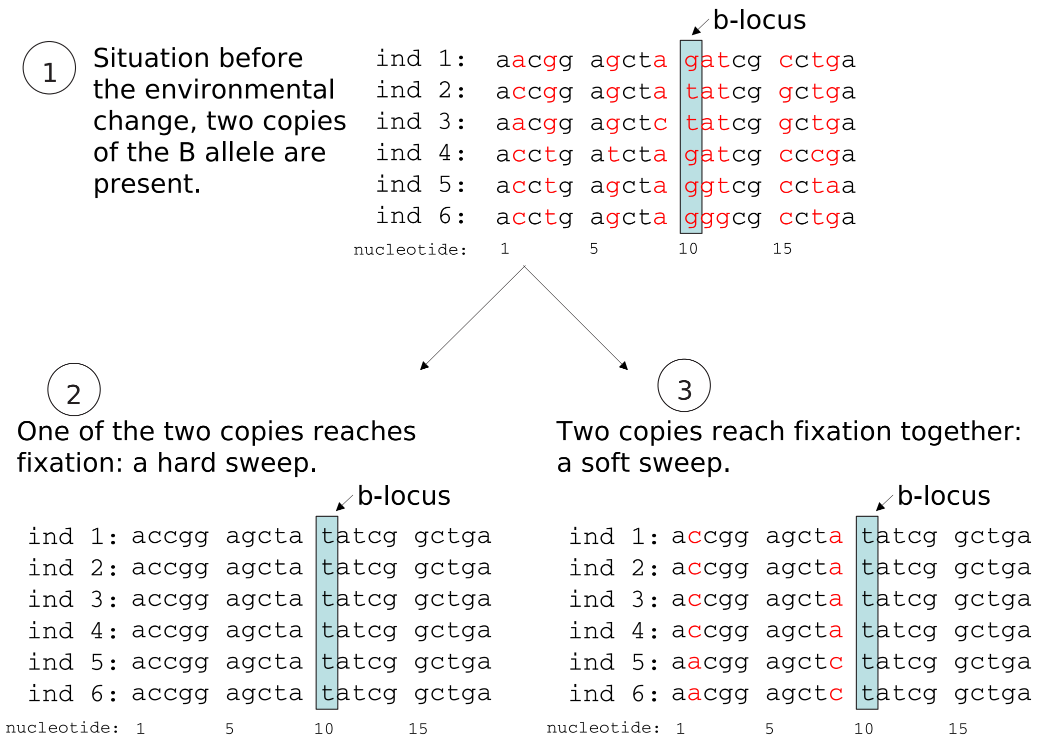
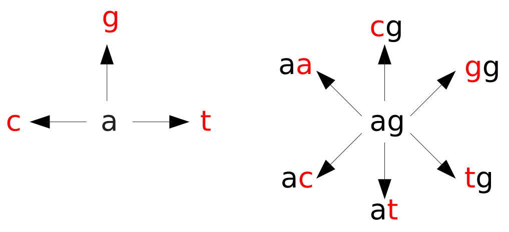

b-locus
1
Situation before the environmental change, two copies of the B allele are present.
ind 1: aacgg agcta gatcg cctga
ind 2: accgg agcta tatcg gctga
ind 3: aacgg agctc tatcg gctga
ind 4: acctg atcta gatcg cccga
ind 5: acctg agcta ggtcg cctaa
ind 6: acctg agcta gggcg cctga
nucleotide: 1 5 10 15
2
One of the two copies reaches fixation: a hard sweep.
b-locus
ind 1: accgg agcta tatcg gctga
ind 2: accgg agcta tatcg gctga
ind 3: accgg agcta tatcg gctga
ind 4: accgg agcta tatcg gctga
ind 5: accgg agcta tatcg gctga
ind 6: accgg agcta tatcg gctga
nucleotide: 1 5 10 15
3
Two copies reach fixation together: a soft sweep.
b-locus
ind 1: accgg agcta tatcg gctga
ind 2: accgg agcta tatcg gctga
ind 3: accgg agcta tatcg gctga
ind 4: accgg agcta tatcg gctga
ind 5: aacgg agctc tatcg gctga
ind 6: aacgg agctc tatcg gctga
nucleotide: 1 5 10 15

Figure 7: The difference between a hard (left) and a soft (right) sweep from the standing genetic variation. Neutral variants can increase in frequency if they are associated with a beneficial mutation. This effect is called genetic hitchhiking. The beneficial allele B is characterized by a t at nucleotide 10. Nucleotides that are polymorphic are in red. For another type of soft sweep see Figure 13. For more explanation see text.



[h]

Figure 8: A single nucleotide can mutate into three other nucleotides, it is said to have three neighbors. A sequence of two nucleotides has six neighbors that can be reached in one mutational step. The mutated nucleotides are in red.

On mutation rates. Note for the reader: if the following gets too technical for you, you can skip it and continue with section 0.5.

The term mutation rate is used in different ways by different biologists.

It is useful to spend a few words explaining what it means throughout the first three chapters of this thesis. The most common use of the term is "the probability[5] that through mutation, an offspring carries a different nucleotide at a given nucleotide position than its parent." If the mother carries a c at a certain position, the "per nucleotide mutation rate" is the probability that the offspring does not carry this c. However, the c can mutate into any of three other nucleotides: a, g, or t. The probability that it mutates to a g, is about one third[6] of the "per nucleotide mutation rate". The c is a sequence of one nucleotide, an it is said to have three direct neighbors, namely, the three other nucleotides. A sequence of two nucleotides has six neighbors that can be reached in one mutational step (see Figure 8) and a sequence of 489 nucleotides (such as the coding region of the *dhfr* gene in *P. falciparum*) has 1467 neighbors[7]. The total mutation probability for a gene, also called the "gene mutation rate" is the "per nucleotide mutation rate" times the number of nucleotides in the gene.

In the case of *dhfr*, I have described one neighbor of the original allele (see Figure 4). This neighbor confers resistance against a malaria drug and differs from the original (wildtype) allele by one nucleotide. The probability to mutate from the original allele to exactly this neighbor is one third of the "per nucleotide mutation rate". However, "sequence space is vast and empty"(van Rheede 2003) and the allele has 488 other neighbors. Most of these neighbors have never been observed in nature (although it is likely that they exist in very low frequency). Many of them (maybe about a quarter of all neighbors) will produce exactly the same protein as the original allele, because the genetic code is redundant. For example, the three nucleotides sequence agt and agc both code for the same amino acid. If a mutation changes agt into agc it will have no effect on the function of the protein at all[8]. Then, there are neighbors that will produce approximately the same protein. Maybe

[5] Technically there is a difference between a rate and a probability, but as long as they are small they can be treated as identical.

[6] In fact, mutation probabilities to the three other nucleotides are not equal. A c is more likely to mutate into a t than into a g or an a, but for the purposes of this section it can be ignored.

[7] For simplicity, I focus only on single nucleotide changes. Mutations can also be, for example, the insertion of one or more nucleotides or the deletion of one or more nucleotides. If one includes such mutations the number of neighbors would be much larger (in fact, infinitely large).

[8] There is a lot of evidence that some codons (combinations of three nucleotides) are in some cases better than others – so a change from agt to agc may have a fitness effect even if it doesn't change the protein sequence.

one hydrophilic amino acid is replaced by another, but the function of the protein is not really affected. Finally, there are neighbors that really change the function of the protein. Most of them will make the protein function worse and those will usually never reach a high frequency in the population and we may never observe them. But some of them will make the protein function better. The resistance allele is an example of such a neighbor. It could be that, say 5 different neighbors have exactly the same improved function e.g., because they would produce exactly the same amino acid, partly because they would produce a different amino acid that would have the same result. We call these neighbors with improved function collectively the B allele.

In the models of chapters 1-3, we simplify the world of alleles drastically and we assume only the original allele exists and the allele or group of alleles that has an improved function. We call the original allele b and the others B. If this group consists of five direct neighbors, then the beneficial mutation rate would be five times the mutation rate for each neighbor. And as we saw earlier, the mutation rate for each neighbor is one third of the “per nucleotide mutation rate”. The probability that one of the neighbors with improved function is reached is called the “beneficial mutation rate”. I will refer to this mutation rate as μ.

The number of mutants and the fixation probability The population of our models in chapters 1 to 3, has discrete generations and consists of N haploid individuals, so that every year all individuals die and N new individuals are born. This may sound strange, but it is the case, for example, in annual plants. It is also possible to show that the outcomes of most models do not depend much on this assumption, but it makes the math much easier. If the beneficial mutation rate is μ, then in a population that consists only of b alleles, the expected number of B mutants in the next generation is $N \cdot \mu$.

Haploid: the individuals have only one set of chromosomes, instead of two.

In an ideal population, offspring are randomly distributed among the potential parents. Such a population is called a Wright-Fisher population, after Sewall Wright and Ronald A. Fisher, two of the founding fathers of the field of population genetics. This random distribution means that the N offspring of the next generation each belong to a random parent, independent of whether or not this parent already has offspring. The result is (approximately) a Poisson distribution of offspring numbers. Some potential parents have no offspring, some have one, some two etc. (see Figure 9).

However, offspring do not have to be Poisson distributed and often, offspring are much less evenly distributed than if they were Poisson distributed. In those cases, many individuals would have no offspring and only few indi-

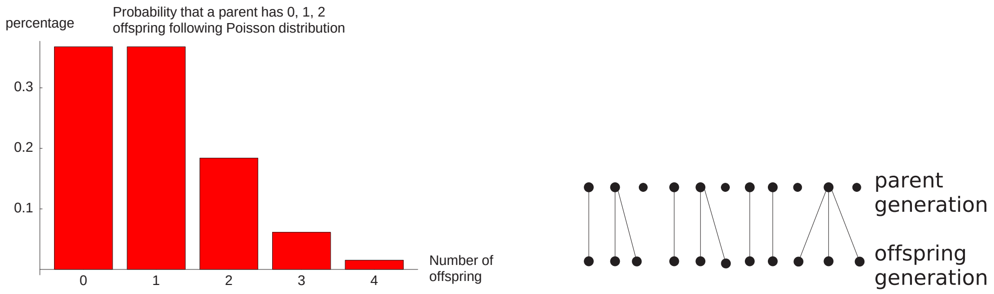


Figure 9: Left: the probability that an individual has 0, 1, 2 etc. offspring. Right: Another way to represent the distribution of offspring among parents. The black dots represent individuals, an individual in the parent population is connected to its offspring by a line.

viduals have many offspring. A measure of how even a distribution is, is the variance[9] in offspring number. In a population where every individual has exactly 1 offspring, the variance is 0; when offspring is Poisson distributed, the variance is 1; and when most individuals have 0 offspring, but some have 4, the variance is 3 (given that he mean number of offspring is 1 in all these cases).

We are interested in the probability that a beneficial allele B, with advantage s goes to fixation. B is said to have advantage s if a B individual has on average $(1 + s)$ times the number of offspring of a b individual. Another population geneticist, J. B. S. Haldane, showed that the fixation probability (P_{fix}) of a beneficial allele is twice its advantage (s) (Haldane 1927).

s is the selective advantage of allele B.

$$P_{\text{fix}} = 2s$$

This result is well known, but it is often forgotten that it only holds when the offspring variance is 1, as is the case when offspring numbers are Poisson distributed. If the variance is larger, the fixation probability is smaller. Roughly, the fixation probability is

$$P_{\text{fix}} = \frac{2s}{\sigma^2}.$$

where σ^2 is the offspring variance (see Figure 10). The fixation probability of a beneficial allele is easily calculated for some simple offspring distributions. Imagine a population where every individual can only have zero or two offspring. If the population size is to stay constant, the mean number of offspring must be 1 and therefore half of the individuals must have zero offspring and the other half must have two offspring. Assume now that there is a mutant individual in the population, and this one mutant has higher fitness than the

[9] Variance: the average of the square of the distance of each data point from the mean.

others. The fixation probability is the probability that eventually, all individuals in the population are descendents of this individual. Let's assume that the mutant has a 10% advantage over the others and the expected number of offspring of this individual is 1.1. This would be the case if the mutant would have 55% chance to have two offspring and only 45% chance to have no offspring. The mutant allele will become fixed in the population if it leaves offspring in every next generation. If it is lost in any generation, it will not fix. Let's call the probability that the allele becomes fixed P_{fix} and the probability that it is eventually lost L. We then have

$$P_{\text{fix}} = 1 - \mathrm{L}$$

The probability that the beneficial allele is lost in the very next generation is simply the probability that the mutant individual has no offspring, and we stated before that this probability is 0.45. With probability 0.55, it will have two offspring. If this happens, there will be two mutant individuals in the next generation. They each carry a copy of the beneficial allele. The probability that each of those copies is lost is the probability that the first copy is lost, times the probability that the second one is lost. But in a large population, each of these copies has the same probability of being lost as the original allele, we called this probability L. The probability that two copies are eventually lost is L^2. And this gives us:

$$\mathrm{L} = 0.45 + 0.55 \cdot \mathrm{L}^2$$

This equation can be solved and we find that the probability that this allele is lost is about 0.82, so the probability that it is not lost, and therefore becomes fixed is 0.18. In the next table (Figure 11), I have done the same calculation for two other examples with larger offspring number variances. You can see that the fixation probability goes down with increasing offspring variance. In Figure 10 I have plotted the prediction from Haldane ($P_{\text{fix}} = \frac{2s}{\sigma^2}$) and the three points that I have just calculated. You can see that it fits pretty well.

In the last paragraphs I have explained that the number of B mutants in a generation is $N\mu$ and that the fixation probability of each of these mutants is $\frac{2s}{\sigma^2}$. The product of these two numbers is the probability that in a generation a mutant arises that will go to fixation. In chapter 1, this probability is called p_{new}. So we have

$$p_{\text{new}} = N\mu\frac{2s}{\sigma^2}.$$

Most population geneticists are used to the notion of effective population size,

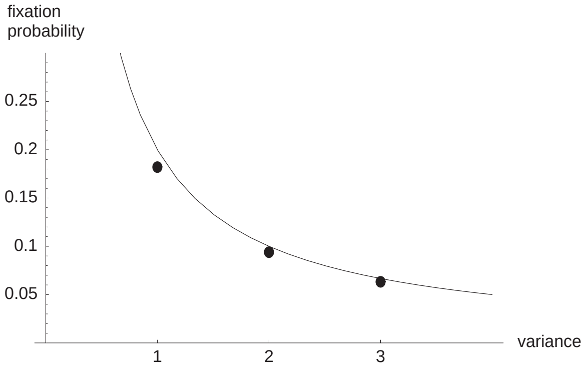


Figure 10: How the fixation probability depends on the offspring variance. The points are from the examples in Figure 11

$N_e = \frac{N}{\sigma^2}$. So out of habit, we would write[10]

$$p_{\text{new}} = N\mu \cdot \frac{2s}{\sigma^2} = \mu \frac{N}{\sigma^2} \cdot 2s = \mu N_e 2s$$

Intuitively, one may expect that the real number of mutants, μN is important for adaptation and not μ times the effective population size. However, if we use the real number of mutants, then we need to take also the real fixation probability. Somewhere, the σ^2 has to enter, and it doesn't matter where.

It may be surprising to learn that a mutant that has a 10% advantage over the others in the population has only a probability of about 20% to go to fixation. A 10% advantage is considered very unrealistic; beneficial alleles are expected to have advantages mostly below 1%, so their fixation probability would be less than 2%. This may be understood in the following way. Imagine a mutation that makes a butterfly better camouflaged. This is definitely a beneficial mutation. It will make it less likely that the butterfly is eaten by a bird. But certainly not impossible! Also, the butterfly or its offspring can die of many causes other than predation, they can die of hunger, or not find a mate, or the eggs can be eaten by a bird. Even with a mutation that gives a clear advantage, an individual has no guarantee to have offspring. If it is very common that individuals have no offspring at all (which is the case with high offspring variance), then it is very likely that beneficial mutations do not fix.

[10] This formula looks a bit different than the one in chapter 1 because in chapter 1 we use a diploid population so we need $2N_e$ and $2hs$, where h is the so-called dominance factor.

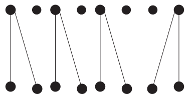
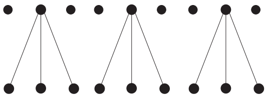
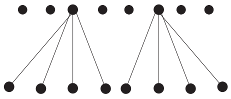
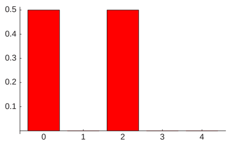
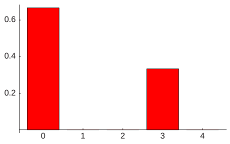
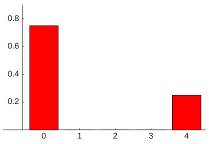
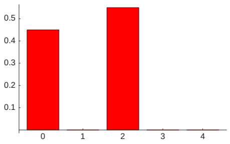
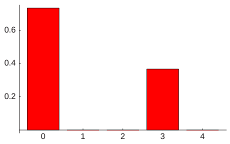
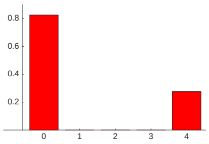
	0 or 2 offspring	0 or 3 offspring	0 or 4 offspring
			
Orig. offspr. distribution			
Mut. offspr. distribution			
Offspring per mutant	$2 \cdot 0.55 = 1.1$	$3 \cdot 0.37 = 1.1$	$4 \cdot 0.27 = 1.1$
Offspring variance σ^2	1	2	3
Prob. of loss $\mathrm{L} =$	$0.45 + 0.55 \cdot \mathrm{L}^2$	$0.63 + 0.37 \cdot \mathrm{L}^3$	$0.73 + 0.27 \cdot \mathrm{L}^4$
Fixation prob. $P_{\text{fix}} =$	0.18	0.10	0.06

Figure 11: Fixation probabilities in populations with three different offspring variances.

0.5 About chapter 2 (Soft sweeps 2)

In this section, I take again the resistance mutation as an example. I assume that at the moment the new malaria medication is introduced, the parasite population consists of only non-resistant individuals that carry the b allele, for example, because the B allele was strongly deleterious before. Then, at some point in time, one individual is born that is resistant because its b allele mutated into a B allele. This mutation can now increase in frequency and eventually become fixed in the population - but this process will take some time. Let's say it takes T_{fix} generations. During these T_{fix} generations, there are still many b alleles in the population and it is possible that one of these b alleles also mutates into a B allele. It is then possible that the second B allele

T_{fix}: the time to fixation of an allele depends on its selective advantage s and the population size N_e. $T_{\text{fix}} \approx \frac{2N_e s}{s}$ (from the appendix of chapter 1).

increases in frequency as well. In the end, half of the population may carry the first B allele and the other half the second B allele. The two B alleles have independent origins because they originate from different mutation events. We call a substitution by two (or more) independent alleles a *soft sweep from recurrent mutation* (see Figure 13 and section 0.6).

Probability of a soft sweep from recurrent mutation. In the second chapter of this thesis, we derive the probability that there is more than one B allele involved in a substitution. Until now, nobody explicitely calculated this probability, because it was assumed to be so small that it could be ignored. However, we find that it is quite likely that more than one B allele substitutes the b allele and we find that the probability that this happens depends only on Θ (the mutation rate times the population size) and not on the selection coefficient (advantage) of the B allele. This result is probably the most surprising result of this thesis (equation 11 in chapter 2). It can be understood intuitively. The time it takes for allele B to go to fixation depends on the advantage of B over b. If this advantage is large, fixation will be fast, if it is small, fixation will take longer. Therefore, the number of new mutants that occurs in the population before fixation is reached goes down when the advantage of the B allele goes up. However, as explained in section 0.4, the probability that such a new mutant goes to fixation goes up with the advantage of B (see Figure 12). Therefore, when B has a small advantage, there will be many mutants during T_{fix}, but each of them will have low probability of going to fixation. When B has a large advantage, T_{fix} is short and there will be few mutants, but they will have a high probability of reaching fixation. The two effects of the advantage of B cancel out and the probability that a mutant arises and goes to fixation is almost independent of the advantage. On the other hand, the probability of a soft sweep from recurrent mutation depends strongly on Θ, because the higher Θ is, the more mutants arise and this will make it more likely that more than one B allele contributes to fixation.

You may ask why it is important whether there is one or more independently derived B allele in the population. If there is only one B allele, then, after substitution there will be no variation in the population in the region around the b locus (as in panel 3 of Figure 13, for a description of the figure see the next section). The genetic background that the B allele occurred on will have spread through the population together with the B allele. This effect is called hitchhiking. The region without variation is called a sweep region. Importantly, such a sweep pattern can be searched for in the genome. It allows us to find genes that have recently undergone a rapid fixation of a new allele.

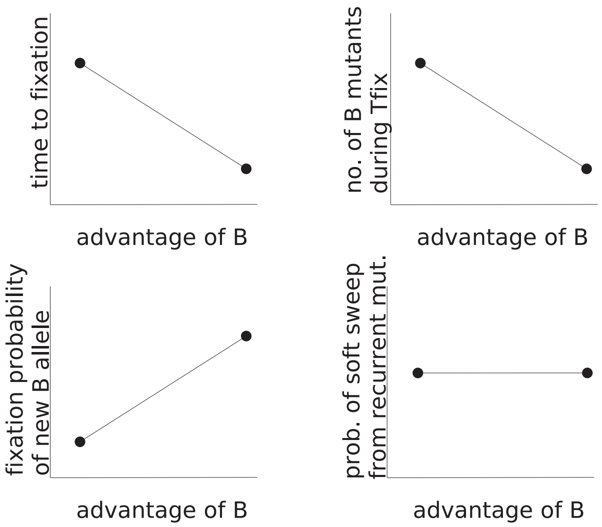
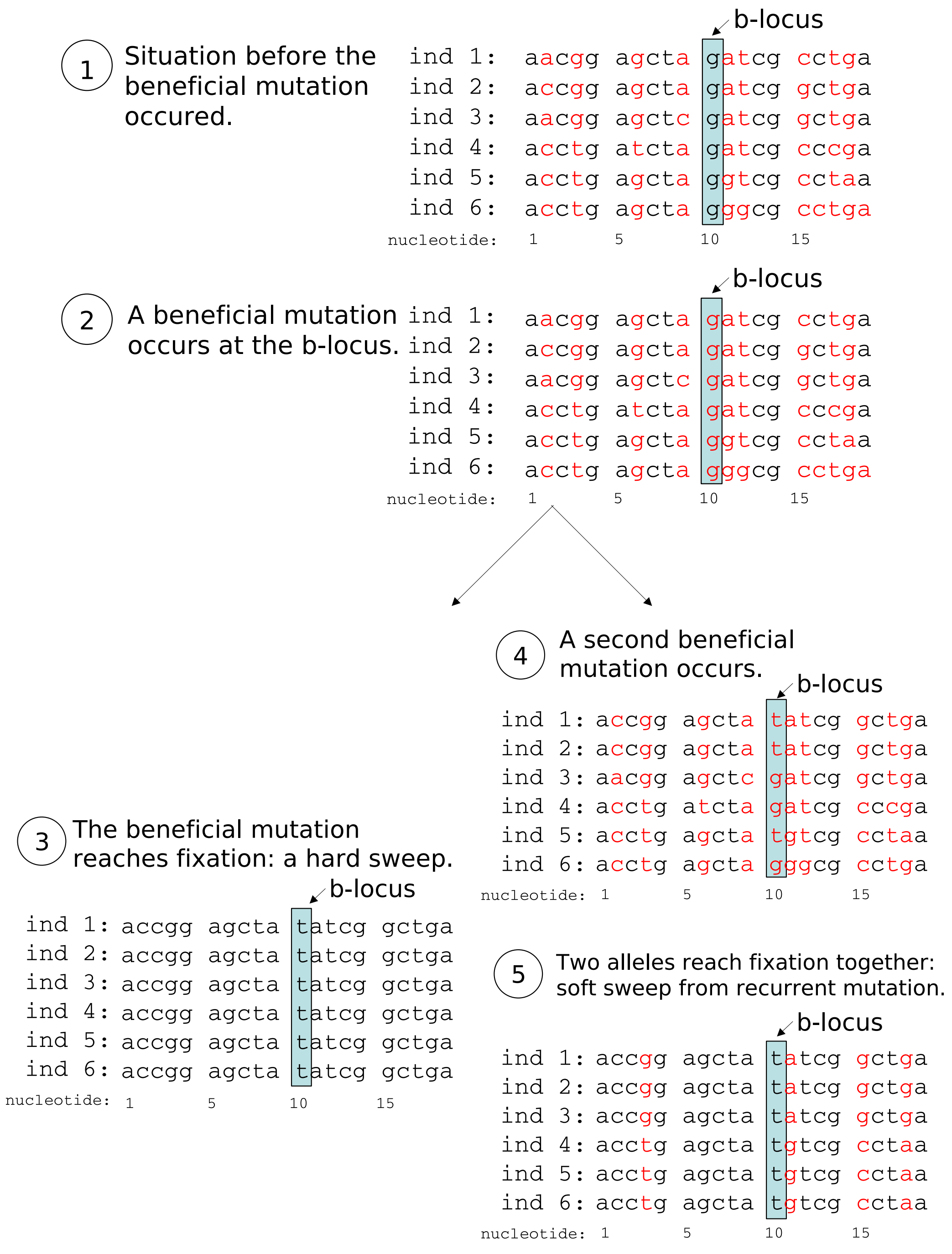


Figure 12: The time it takes for a B allele to go to fixation, T_{fix}, decreases as the advantage of B increases. The number of mutants that arises during T_{fix} therefore also decreases with increasing advantage of B. The probability that such a mutant contributes to fixation increases with the advantage of B. The probability that more than one B allele contributes to fixation doesn't depend on the advantage of B.

These genes, in turn, are of interest to us because it is these genes that have contributed to the recent adaptation of the population that we study. However, if there are two independent B alleles, there are also two independent backgrounds. If these backgrounds are not the same, then there will still be variation in the region (as in panel 5 of Figure 13). In order to find such genes where a soft sweep from recurrent mutation has happened, we need to search for a different pattern. This pattern and how to search for it is the topic of chapter 3.

Figure 13: (right page) The difference between a hard sweep (left) and a soft sweep from recurrent mutation (right). Neutral variants can increase in frequency if they are associated with a beneficial mutation, this is called genetic hitchhiking. In the case of a hard sweep the result is that there is no variation left after substitution, in the case of a soft sweep, variation can remain, in this case there are two backgrounds (haplotypes) left. The difference between the soft sweep in this figure and the one in Figure 7 is that here, there are two completely independent beneficial mutations. The backgrounds on which these mutations happen can therefore be very different. In the case of a soft sweep from the standing variation (Figure 7) it that it is possible that the two copies of the beneficial allele are identical by descent, which means that they originate from one mutational event. The result is that their backgrounds are only different because new (neutral) mutations have happened on the background. The backgrounds will be much more similar in this case. Nucleotides that are polymorphic are in red.



1 Situation before the beneficial mutation occured.

```
           b-locus
ind 1:  aacgg agcta gatcg cctga
ind 2:  accgg agcta gatcg gctga
ind 3:  aacgg agctc gatcg gctga
ind 4:  acctg atcta gatcg cccga
ind 5:  acctg agcta ggtcg cctaa
ind 6:  acctg agcta gggcg cctga
nucleotide: 1   5    10   15
```

2 A beneficial mutation occurs at the b-locus.

```
           b-locus
ind 1:  aacgg agcta gatcg cctga
ind 2:  accgg agcta gatcg gctga
ind 3:  aacgg agctc gatcg gctga
ind 4:  acctg atcta gatcg cccga
ind 5:  acctg agcta ggtcg cctaa
ind 6:  acctg agcta gggcg cctga
nucleotide: 1   5    10   15
```

3 The beneficial mutation reaches fixation: a hard sweep.

```
           b-locus
ind 1: accgg agcta tatcg gctga
ind 2: accgg agcta tatcg gctga
ind 3: accgg agcta tatcg gctga
ind 4: accgg agcta tatcg gctga
ind 5: accgg agcta tatcg gctga
ind 6: accgg agcta tatcg gctga
nucleotide: 1   5    10   15
```

4 A second beneficial mutation occurs.

```
           b-locus
ind 1: accgg agcta tatcg gctga
ind 2: accgg agcta tatcg gctga
ind 3: aacgg agctc gatcg gctga
ind 4: acctg atcta gatcg cccga
ind 5: acctg agcta tgtcg cctaa
ind 6: acctg agcta gggcg cctga
nucleotide: 1   5    10   15
```

5 Two alleles reach fixation together: soft sweep from recurrent mutation.

```
           b-locus
ind 1: accgg agcta tatcg gctga
ind 2: accgg agcta tatcg gctga
ind 3: accgg agcta tatcg gctga
ind 4: acctg agcta tgtcg cctaa
ind 5: acctg agcta tgtcg cctaa
ind 6: acctg agcta tgtcg cctaa
nucleotide: 1   5    10   15
```

0.6 About chapter 3 (Soft sweeps 3)

Evolution (in the sense of the changing of a species) is usually a slow process and therefore not easy to study. One way population geneticists study evolution, is by using patterns in DNA polymorphism that are left by certain processes. Fast substitution of an allele is such a process. If selection favors allele *B* over allele *b*, then, given that *B* becomes fixed, fixation will be fast. It was John Maynard-Smith and John Haigh who realized that fast substitution leaves a distinct pattern in the DNA (Maynard Smith and Haigh 1974). In order to find such pattern in data, we need to describe the pattern accurately and quantify aspects of it so that we can search for it.

Figure 13 shows what happens when a *B* allele goes to fixation. Each line in the figure represents a short fragment of DNA from an individual. In the middle of the fragment is the nucleotide that determines whether an allele is a *b* or a *B* allele. The *b* alleles carry a $\underline{\text{g}}$ at the *b* locus and the *B* alleles a $\underline{\text{t}}$. The $\underline{\text{t}}$ is increasing in frequency and becomes fixed. The classical case is shown by the left part of Figure 13. No variation at the *b* locus is available at the time of the environmental change (panel 1). There is variation at some of the other nucleotides (at 2, 4, 6, 9, 11, 12, 15, 17 and 18). After the environmental change, a single *B* allele occurs and goes to fixation (panel 3). After fixation there is no variation left around the *b* locus. What you don't see in this figure is that further away from the *b* locus, there will be variation again. This is because an individual that carries the *B* allele can exchange parts of the chromosome with another individual by recombination (crossing-over). The pattern, that is caused by fast fixation of a single *B* allele is well described and often used to find loci of interest.

In chapter 2 of this thesis we show that it is possible that more than one *B* allele contributes to fixation. The right part of Figure 13 shows what happens if the *B* allele occurs more than once. The starting point is the same, a *B* allele occurs in the population. However, in this case the mutation from $\underline{\text{g}}$ to $\underline{\text{t}}$ occurs a second time (panel 4). Because it occurs in two different individuals, it is linked to two different backgrounds (genetic backgrounds in this sense are often called haplotypes). As you can see from panel 5, there is still variation left after the *B* alleles have reached fixation. There are polymorphisms left at nucleotides 4, 11, 15 and 18. This is very different from what you see in panel 3 where there is no variation left at all. What is also important to notice, is that the first three individuals all carry $\underline{\text{g}}$, $\underline{\text{a}}$, $\underline{\text{g}}$, $\underline{\text{g}}$ at nucleotides 4, 11, 15 and 18, whereas the last three individuals all carry $\underline{\text{t}}$, $\underline{\text{g}}$, $\underline{\text{c}}$, $\underline{\text{a}}$ at those nucleotides. This is because the combination of $\underline{\text{g}}$, $\underline{\text{a}}$, $\underline{\text{g}}$, $\underline{\text{g}}$ was associated with the first *B* allele and the combination t, g, c, a was associated with the second B allele. The result is that individuals 1, 2 and 3 are identical in this DNA fragment, and individuals 4, 5, and 6 as well, but between the two groups there are four differences. The polymorphisms are said to be in linkage disequilibrium and this is one of the aspects of this pattern that we can use to detect it.

The K test. If one would find such a pattern in data from a real population, the first question to ask would be: Does this pattern deviate from what we expect under normal circumstances? Normal circumstances would be, for example, the absence of selection. To decide whether a pattern deviates from what is normal, we need two things, first we need a way to quantify the pattern, and second we need to know what can be considered normal values of this quantity.

One way to quantify the pattern that I described in the last paragraph is to count the number of polymorphisms and the number of different sequences (haplotypes). A polymorphism that shows the same distribution of states as another polymorphism does not create any new sequences, it only makes sequences that are already different more different. In panel 5 of Figure 13, individuals 1, 2 and 3 carry the same haplotype, they have exactly the same sequence. Individuals 4, 5 and 6 carry a second haplotype. The number of haplotypes is often indicated by K. For panel 5, we have $K = 2$ and $S = 4$ (S is the number of polymorphic sites). One can also count the number of haplotypes in panel 1 (which shows the equilibrium population before selection started). In panel 1, I will only consider the first four polymorphisms (which corresponds to the first 10 nucleotides), so that there is the same number of polymorphisms as in the last panel. Individuals 5 and 6 have the same sequence, but all other individuals have different sequences. The equilibrium population has five different haplotypes ($K = 5$) after four polymorphisms. The population after the soft sweep had only two haplotypes with the same number of polymorphisms. Two is less than five, but the question now is is it significantly less?

K is the number of haplotypes or the number of different sequences in a sample.

Note that 4 polymorphisms lead to at least 2 haplotypes and at most 5 haplotypes in the sample.

By doing extensive simulations of equilibrium populations (without selection) we can determine which K values can be considered normal and which values too low compared to the number of polymorphisms. For example, for four polymorphisms, and if there are 20 individuals, then the expected number of different haplotypes (K) is 4.15. About 3% of the simulated samples has only two haplotypes. 97% of the simulated samples has more than two haplotypes. We can therefore say that a K value of two is significantly low ($p < 0.03$), and a K value larger than two is normal. The distribution of K values is shown in Figure 14.

Power analysis. We have seen that we can test whether the number of haplotypes, K, is significantly low given the number of polymorphic sites in the sample. The test is called the K test. The next thing we can do is to try to determine the power of this K test. For this we do again many simulations, but this time not of equilibrium populations, but of populations in which a B allele substitutes a b allele. We take only those populations where we know that at least two independent B alleles have contributed to fixation. For these populations we now look at the number of polymorphisms and the number of

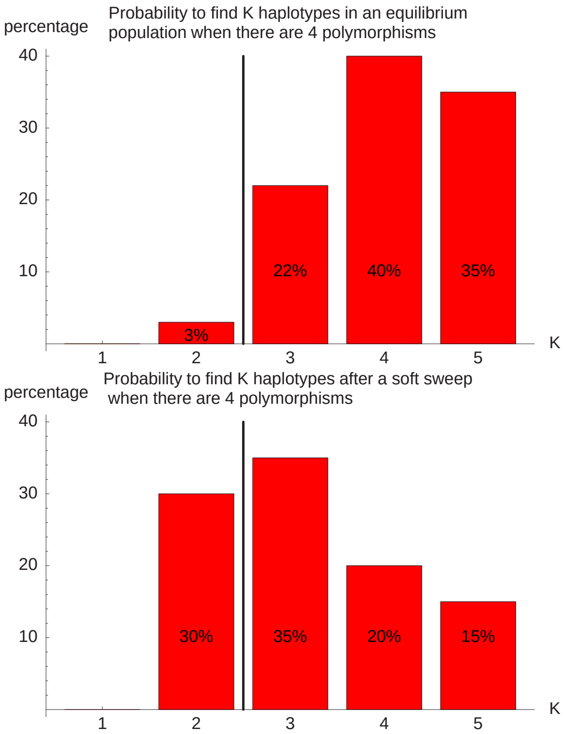


Figure 14: Upper figure: the distribution of K values in simulations of equilibrium populations each time for four polymorphisms. This distribution determines the boundary of the 5% significance (black line), values left from the 5% boundary are significantly low, values right of the 5% boundary are not. The lower figure shows results from simulation where at least two B alleles have reached fixation together, and where there are four polymorphisms. 30% of these simulations showed only two haplotypes, so in 30% of the cases we can reject the null hypothesis that no selection has happened. We know that in all the simulations selection has happened, but the test cannot detect this in all cases. In this example the power of the K test to detect a soft sweep is 30%.

haplotypes in a stretch of DNA. And we check whether the K value that we find in a simulation run is significantly low. For all simulations we then count the number of significant test results and we get the percentage of simulation runs that gave a significant result (see Figure 14). This percentage is what we call the power of the test. If it is high it means that the substitution by two B alleles is often recognized because it has too few haplotypes. If the power is low, it means that the test can only detect some cases. Figure 7a in chapter 3 shows the results for the power analysis that I have just described.

0.7 About chapter 4 (Sympatric speciation)

One of the main aims of evolutionary biology is to explain the species diversity that we see today and in the fossil record. Speciation apparently takes place often enough to give rise to a high species diversity, but not so often that we cannot distinguish species anymore. Understanding the process of speciation is therefore a central theme in evolutionary biology.

During *allopatric speciation*, a population splits into two geographically isolated populations, for example, when the habitat is split in two. The two isolated populations then evolve independently and when they come back into contact, they may have evolved such that they are reproductively isolated and they are no longer capable of mating and producing viable offspring. In *sympatric speciation*, species diverge while inhabiting the same habitat. For this to work, the species must split up in two groups that do not mate with each other. Only then can the two groups be considered *biological species.* At the same time the two groups must diverge ecologically, for example by using different food resources, otherwise one group would outcompete the other group. This is because of the law of competitive exclusion: two species cannot coexist if they use exactly the same resources. Allopatric speciation is considered much easier than sympatric speciation, because the geographic isolation gives time for ecological differentiation and reproductive isolation to evolve. However, it is not clear how much time is needed for these things to evolve, especially since there is no selective force that promotes this evolution. Sympatric speciation is considered less likely, yet also possible. There are some convincing examples of sympatric speciation. It is quite clear, for example, that at least some of the cichlids in Lake Victoria in East Africa have speciated in the lake without any geographical barriers (see Figure 16).

allopatric speciation

sympatric speciation

The biological species concept: species are groups of interbreeding natural populations that are reproductively isolated from other such groups (Mayr 1942).

Many theoretical biologists have worked on sympatric speciation. They have built models to understand under what conditions sympatric speciation

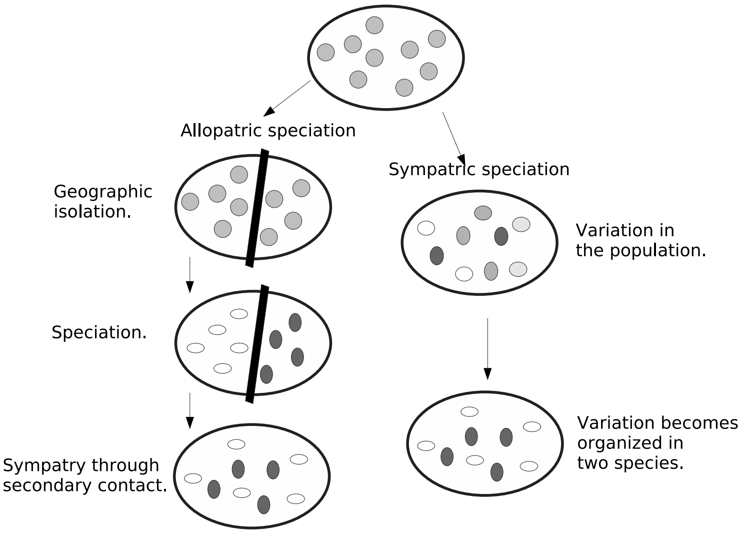

Allopatric speciation
Sympatric speciation
Geographic isolation.
Speciation.
Sympatry through secondary contact.
Variation in the population.
Variation becomes organized in two species.

Figure 15: Allopatric speciation (left) involves a period in which the two populations are geographically isolated, whereas during sympatric speciation (right) there is no geographic isolation. Another mechanism, such as assortative mate choice, is needed to induce reproductive isolation.

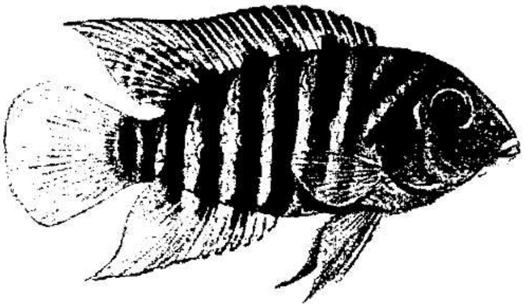


Figure 16: A cichlid from Lake Victoria in Africa. At least some of the enormous species diversity of cichlids in Lake Victoria is thought to be the result of sympatric speciation.

is possible. In 1999 Ulf Dieckmann and Michael Doebeli published a paper in *Nature* that has raised a lot of controversy (Dieckmann and Doebeli 1999). In this paper they show that, in their model, sympatric speciation occurs easily. Since their paper was published, at least ten papers have been written stating that Dieckmann and Doebeli did something wrong in their model and that in more realistic models sympatric speciation does not occur so easily. Chapter 4 of his thesis consists of a new analysis of the Dieckmann and Doebeli model. We simplified the model so that the important features are still there, but a more thorough analysis is possible. With our results we can resolve some of the controversy by showing exactly why some models give different results than others. Critics of the Dieckmann and Doebeli paper wrote that the reported results are only possible because of the high mutation rate, the small phenotypic range and the availability of variation in their model. We show that what is important for their result is indeed the phenotypic range, but also how much the individuals in the population compete for resources. Much less important, and not crucial to their result, is the availability of variation and the mutation rate.

The model we analyzed. The model describes a population of individuals, say fish, and each fish is characterized by two traits. The first trait can be a preferred food particle size, the second trait is the level of choosiness for partner choice. Food comes in different sizes, and most food particles are of intermediate size. Individuals have different food preferences, but initially most individuals prefer intermediate size food particles. Females can choose their mate, but initially they are not choosy, they will simply mate with the first male they find. If there are many fish eating the intermediate size food, then it is possible that there is more food left that is large or small rather than medium sized.

Food size preference is genetically determined by a single gene with alleles a and A. Individuals that carry aa[11] prefer small food particles, Aa prefer medium size food particles and AA prefer large food. Food preference can be genetically determined if, for example, it depends on the size of the individual itself. Large individuals have large mouths and are better at eating large food particles. If a large female (that prefers large food) mates with a large male that also prefers large food particles, then the offspring will also prefer large food particles. If there is a shortage of intermediate sized food, then it is good for a female to mate with a male that is similar to her. If she is large, but she mated with a male that is small, then the offspring would be medium size. This offspring would prefer intermediate food particles and it would not have enough to eat.

In our model we allow choosiness to evolve. Mutations can happen so that females are slightly more or slightly less choosy than their parents. If mutations that make the fish more choosy spread through the population then in the end fish will only mate with their own type. Whether or not this happens depends on the exact parameter values of the model, for example on how much the individuals compete for food. If in the end, the fish only mate with their

[11] In chapter 4 individuals are diploid so they carry two copies of each gene.

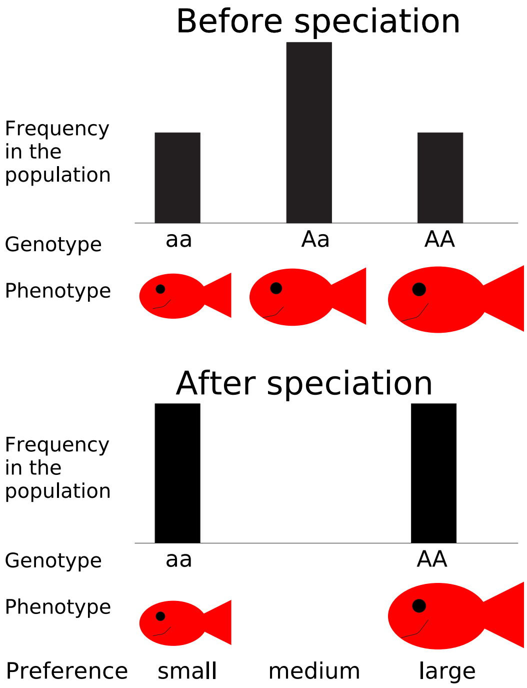


Figure 17: Speciation in the model of chapter 4. Before speciation mating is random and the population is in Hardy-Weinberg equilibrium. In this equilibrium, and if the two alleles (a and A) have equal frequency, then the frequency of heterozygotes (Aa) is 0.5, and of each of the homozygotes 0.25. After speciation there are only homozygotes left. They breed only among themselves so no heterozygotes are born.

own type, then the population has speciated (see Figure 17). One species will consist of only large individuals that eat large food particles, and the females of this species will be very choosy and mate only with large males that also prefer large food. In this population there will only be A alleles. The other species will consist of small individuals that eat small food and females that want to mate only with small males that prefer small food. This population will consist of only individuals with a alleles.

One can imagine the speciation problem also from the opposite direction. What if there are two species, one with genotype aa and one with genotype AA. If this is the case, wouldn't there be a lot of medium sized food that is not eaten by anyone? In other words, is there a niche in the middle of the food size spectrum? The answer is given in detail in chapter 4, but I will give a short version here. First of all, the individuals that prefer a certain food size do not only eat food of exactly that size. They will eat mostly that food size, but they can also eat food that has a slightly different size (they

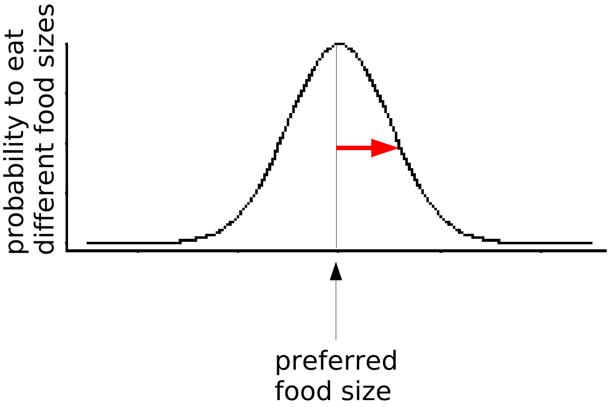
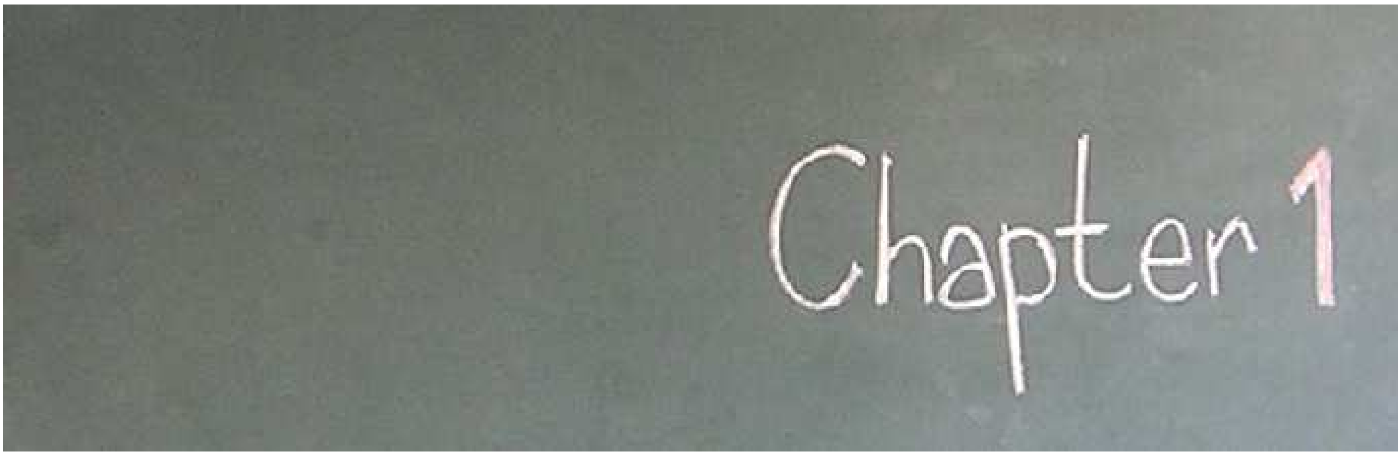


Figure 18: The distribution of food eaten by an individual, if there would be food available of every size. The preferred food size is most eaten, but the other food sizes are also eaten. The red arrow shows the width of the distribution, or how picky the individual is. In chapter 4 this parameter is σ_c.

eat following a Gaussian distribution, see Figure 18). This means that if the small individuals have a preference that is not so different from the medium individuals, then the small individuals will eat also part of the preferred food of the medium individuals. And the large individuals would do the same. The result is that there is no food left in the middle and hence there is no niche in the middle. However, there is two situations in which this is not the case. In those cases speciation will not (or not always) happen. The first situation is when the fish are so picky in what they eat, that both the small and large fish eat almost no medium sized food (in chapter 4 this is the case when σ_c is small). The second situation is that the fish are not very picky, but their preferred food sizes are so wide apart that again the small and large fish eat almost no medium sized food (in the chapter this is the case when x is large). How far apart the preferences of the different types of fish are is determined by the parameter x in our model. The explicit use of this parameter x is one of the reasons why we could get a clearer picture of the behavior of the model than some papers before us.



Chapter 1

Soft Sweeps – Molecular Population Genetics of Adaptation from Standing Genetic Variation

Joachim Hermisson and Pleuni S. Pennings

Published in Genetics (2005) 169: 2335-2352

There are two ways in which a population can adapt to a rapid environmental change or habitat expansion. It may either adapt through new beneficial mutations that subsequently sweep through the population or by using alleles from the standing genetic variation. We use diffusion theory to calculate the probabilities for selective adaptations and find a large increase in the fixation probability for weak substitutions, if alleles originate from the standing genetic variation. We then determine the parameter regions where each scenario – standing variation vs. new mutations – is more likely. Adaptations from the standing genetic variation are favored if either the selective advantage is weak or the selection coefficient and the mutation rate are both high. Finally, we analyze the probability of "soft sweeps", where multiple copies of the selected allele contribute to a substitution and discuss the consequences for the footprint of selection on linked neutral variation. We find that soft sweeps with weaker selective footprints are likely under both scenarios if the mutation rate and/or the selection coefficient is high.

1.1 Introduction

There are two contrasting ways in which evolutionary biologists envisage the adaptive process following a rapid environmental change or the colonization of a new niche. On the one hand, it is well known from breeding experiments and artificial selection that most quantitative traits respond quickly and strongly to artificial selection (see e.g. Falconer and Mackay 1996). In these experiments, there is almost no time for new mutations to occur. Evolutionists who work with phenotypes therefore tend to hold the view that also in natural processes a large part of the adaptive material is not new, but already contained in the population. In other words, it is taken from the standing genetic variation. Consequently, standard predictors of evolvability, such as the heritability, the coefficient of additive variation, or the G matrix are derived from the additive genetic variance of a trait, cf. e.g. Lande and Arnold (1983); Houle (1992); Hansen *et al.* (2003), and Lynch and Walsh (1998); Steppan *et al.* (2002) for review. On the other hand, in the molecular literature on the adaptive process and on selective sweeps adaptation from a single new mutation is clearly the ruling paradigm (e.g. Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Barton 1998; Kim and Stephan 2002). In conspicuous neglect of the quantitative genetic view, the standing genetic variation as a source for adaptive substitutions is generally ignored, with only few recent exceptions (Orr and Betancourt 2001; Innan and Kim 2004).

The difference that is expressed in these two views could have important evolutionary consequences. If adaptations start out as new mutations the rate of the adaptive process is limited by the rates and effects of beneficial mutations. In contrast, if a large part of adaptive substitutions derives from standing genetic variation, the adaptive course is modulated by the quality and amount of the available genetic variation. Because this variation is shaped by previous selection, the future course of evolution will not only depend on current selection pressures, but also on the history of selection pressures and environmental conditions that the population has encountered. Clearly, quite different sets of parameters could be important under the two scenarios if we want to estimate past and future rates of evolution. In order to assess which alternative is more prevalent in nature, population genetic theory can be informative in two ways. First, it allows us to determine the probabilities for selective adaptations in both scenarios. Second, theory can be used to predict whether and how these different modes of adaptation can be detected from population data. In this article, we address these issues in a model of a single locus.

We study the fixation process of an allele that is beneficial after an environmental change, but neutral or deleterious under the previous conditions. The population may experience a bottleneck following the shift of the environment. Assuming that the allele initially segregates in the population at an equilibrium of mutation, selection, and drift, we calculate the probability that it spreads to fixation after positive selection begins. We compare this probability with the fixation rate of the same allele, given that it only appears after the environmental change as a new mutation. This allows us to determine the parameter space, in terms of mutation rates, selection coefficients and the demographic structure, where a substitution that is observed some time after an environmental change is most likely from the standing genetic variation. We also analyze how the distribution of the effects of adaptive substitutions changes if the standing genetic variation is a source of adaptive material. Our main finding is that adaptations with a small effect are much more frequent in this case than predicted in a model that only considers adaptations from new mutations.

We then ask whether adaptations from standing genetic variation can be detected from the sweep pattern on linked neutral variation. If a selective sweep originates from a single new mutation, all ancestral neutral variation that is tightly linked to the selected allele will be eliminated by hitch-hiking. We call this scenario a *hard sweep* in contrast to a *soft sweep* where more than a single copy of the allele contributes to an adaptive substitution. The latter may occur if the selected allele is taken from the standing genetic variation, where more than one copy is available at the start of the selective phase, or if new beneficial alleles occur during the spread to fixation. With a soft sweep, part of the linked neutral variation is retained in the population even close to the locus of selection. We calculate the probability for soft sweeps under both scenarios of the adaptative process and discuss the impact on the sweep pattern. We find that soft sweeps are likely for alleles with a high fixation probability from the standing variation, in particular for alleles that are under strong positive selection. Already for moderately high mutation rates, however, fixation of multiple independent copies is also likely if the selected allele only enters the population as a recurrent new mutation. We therefore predict that unusual sweep patterns compatible with soft sweeps may be frequent under biologically realistic conditions, but they cannot be used as a clear indicator of adaptation from standing genetic variation.

1.2 Model and Methods

Assume that a diploid population of effective size N_e experiences a rapid environmental shift at some time T that changes the selection regime at a given locus. We consider two alleles (or classes of physiologically equivalent alleles) at this locus, a and A. a is the ancestral "wildtype" allele and A derived, in the sense that the population was never fixed for A prior to T. A is favorable in the new environment with homozygous fitness advantage s_b. The dominance coefficient is h, i.e. the heterozygous fitness is $1 + hs_b$. Assuming that the population was well-adapted in the old environment, A was either effectively neutral or deleterious before T, with selection coefficient s_d measuring its homozygous disadvantage and dominance coefficient h'. A is generated from a by recurrent mutations at rate u. In the following, it will be convenient to work with scaled variables for selection and mutation, defined as $\alpha_b = 2N_e s_b$, $\alpha_d = 2N_e s_d$, and $\Theta_u = 4N_e u$. We will initially assume that the population size N_e stays constant over the time period under consideration, but relax this condition later. We restrict our analysis to a single adaptive substitution, which is studied in isolation. This assumption means that different adaptive events do not interfere with each other due to either physical linkage or epistasis.

Simulations

We check all our analytical approximations by full-forward computer simulations. For this, a Wright-Fisher model with $2N_e$ haploid individuals is simulated. Every generation is generated by binomial or multinomial sampling, where the probability of choosing each type is weighted by its respective fitness. No dominance is assumed ($h = h' = 0.5$) and $2N_e$ is 50000. Data points are averaged over at least 12000 runs for $\Theta_u = 0.4$ and all data points in Fig. 1.6, 20000 runs for $\Theta_u = 0.04$, and 40000 runs for $\Theta_u = 0.004$.

Each simulation is started $6N_e = 150000$ generations before time T in order to let the population reach mutation-selection-drift equilibrium. Longer initial times did not change the results in trial runs. At the start, the population consists of only ancestral alleles "0", the derived allele "1" is created by mutation. Whenever the derived allele reaches fixation by drift, it is itself declared "ancestral", i.e. the population is set back to the initial state.

After $6N_e$ generations, the selection coefficient of the derived allele changes from neutral or deleterious (s_d) to beneficial (s_b). Mutations now convert ancestral alleles into new derived alleles (using a different symbol "2") with the same selection coefficient s_b. Simulations continue until eventual loss or fixation of the ancestral allele, where new mutational input is stopped $G = 0.1N_e = 2500$ generations after the environmental change. Each run has four possible outcomes: Fixation of "0", "1", or "2", or of "1" and "2" together.

Bottleneck: In the bottleneck scenario, the population is reduced to 1% at time T ($N_T = 250$). After time T, the population is allowed to recover logistically following $N_{t+1} = N_t + rN_t(1 - N_t/K)$ where $r = 5.092 \cdot 10^{-2}$ and the carrying capacity is $K = 2546$. This results in an average population size of $N_{\mathrm{av}} = 2500$ (10% of the original size) after the environmental change until new mutational input is stopped at $G = 0.1N_e$ generations. For $\Theta_u = 0.004$ only realizations with more than 10 fixation events in 40000 runs are included in the figures.

Number of (independent) copies: To determine the number of independent copies that contribute to a fixation, each mutation is given a different name and followed separately. Runs are done with and without new mutational input after the environmental shift and continued until fixation of the selected allele or all copies from the standing variation are lost. Additionaly, also runs with only new mutations are done. When fixation of the selected allele occures, we count the number of descendents from different origins in the population. A similar procedure is followed to determine the number of copies from the standing variation that contribute to a substitution. For this, all copies of the selected allele that are present at the time of the environmental change are given a different name. In the case of fixation, the number of different copies in the population is counted. Only realizations with more than 10 fixtions are included in the figures.

1.3 Results

Fixation probability from the standing genetic variation

The fixation probability of an allele A with selective advantage s_b that segregates in a population at frequency x is given by Kimura's diffusion approximation result (Kimura 1957)

$$\Pi_x(\alpha_b, h) \approx \frac{\int_0^x \exp[-\alpha_b(2hy + (1-2h)y^2)]dy}{\int_0^1 \exp[-\alpha_b(2hy + (1-2h)y^2)]dy}\,. \tag{1.1}$$

In the following, we will assume that selection on the heterozygote is sufficiently strong (formally, we need that $2h\alpha_b \gg (1-2h)/2h$). We can then ignore the term proportional to y^2 in Eq. (1.1) and Π_x is approximately

$$\Pi_x(h\alpha_b) \approx \frac{1-\exp[-2h\alpha_b x]}{1-\exp[-2h\alpha_b]} \,. \tag{1.2}$$

If A enters the population as a single new copy, $x = 1/2N_e$, and if $2N_e \gg 2h\alpha_b \gg 1$, we recover Haldane's classic result that the fixation probability is twice the heterozygote advantage, $\Pi_{1/2N_e} \approx 2hs_b$ (Haldane 1927). This relation underlines the importance of genetic drift: It is not sufficient for an advantageous allele to arrive in a population, it also needs to escape stochastic loss. Due to the strong linear dependence of the fixation probability on the selection coefficient, alleles with a small beneficial effect are less likely to escape such loss. The fixation process thus acts like a stochastic sieve that favors adaptations with large effects. This was stressed in particular by Kimura (1983). According to Eq. (1.2), an approximately linear dependence of Π_x on $h\alpha_b$ holds more generally as long as either the initial frequency x or the heterozygote advantage $h\alpha_b$ are small, such that $2h\alpha_b x < 1$.

Let us now compare this view of the fixation process with the alternative scenario of adaptation from the standing genetic variation. In the most simple case, the allele A again originates from a single mutation, but *before* the environmental change, and already segregates in the population under neutrality when positive selection sets in. Standard results (e.g. Ewens 2004) show that under these conditions the probability for an allele to segregate at a given frequency is proportional to the inverse of the frequency, $\rho(x_k) = a_{N_e}^{-1}k^{-1}$, where $x_k = k/2N_e$ and $a_{N_e} = \sum_{k=1}^{2N_e-1}(1/k)$. The average fixation probability then is $\Pi_{\text{seg}} = \sum_{k=1}^{2N_e-1}\Pi_{x_k}\rho(x_k)$. We derive an exact result for Π_{seg} in terms of a hypergeometric function in the Appendix; for $2N_e \gg 2h\alpha_b \gg 1$ we obtain the approximation

$$\Pi_{\text{seg}}(h\alpha_b, N_e) \approx 1 - \frac{|\ln(2hs_b)|}{\ln(2N_e)} = \frac{\ln(2h\alpha_b)}{\ln(2N_e)} \,. \tag{1.3}$$

We can make two interesting observations from this result. First, as may be seen from Fig. 1.1, there is a large increase in the (average) fixation probability if an allele does not arise as a single new copy, but already segregates in the population. This increase is particularly large for small adaptations, which points to the second observation: For alleles from the standing genetic variation, the fixation probability depends only weakly (logarithmically) on the

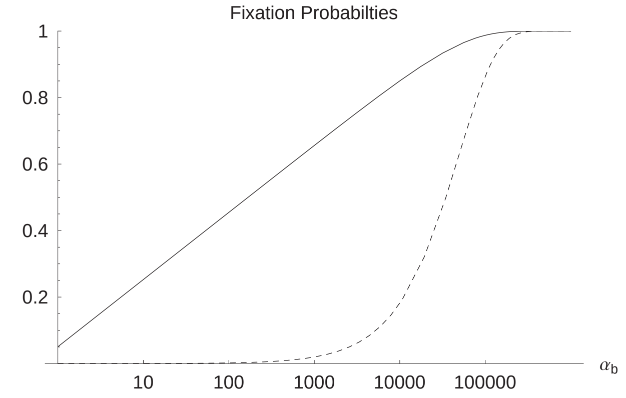


Figure 1.1: Fixation probabilities from a single new mutation (dashed line) and from a single segregating allele (solid line). Note that α_b is measured on a logarithmic scale.

selection coefficient. Indeed, Π_{seg}, unlike Π_x, does not show a linear dependence on $h\alpha_b$ even if $h\alpha_b$ is very small. The reason is that, *conditioned on later fixation*, the average frequency of the allele at the time of the environmental change, $\bar{x}_k$, increases with decreasing $h\alpha_b$, such that $2h\alpha_b\bar{x}_k > 1$ for all $h\alpha_b$ (a simple calculation in the Appendix reveals that $\bar{x}_k \approx 1/\ln(2h\alpha_b)$). The usual linear approximation of Π_x is therefore never appropriate.

Consider, now, an allele A that segregates in the population at an equilibrium of mutation, (negative) selection, and drift when the environment changes at time T. For $t > T$, positive selection sets in. We are interested in the net probability P_{sgv} that the allele is available in the population at time T *and* subsequently goes to fixation. In the continuum limit for the allele frequencies, P_{sgv} is given by the integral

$$P_{\text{sgv}} = \int_0^1 \rho(x)\Pi_x dx \tag{1.4}$$

where Π_x is the fixation probability (Eq. 1.2) and $\rho(x)$ is the density function for the frequency of a derived allele in mutation-selection-drift balance. Approximations for $\rho(x)$ can be obtained from standard diffusion theory; all derivations are given in the Appendix. In the neutral case ($\alpha_d = 0$) the distribution of derived alleles is approximately

$$\rho(x) \approx C_0 x^{\Theta_u - 1} \frac{1 - x^{1-\Theta_u}}{1-x}\,. \tag{1.5}$$

For a previously deleterious allele, and $2h'\alpha_d \gg (1-2h')/2h'$, we obtain

$$\rho(x) \approx C_\alpha x^{\Theta_u - 1} \exp(-2h'\alpha_d x) \frac{1-\exp[2h\alpha_d(x-1)]}{1-x} \,. \tag{1.6}$$

C_0 and C_α are normalization constants. $\rho(x)$ includes a probability Pr_0 that A is not present in the population at time T. For $\Theta_u < 1$, this probability is approximately

$$\mathrm{Pr}_0(h'\alpha_d, N_e) \approx \left(\frac{2N_e}{2h'\alpha_d+1}\right)^{-\Theta_u} = \exp\Big(-\Theta_u \ln[2N_e/(2h'\alpha_d+1)]\Big) \,. \tag{1.7}$$

For the probability that the population successfully adapts from the standing variation we derive the following simple approximation

$$P_{\mathrm{sgv}}(h\alpha_b, h'\alpha_d, \Theta_u) \approx 1 - \left(1 + \frac{2h\alpha_b}{2h'\alpha_d+1}\right)^{-\Theta_u} = 1 - \exp\Big(-\Theta_u \ln[1+R_\alpha]\Big) \,, \tag{1.8}$$

where $R_\alpha := 2h\alpha_b/(2h'\alpha_d+1)$ is the *relative selective advantage.* R_α measures the selective advantage of A in the new environment relative to the forces that cause allele frequency changes in the ancestral environment, deleterious selection and drift (represented by the 1). We will refer to $R_\alpha < 1$ and $R_\alpha > 1$ as cases of small and large relative advantage, respectively. If the allele A is completely recessive in the old environment ($h' = 0$), similar approximations hold here and below if $2h'\alpha_d + 1$ in R_α is formally replaced by $\sqrt{\alpha_d} + 1$ (see again the Appendix for details). In order to relate Eq. (1.8) to Eq. (1.3), we need to calculate the fixation probability for a segregating allele that is derived from a single mutation prior to the environmental change. This probability is obtained from (1.8) and (1.7) by conditioning on segregation of the allele in the limit $\Theta_u \to 0$. We find

$$\Pi_{\mathrm{seg}}(h\alpha_b, h'\alpha_d, N_e) \approx \frac{\ln[1+R_\alpha]}{\ln[2N_e/(2h'\alpha_d+1)]} \tag{1.9}$$

For $\alpha_d = 0$ and $h\alpha_b \gg 1$ this reduces to Eq. (1.3).

All further results of our study depend on Eq. (1.8). Computer simulations show that this simple analytical expression is quite accurate over a large parameter range (assuming $\Theta_u < 1$ and $h\alpha_b, h'\alpha_d \ll 2N_e$; see Figure 1.2). Slightly better approximations (which coincide with 95% confidence intervals of all our simulation runs) can be obtained by numerical integration of Eq. (1.4) using the allele frequency distributions Eq. (1.5) and Eq. (1.6). It is instructive to compare the stochastic result Eq. (1.8) with the deterministic approximation used by Orr and Betancourt (2001). If we set $x \equiv \Theta_u/2h'\alpha_d$ in Eq. (1.2) (the equilibrium value at mutation-selection balance), the fixation probability from the standing variation becomes

$$P_{\mathrm{sgv}}(h\alpha_b, h'\alpha_d, \Theta_u) \approx 1 - \exp(-\Theta_u h\alpha_b/h'\alpha_d)\,. \tag{1.10}$$

Eq. (1.8) reduces to Eq. (1.10) if and only if there is relatively strong past deleterious selection such that $R_\alpha \ll 1$. In this limit, the initial frequency of the selected allele is sufficiently reduced that the fixation probability Π_x (Eq. 1.2) is approximately linear in x over the range of $\rho(x)$, $\Pi_x \approx 2h\alpha_b x$. In the integral (1.4) then only the average allele frequency $\bar{x}$ enters, which (almost) coincides with the deterministic approximation. For $R_\alpha \geq 1$, the distribution $\rho(x)$ feels the concavity of Π_x and the true value of P_{sgv} drops below the deterministic estimate. This is captured by Eq. (1.8), see Fig. 1.2. For $R_\alpha \leq 1$ the fixation probability does not approach the "deterministic" approximation even if N_e, and thus α_d, α_b and Θ_u, get large. The reason is that it is the variance of $2h\alpha_b x$ that matters, which does not go to zero even if the variance of the allele frequency $\mathrm{Var}[x] \to 0$ for large Θ_u and α_d.

Eq. (1.8) and Eq. (1.9) confirm a weak dependence of the fixation probability on α_b. For fixed α_d, the fixation probability depends logarithmically on α_b (and on R_α) as long as $R_\alpha > 1$. In the "deterministic limit" $R_\alpha \ll 1$, this dependence goes back to linear. However, this is only true if α_b varies independently of α_d. If stronger selected alleles have larger trade-offs, i.e. α_b and α_d are positively correlated, R_α and thus P_{sgv} and Π_{seg} will increase less than linearly with α_b even if $R_\alpha \ll 1$. Using the deterministic aproximation, Orr and Betancourt (2001) previously found that the dominance coefficient drops out of P_{sgv} if dominance does not change upon the environmental shift, $h = h'$. The stochastic result Eq. (1.8) confirms this finding and extends it beyond the limits of validity of the deterministic approximation as long as $h\alpha_b$ and $h'\alpha_d$ are both large.

Standing variation versus new mutations

We want to compare the fixation probability from the standing variation with the probability that an adaptive substitution occurs from new mutation. The probability for a new allele to occur in the population that is destined for fixation is approximately $p_{\mathrm{new}} = 2N_e u 2hs_b$ per generation. Using a Poisson approximation, the probability that such a mutation arrives within G genera-

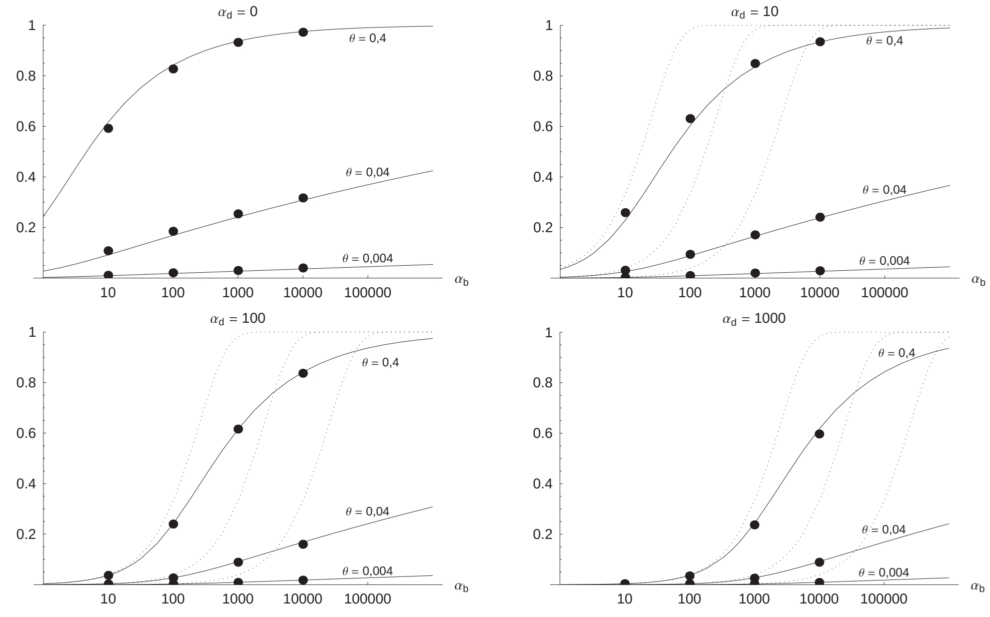


Figure 1.2: The probability of fixation from mutation-selection-drift balance, P_{sgv}, for a range of mutation and selection parameters. Solid lines show approximation Eq. (1.8), dotted lines show the deterministic approximation Eq. (1.10). Large dots are simulation results. 95% confidence intervals are contained in the symbols.

tions is

$$P_{\text{new}}(G) = 1 - \exp[-\Theta_u h\alpha_b G], \tag{1.11}$$

where G is measured in units of $2N_e$. We can now determine the number of generations G_{sgv} that it takes for $P_{\text{new}}(G_{\text{sgv}}) = P_{\text{sgv}}$. This value serves as a measure of the relative adaptive potential of the standing variation. Using Eq. (1.8) we obtain

$$G_{\text{sgv}}(h\alpha_b, h'\alpha_d) \approx \frac{\ln[1 + R_\alpha]}{h\alpha_b}\,. \tag{1.12}$$

This value is independent of Θ_u and depends only on the selection parameters of the allele. One can relate G_{sgv} to the average fixation time t_{fix} of an allele with selective advantage $h\alpha_b$. In the Appendix, we derive t_{fix} in units of $2N_e$,

$$t_{\text{fix}}(h\alpha_b) \approx \frac{2(\ln[2h\alpha_b] + 0.577 - (2h\alpha_b)^{-1})}{h\alpha_b}\,. \tag{1.13}$$

The approximation is very accurate for $h = 0.5$ and $h\alpha_b \gtrsim 2$. For $h \neq 0.5$ it defines a lower bound. We see that $G_{\text{sgv}} < t_{\text{fix}}$ for arbitrary R_α. This holds even if we account for the fact that the average fixation time from the standing variation may be shorter (but $\geq t_{\text{fix}}/2$), since the allele starts at a higher frequency. This result means that in a time span that an allele from the standing variation needs to reach fixation, it is at least as likely that the allele alternatively appears as a new mutation destined for fixation only after the environmental change.

Next, we consider the case that a derived beneficial mutation A is found in a population some time after the environmental change. There are three possibilities: Either A derives from the standing genetic variation at time T, or from new mutation(s) that occured after the environmental change, or both. Computer simulations that include new mutations after time T show that hybrid fixations that use material from both sources are quite frequent for high Θ_u, but also that the contribution of the standing variation generally dominates in this case (for $\Theta_u = 0.4$ on average $67\% - 97\%$, depending on α_b and α_d). In the following, we combine hybrid fixations with fixations that use only alleles from the standing variation and define P_{sgv} more broadly as the probability that an adaptive substitution uses material from the standing genetic variation. With this definition, simulation results are closely matched by the theoretical prediction in Eq. (1.8).

We can now ask for the probability that a derived allele A, which is found in the population some time G after T, and either fixed or destined to go to fixation at this time, originated (at least partially) from alleles in the standing genetic variation. Measuring G in units of $2N_e$ generations, this probability may be expressed as $\text{Pr}_{\text{sgv}} = P_{\text{sgv}}/(P_{\text{sgv}} + (1 - P_{\text{sgv}})P_{\text{new}})$. With Eq. (1.8),

$$\text{Pr}_{\text{sgv}}(\alpha_b, \alpha_d, \Theta_u) \approx \frac{1 - \exp\{-\Theta_u \ln[1 + R_\alpha]\}}{1 - \exp\{-\Theta_u(\ln[1 + R_\alpha] + h\alpha_b G)\}}\,. \tag{1.14}$$

In Figure 1.3, this is shown for $G = 0.05$, i.e. for a time of $0.1N_e$ generations after the environmental change. This time should be sufficiently long for significant adaptive change, but still short enough for a selective sweep to be detected in DNA sequence data (Kim and Stephan 2000; Przeworski 2002). For *Drosophila melanogaster*, $0.1N_e$ generations approximately corresponds to the time since it expanded its range out of Africa into Europe after the last glaciation (i.e. about $10,000 - 15,000$ years ago).

There are two advantages of the standing variation over adaptations purely from new mutations. First, the standing genetic variation may already contain multiple copies of the later-beneficial allele, reducing the probability of a stochastic loss relative to a single copy. This advantage is measured in the relative adaptive potential G_{sgv} above. A second, independent advantage is

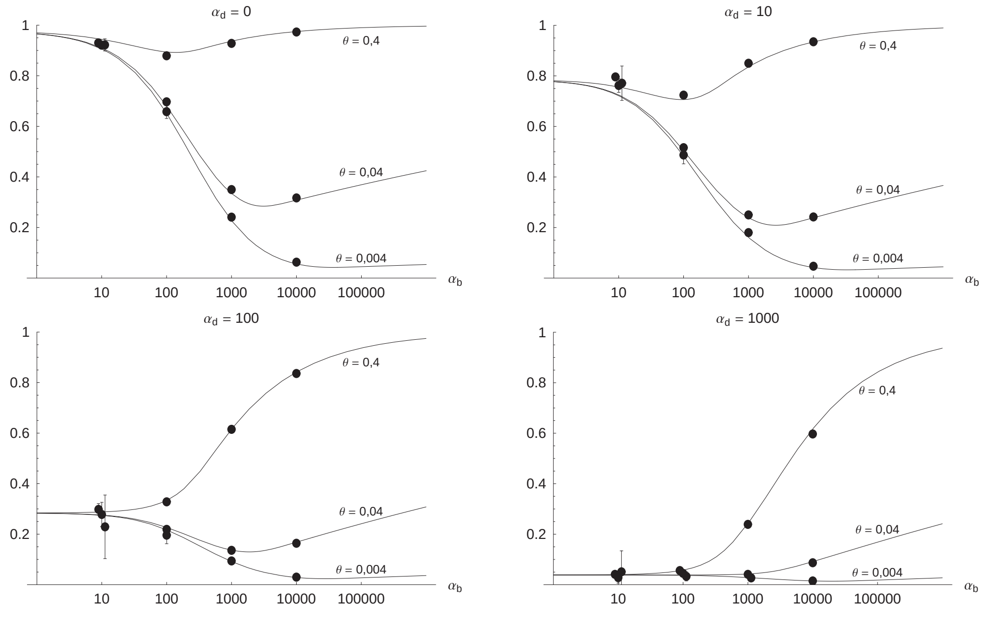


Figure 1.3: The probability that an adaptive substitution is from the standing genetic variation ($\mathrm{Pr}_{\mathrm{sgv}}$). Simulation data with 95% confidence intervals are compared to the analytical approximation Eq. (1.14).

that alleles from the standing variation are immediately available and may outcompete new mutations due to this headstart. Consequently, we see that substitutions from the standing variation dominate in two parameter regions. First, they dominate for small $h\alpha_b$ as long as selection before the environmental change was also weak because $P_{\mathrm{sgv}} > P_{\mathrm{new}}$ in this range. (P_{sgv} is larger than P_{new} for $h\alpha_b < \ln[1 + R_\alpha]/G$; for small $h\alpha_b$ this needs $h'\alpha_d < 1/G$, i.e. $\alpha_d < 40$ for $h' = 0.5$ and $G = 0.1N_e$). The second parameter region is if $h\alpha_b$ and the mutation rate Θ_u are both high. In this case, the crucial advantage of the alleles from the standing genetic variation is their immediate availability: The probability for fixation from the standing variation is already sufficiently high that there is no need to wait for a new mutation to occur.

For practical application of this result, remember that $\mathrm{Pr}_{\mathrm{sgv}}$ does not only count alleles that are fixed at time $T + G$, but also alleles that are destined to go to fixation. Consequently, simulations in Fig. 1.3 are continued until loss or fixation of the allele even beyond $T+G$. This makes almost no difference as long as the average fixation time t_{fix} of an allele is much smaller than G. However, if $t_{\mathrm{fix}} \geq G$, Eq. (1.14) can no longer be used to predict full substitutions. For

$G = 0.1N_e$, $t_{\text{fix}} > G$ if $h\alpha_b \lesssim 275$. If we only count substitutions that are completed at time $T + G$, P_{new} is more strongly reduced than P_{sgv}. For alleles with $t_{\text{fix}} \approx G$, predominance of the standing genetic variation is larger than predicted by Eq. (1.14) (confirmed by simulations, results not shown). For alleles with $t_{\text{fix}} \gg G$ practically all substitutions that are completed at time $T + G$ contain material from the standing variation; however, there are then only very few fixations at all.

Population bottlenecks

So far, we have assumed that the effective population size before, during, and after the environmental change is constant. For many evolutionary scenarios, however, it may be more realistic to assume that the shift of the environmental conditions is accompanied by a population bottleneck. Examples include colonization events and human domestication, but also the (temporary) reduction of the carrying capacity of a maladapted population in a changed environment.

Suppose that a population of ancestral size N_0 goes through a bottleneck directly after the environmental change and recovers afterwards until it reaches its carrying capacity in the new environment. We want to know how these demographic events change the probability Pr_{sgv} that a substitution is derived from the standing genetic variation. We expect two factors to play a role. On the one hand, a deep and long-lasting bottleneck may significantly reduce the standing variation and the potential of the population to adapt from it. On the other hand, a slow or incomplete recovery reduces the opportunity for new mutations to arrive in the population and thus the probability of adaptation from new mutations.

It is therefore instructive to distinguish two elements of a bottleneck, population size reduction and subsequent recovery, and discuss their effects separately. The simplest case is a pure reduction of N_0 by a factor $B > 1$ at time T, with no recovery. For matters of comparison, we continue to use the ancestral population size N_0 in the definitions of $\Theta_u, \alpha_b, \alpha_d$, and G. In our formulas for the fixation probabilities from new or standing variation (Eqs. 1.8, 1.11, and 1.14) population size reduction is then simply included by a rescaling of the selection parameter α_b to α_b/B. (For adaptations from the standing genetic variation note that a sampling step to generate a bottleneck does not change the frequency distribution of the later-beneficial allele, leaving α_b in Eq. (1.2) the only parameter subject to change. For adaptation from new mutations the rescaling argument follows if we express the probability for a new mutation destined for fixation per generation as $p_{\text{new}} = (2N_e/B)u2hs_b = 2uh\alpha_b/B$.)

Consequently, the graphs in Fig. 1.3 are simply shifted to the right. A pure reduction of the population size at time T thus reduces the relative advantage of the standing genetic variation for strongly selected alleles with a large mutation rate, but enhances its advantage for weakly selected alleles. Note that the adaptive potential G_{sgv} increases by a factor of B relative to t_{fix} and can now be much larger than the fixation time.

Relative to a simple reduction in population size, recovery increases the adaptation probability from the standing variation, P_{sgv}, and from new mutations, P_{new}, in different ways. First, recovery increases P_{new} (but not P_{sgv}) simply due to the fact that the opportunity for new mutations increases with increasing population size. Second, the fixation probability of beneficial alleles is increased due to population growth. For further progress, we use results on the fixation probability in populations of changing size by OTTO and WHITLOCK (1997). We assume that the population experiences logistic growth according to $dN/dt = \lambda(1 - N/K)N$ after an initial reduction to N_T. Here, λ is the intrinsic growth rate (for t in units of $2N_0$), and K the carrying capacity. There are two things to note. First, the effect of recovery on the fixation probability is only significant if it is sufficiently fast on a scale set by the selection strength. For logistic recovery, this is the case if $\lambda \gtrsim h\alpha_b$. Second, the increase of the fixation probability due to recovery is much more important for P_{sgv} than for P_{new}. The reason is that only alleles that are already present during the bottleneck will be affected. While this is the case for all alleles from the standing variation that survive population size reduction, only relatively few new mutation will occur in the small bottleneck population (at least if recovery is sufficiently fast to matter). More formally, one can show that the increase in the fixation probability due to recovery can be neglected in P_{new} if $\lambda G \gg 1$. This leaves only a very restricted parameter space of $h\alpha_b \lesssim \lambda \lesssim 1/G$ where an increase in fixation probability plays a role for P_{new} (confirmed by simulations, not shown).

In the following, we concentrate on fast recovery on a scale of G, i.e. $\lambda \gg 1/G$ (results for slow recovery are intermediate between fast and no recovery). As a measure for the opportunity for new beneficial mutations to arrive in the population, let N_{av} be the average population size from time T to time $T + G$ where the substitutions are censused. We then define a bottleneck parameter for new mutations $B_{\text{new}} := N_0/N_{\text{av}}$ and rescale α_b to α_b/B_{new} in P_{new} (Eq. 1.11). For fixations from the standing genetic variation, we define the bottleneck strength as $B_{\text{sgv}}(h\alpha_b) = N_0/N_{\text{fix}}(h\alpha_b)$ and rescale the relative selection strength $R_\alpha \to R_\alpha/B_{\text{sgv}}$ in Eq. (1.8) and (1.14). Here, N_{fix} is an average "fixation effective population size" that is felt by a beneficial allele on its way to fixation or loss. Since the sojourn time of a strongly selected allele is shorter than of a weakly selected allele, N_{fix} and B_{sgv} depend on the selection coefficient of the allele. For logistic growth, Eq. (19) in Otto and Whitlock (1997) leads to

$$B_{\text{sgv}}(h\alpha_b) = \frac{N_0}{N_T} \cdot \frac{h\alpha_b + \lambda N_T/K}{h\alpha_b + \lambda}\,. \tag{1.15}$$

Fig. 1.4 shows the precentage of fixations from the standing variation for a bottleneck with $N_T = N_0/100$ and logistic recovery with about 5% initial growth per generation and carrying capacity $K = 2546$. More precisely, we choose $\lambda = 0.05092 \cdot 2N_0 = 2546$ for the growth rate per $2N_0 = 50000$ generations, such that the average size after the environmental change until $0.1N_0$ generations (i.e. $G = 0.05$) is $N_{\text{av}} = N_0/10 = 2500$.

From Eq. (1.15) and Fig. 1.4, we can distinguish three parameter regions

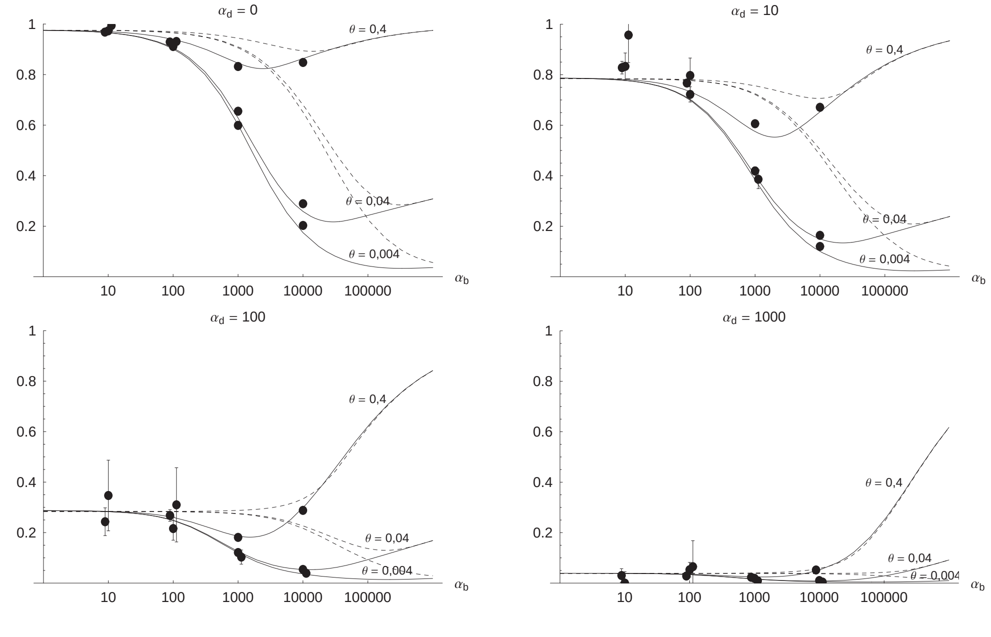


Figure 1.4: The probability that an adaptive substitution stems from the standing genetic variation Pr_{sgv} in a population with a bottleneck at the time of the environmental change. Dashed lines show a simple reduction in population size by a factor 100 without recovery. Simulation dots and solid lines are for the opposite case of strong logistic recovery (parameters see main text). The lines follow from the simple analytical approximation Eq. (1.14) with the bottleneck correction $R_\alpha \rightarrow R_\alpha/B_{\text{sgv}}$ and $\alpha_b \rightarrow \alpha_b/B_{\text{new}}$ in the term proportional to G. Direct numerical integration of Eq. (1.5) and Eq. (1.6) with the same bottleneck correction produces a slightly better fit.

for the effect of a bottleneck. Firts, for $h\alpha_b > \lambda$, the fixation probability of individual alleles is not substantially increased by population growth as compared to the case without recovery. However, population growth increases the opportunity for new mutations and thus $B_{\text{new}} < B_{\text{sgv}}$. For large Θ_u, there is nevertheless almost no change in Pr_{sgv} relative to no recovery. The reason is that fixation is then almost certain, with $P_{\text{new}} \approx 1$ and thus $\text{Pr}_{\text{sgv}} \approx P_{\text{sgv}}$ (see the definition of Pr_{sgv} above Eq. (1.14)). Second, for very small selection coefficients, $h\alpha_b < \lambda N_T/K$, all alleles feel the new carrying capacity K as their "fixation effective population size". If $\lambda \gg 1/G$, the bottleneck then acts like a single change in the population size from N_0 to K. Finally, for intermediate selection coefficients, P_{new} generally profits more from the recovery than P_{sgv}, leading to a reduction in Pr_{sgv} if compared to no recovery.

Compared with the results of the previous section, we can summarize the effect of a bottleneck as follows. There is a tendency to further increase the predominance of the standing variation for weakly selected alleles, and to decrease its advantage for high $h\alpha_b$ and Θ_u. However, unless the bottleneck is very strong, there is no qualitative change in the overall pattern.

Footprints of soft sweeps

Since adaptations from the standing genetic variation start out with a higher copy number of the selected allele, more than one of these copies may escape stochastic loss and eventually contribute to fixation. Depending on whether one or multiple copies are involved in the substitution, one may expect differences in the footprint of the adaptation on linked neutral variation. In order to derive the probability that n copies of the allele A that segregate in the population at time T contribute to its fixation, we follow Orr and Betancourt (2001) and assume that individual copies enjoy an independent probability to escape stochastic loss. We may then apply a Poisson approximation. If the frequency of A at the time of the environmental change is x, the probability that $k = n$ copies survive and contribute to fixation is approximately

$$Pr(k = n; x) = \exp[-2h\alpha_b x]\,\frac{(2h\alpha_b x)^n}{n!}\,. \tag{1.16}$$

This approximation is consistent with Eq. (1.3) if $2h\alpha_b \gg 1$. The probability that more than one copy contributes to the substitution (i.e. the probability for a "soft sweep") then is $Pr(k > 1; x) = 1 - (1 + 2h\alpha_b x)\exp[-2h\alpha_b x]$. Averaging over the allele frequency distribution at time T, $\rho(x)$, and conditioning on the case that fixation did occur, we obtain the probability for a soft sweep for adaptations from the standing genetic variation,

$$P_{\text{mult}} \approx 1 - \frac{2h\alpha_b}{P_{\text{sgv}}} \int_0^1 x \exp[-2h\alpha_b x] \rho(x) dx \,. \tag{1.17}$$

Using the approximations Eq. (1.5) and Eq. (1.6) for the allele distribution, and Eq. (1.8) for P_{sgv}, this gives

$$P_{\text{mult}}(R_\alpha, \Theta_u) \approx 1 - \frac{\Theta_u R_\alpha/(1+R_\alpha)}{(1+R_\alpha)^{\Theta_u} - 1} \,. \tag{1.18}$$

which reduces to $P_{\text{mult}} \approx 1 - R_\alpha/((1+R_\alpha)\ln[1+R_\alpha])$ in the limit $\Theta_u \to 0$. This limit is essentially reached for $\Theta_u \lesssim 0.004$. We can again compare the stochastic result with the deterministic approximation that is obtained from Eq. (1.17) assuming $x \equiv \Theta_u/2h'\alpha_d$,

$$P_{\text{mult}} \approx \frac{\exp[\Theta_u h\alpha_b/h'\alpha_d] - 1 - \Theta_u h\alpha_b/h'\alpha_d}{\exp[\Theta_u h\alpha_b/h'\alpha_d] - 1} \approx \frac{1}{2}\Theta_u h\alpha_b/h'\alpha_d \,. \tag{1.19}$$

Both approximations Eq. (1.18) and Eq. (1.19) are compared to simulation data in Figure 1.5. The deterministic approximation reproduces the stochastic result only for very large mutation rates, $\Theta_u \gg 1$, outside the parameter space in the figure. For low mutation rates, where Eq. (1.19) predicts a zero limit for $\Theta_u \to 0$ it severely underestimates P_{mult}. The stochastic approximation produces a reasonable fit unless $h'\alpha_d$ and $h\alpha_b$ are both small. In this parameter range with relatively high initial allele frequency of the allele and weak positive selection, the Poisson approximation is no longer valid.

In order to estimate the impact of a soft sweep on linked neutral variation we are also interested in the number of *independent* copies that contribute to the fixation of the allele, i.e. copies that are not identical by descent. Concentrating on copies that segregate in the population at the time T of the environmental change, we can again use a Poisson approximation, $\tilde{Pr}(k = n) = \exp(-\lambda)\lambda^n/n!$. With this conjecture, $1 - \exp(-\lambda)$ is the fixation probability from the standing genetic variation. Equating with P_{sgv} as given in Eq. (1.8), we obtain $\lambda = \Theta_u \ln[1 + R_\alpha]$. The probability of fixation of multiple independent copies, conditioned on the cases where fixation occurs then is

$$P_{\text{ind}}(R_\alpha, \Theta_u) \approx 1 - \frac{\Theta_u \ln[1+R_\alpha]}{(1+R_\alpha)^{\Theta_u} - 1} \,. \tag{1.20}$$

Alternatively, we obtain Eq. (1.20) from Eq. (1.18) using the relation $1 - P_{\text{mult}}(\Theta_u) = (1 - P_{\text{ind}}(\Theta_u))(1 - P_{\text{mult}}(\Theta_u = 0))$. This equation expresses the

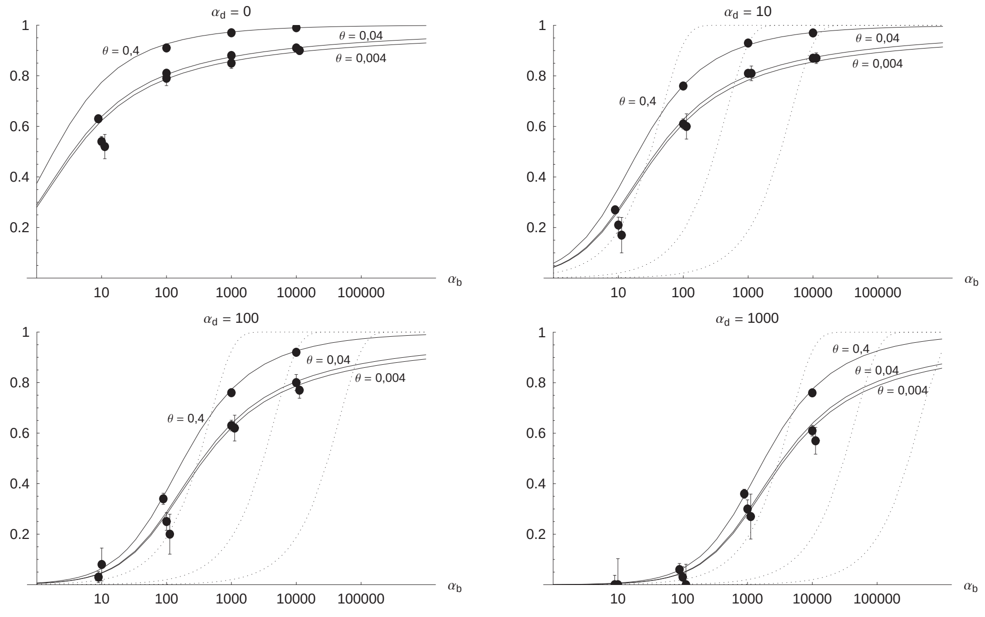


Figure 1.5: The probability, P_{mult}, that multiple copies from the standing genetic variation contribute to a substitution. Solid lines correspond to the approximation Eq. (1.18), dotted lines to the deterministic approximation Eq. (1.19).

probability for fixation of a single copy ("no multiple fixation given fixation") as the probability of fixation from a single origin times the probability of fixation of a single copy given that all successful copies are from a single origin (a single origin is enforced in P_{mult} by $\Theta_u \to 0$). This alternative derivation shows that Eq. (1.18) and Eq. (1.20) follow from the same assumption: independent fixation probability for different copies. To the order of our approximation, P_{mult} and P_{ind} depend on selection only through the relative selective advantage $R_\alpha = 2hs_b/(2h's_d + 1/(2N_e))$. This parameter combines two effects. The denominator of R_α takes into account that multiple fixations are less likely if the initial frequency of the allele at time T is low. This frequency decreases with deleterious selection $h's_d$ and drift, represented by the $1/2N_e$ term. Secondly, the numerator of R_α accounts for the fixation probability of the allele: The probability that the allele is maintained during the adaptive phase increases with hs_b. For $h\alpha_d \gg 1$, the result depends only on the ratio of the selection coefficients as also predicted by the deterministic approximation (Orr and Betancourt 2001). If the environmental change is followed by a bottleneck, Eq. (1.18) and Eq. (1.20) can be used with $R_\alpha \to R_\alpha/B_{\text{sgv}}$

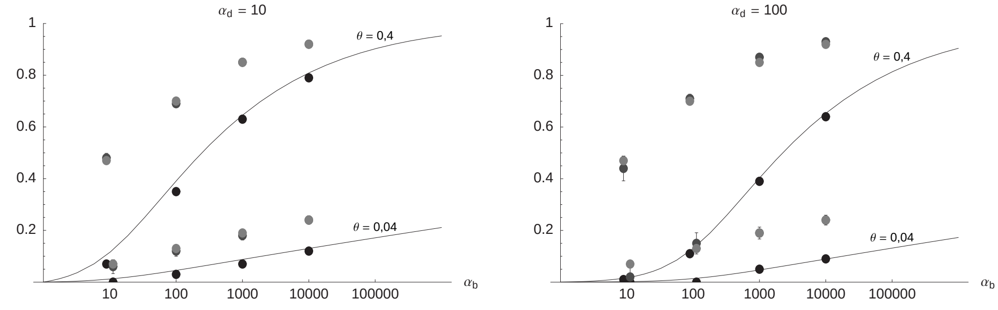


Figure 1.6: The probability that multiple copies with independent origin contribute to a substitution, P_{ind}. Black simulation dots are for fixations from the standing variation without new mutational input after time T, dark grey dots include new mutations. Light grey dots are for fixations from recurrent new mutations only. In several cases, light grey dots are exactly on top of dark grey dots. Lines correspond to the approximation Eq. (1.20).

with the bottleneck factor introduced above. In contrast to P_{mult}, the fixation probability of multiple independent copies depends strongly on the mutation rate Θ_u and vanishes for $\Theta_u \to 0$. In Fig. 1.6, Eq. (1.20) is compared with simulation data. The approximation produces a good fit for $\alpha_d \geq 10$ where the Poisson approximation is valid.

By construction, both approximations (1.18) and (1.20) account only for the fixation of copies of the allele that were already in the population at time T. It is, however, also possible that a successful copy first arises for $t > T$ as a new mutation during the adaptive phase. Since the origin of these new copies is necessarily independent, this effect contributes to P_{ind}. The size of this contribution depends on the population-level mutation rate $\Theta_{u,t>T}$ directly after the environmental change. $\Theta_{u,t>T}$ can be smaller than the original Θ_u that appears in Eqs. (1.18) and (1.20) if there is a bottleneck at T. For $\Theta_{u,t>T} = \Theta_u$ our simulation results show that the contribution of new mutations to P_{ind} is substantial (dark grey dots in Fig. 1.6). One consequence of mutational input after T is that P_{ind} becomes almost independent of α_d. Even more importantly, we see that the fixation of multiple independent copies is not particular to adaptations from the standing genetic variation. It occurs with basically the same probability if the selected allele enters the population only after the environmental change as a recurrent new mutation (see Fig. 1.6, light grey dots).

For recurrent new mutations, the simulation data show that the total fixation rate of multiple independent copies, $r_{\text{ind}} = -\ln[1 - P_{\text{ind}}]$, increases logarithmically with α_b and linearly with Θ_u. For a heuristic understanding of this dependence, assume $h = 0.5$ and let $x(t)$ be the frequency of a first copy of the selected allele on its way to fixation in absence of further mutation. For small u, the probability for a second copy of the beneficial mutation to arise while a first copy spreads to fixation then is $p_2 = 2N_e u \int_0^\infty (1 - x(t))dt = 2N_e u(t_{\text{fix}}/2)$. Here, t_{fix} is the average fixation time and we have used that the first copy spends on average equal times in frequency classes x and $(1 - x)$. By far the largest contribution to p_2 comes from the early phase of the sweep where the frequency x of the first copy is very low. The probability of the second copy to survive until fixation of the allele depends on x, but to leading order only the survival probability for $x \to 0$ matters, which is approximately s_b. With t_{fix} from Eq. (1.37) we then obtain $r_{\text{ind}} = \Theta_u \ln(\alpha_b) + \mathcal{O}\big(\alpha_b^0\big)$. A more detailed account will be given elsewhere.

P_{ind} is the probability that descendents of multiple independent copies of the selected allele segregate in the population at the time when this allele reaches fixation. Consequently, the number of copies in our simulation runs was counted at the time of fixation (same for P_{mult}). In practical applications, however, one is often interested in the probability of observing descendents from independent origins a fixed time G after an environmental change. This probability will decrease with G, since copies get lost by drift until, eventually (in the absence of back-mutation), all copies derive from a single mutation as their common ancestor. The drift phase from the time of fixation to the time of observation G depends on the selection coefficient and will be longer for strongly selected alleles with short fixation times. In principle, this could affect the dependence of the probability of observing multiple fixed copies in a population on $h\alpha_b$. In order to test this, we ran additional simulations to measure the probability for the survival of multiple (independent) copies $G = 0.1N_e$ generations after the environmental change (results not shown). For alleles with fixation time $t_{\text{fix}} < 0.1N_e$, we did not detect any difference to the data displayed in Fig. 1.5 and Fig. 1.6, meaning that fixation of a single copy in the neutral drift phase after initial fixation of multiple copies is rare. This is not surprising considering that the average fixation time under neutral drift exceeds $0.1N_e$ generations even if the frequency of the major copy is initially at 99%.

Another question is whether multiple copies of the selected allele are likely to be found in a small experimental sample, even if they exist in the population. We tested this by arbitrarily drawing 12 chromosomes in each case of a soft sweep. Multiple copies in the sample were found in 70% – 80% of all cases (for $\Theta_u = 0.4$). Summarizing our results for the fixation probabilities of multiple copies and of multiple independent copies, we can distinguish three parameter regions:

- Low mutation rate, relatively strong past selection. If the mutation rate is low ($\Theta_u \ll 0.1$) fixation of multiple independent copies of the selected allele is unlikely. If multiple copies fix, they are most likely identical by descent. If past deleterious selection is strong, however, also the fixation of multiple homologous copies is rare. For $\Theta_u = 0$, Eq. (1.18) indicates that less than 5% and less than 30% of fixations originate from multiple copies for $R_\alpha \leq 0.1$ and $R_\alpha = 1$, respectively (Fig. 1.5).

- Low mutation rate, relatively weak past selection. With increasing relative advantage R_α the fixation of multiple homologous copies increases. For $\Theta_u \to 0$, fixation of multiple copies occurs in more than 50% of the cases ($P_{\text{mult}} > 0.5$) if $R_\alpha \gtrsim 4$ (Fig. 1.5).

- High mutation rate. For mutation rates $\Theta_u \gtrsim 0.1$ fixations from independent origins are much more frequent and become more likely than the fixation of single copies. This holds true for whether the origin of the selected allele is from the standing variation or from recurrent new mutations. The fixation probability for multiple independent copies increases logarithmically with $h\alpha_b$. For $\Theta_u = 0.4$, 50% − 90% of a substitutions involve multiple independent copies (Fig. 1.6).

Imagine that we observe a DNA region where an adaptive substitution has happened following an environmental change at time T. Suppose that we observe this region G generations after the environmental change, and $2 \gg G \gg t_{\text{fix}}$, such that the advantageous allele has reached fixation, but G (in units of $2N_e$) is much shorter than the average neutral coalescent time. We want to analyze whether and how the contribution of multiple copies to an adaptive substitution affects the signature of selection on linked neutral variation. For this, it is helpful to distinguish two aspects of a selective footprint, its width in basepairs along the sequence and its maximum depth in terms of the extent of variation lost in a region close to the locus of selection.

For a hard sweep, the coalescent at the selected site itself does not extend beyond time T. Ancestral variation that has existed prior to T can only be maintained if there is recombination between the selected site and the site studied. In a core region around the selected site, where no recombination has happened, all ancestral variation is lost. Recombination therefore modulates the width of the sweep region, but in general does not affect its maximum depth. Since only recombination in the selective phase matters, and since the adaptive phase is much shorter for a strongly selected allele, the width of a selective footprint decreases with larger α_b.

For a soft sweep, the coalescent at the selected site itself extends into the ancestral environment. As compared with a hard sweep, a soft sweep therefore has a reduced maximum depth. Our results show that soft sweeps with shallower footprints are more likely for large α_b. This does not contradict the fact that selective footprints get weaker and eventually vanish as $\alpha_b \to 0$, for two reasons. First, even if it is more likely for lower α_b that all ancestral variation is eliminated close to the selection center, the width of the window where this holds true gets smaller at the same time. If this width drops below the average distance of polymorphic sites, the footprint of selection becomes undetectable. Second, if we observe the sweep region G generations after positive selection begins, we can only compare selective footprints of alleles that have reached fixation by this time. If we want to study very weakly selected alleles, G needs to be so large that any footprint of selection will be washed out by new mutations that arise after time T.

The impact of a soft sweep on the molecular signature depends on whether the surviving copies are independent by descent or not. Copies from different origins are related by a neutral coalescent and represent independent ancestral haplotypes. If these haplotypes are sampled close to the locus of selection, this should mark a clearly visible difference to the classic pattern of a hard sweep. A detailed quantitative analysis with estimates of the impact on summary statistics for nucleotide variability exceeds the aims of this study and will be given elsewhere.

If multiple surviving copies are identical by descent, the expected change in the molecular footprint relative to a hard sweep depends on the strength of deleterious selection that the allele has experienced prior to the environmental change. We expect a shallower footprint (and larger deviation from the hard sweep) for weaker deleterious selection. The reason is that it is more likely for a weakly deleterious allele to segregate in a population for a long time, i.e. the average time to the most recent common ancestor in the core region of the sweep is larger for smaller α_d. Indeed, this intuition can be made more precise.

A remarkable property of the Markov process that underlies the Wright-Fisher model is that, *conditional on* an allele A having reached some frequency x in a population, this process is independent of the *sign* of the selection coefficient of A (cf. Chap. 4.6 and 5.4 in Ewens 2004, for simplicity, we assume $\Theta_u = 0$ and $h = h' = 0.5$). This has interesting consequences for adaptations from mutation-selection-drift balance. Assume that an allele A with selective disadvantage s_d that is derived from a single mutation segregates in the population at frequency x at the time T of the environmental change. Then the mean age of this allele, and more generally, the average time that it spent in each frequency class in the past is the same as if it had a selective *advantage* of the same absolute size prior to T. Assume that A spreads to fixation under positive selection with selection coefficient s_b after the environmental change and compare this with a sweep of an (imaginary) allele A' with the same frequency x at time T, but selective advantage s_b throughout. For $s_d = s_b$, the total fixation time of the alleles, and their sojourn times in every frequency class are the same, for $s_d < s_b$ (resp. $s_d > s_b$) they are longer (shorter) for A.

The above argument shows that the footprint of a sweep from the standing genetic variation is identical to a "usual" sweep pattern if the selection coefficient changes its sign, but not its absolute value upon the environmental change. If we observe the sweep region at time G, the only difference to a sweep that has originated from a new mutation after time T is the somewhat older age of the sweep from the standing variation. For $s_d \neq s_b$, the change in the selection regime leads to differences in the expected footprint of alleles A and A'. Clearly, this difference is due to the cases where the coalescent of A (and A') extends into the old environment, i.e. where the sweep is "soft". For $s_d > s_b$, the expected coalescense in the ancestral environment is faster for A than for A', leading to stronger footprint of selection. However, since soft sweeps are very rare for $s_d > s_b$, this will hardly lead to a detectable difference in the average footprint.

Let us now concentrate on the case $s_b > s_d$, or $R_\alpha > 1$ where soft sweeps are frequent. In this case, the coalescense in the ancestral environment is slower and the selective signature for A reduced in depth and width relative to A' (due to the increaesed opportunity for mutation and recombination until the allele is fully coalesced). If the frequency x of the allele at time T is large, the sweep pattern of A will look more like a sweep of an advantageous allele with a selection coefficient of size $s_d < s_b$. We therefore also expect to find a larger difference between the footprints of soft sweeps and hard sweeps from a new mutation in this case. For a rough estimate of when this difference should be detectable, we compare the total fixation times of the allele A in the case of a soft sweep, $t_{\text{fix,soft}}(s_d, s_b)$, with the average duration of a sweep from a new mutation $t_{\text{fix}}(s_b)$ (cf. Eq. 1.13). For an optimal (that is minimal) time of observation $G \approx t_{\text{fix}}(s_b)$, we expect a clear difference in the selective signatures if the increase in coalescence time is of the same order of magnitude as the original coalescence time. Estimating the relative change in coalescence time by the change in fixation time, this means $t_\Delta = t_{\text{fix,soft}}(s_d, s_b) - t_{\text{fix}}(s_b) \gtrsim t_{\text{fix}}(s_b)$. We derive t_Δ from the frequency distribution of the allele at the time

T conditional on multiple fixation and results from diffusion theory on the expected age of an allele given its frequency; details are given in the Appendix. The results (not shown) predict visible changes in the sweep pattern for a minimum of R_α between 20 and 100.

1.4 Dicussion

The adaptive process is the genetic response of a population to external challenges. In nature, these challenges may be due to changes in climate or food resources, or arise with the advent of a new predator or parasite. They either affect the original habitat of the population, or are a consequence of the colonization of a new niche, or of human artificial selection. In this article, we are interested in the adaptive response of a previously well-adapted population to a sudden and permanent change. We concentrate on a single locus with two (classes of) alleles, one, a, ancestral, and the other, A, derived. Allele A is either neutral or deleterious under the original conditions, but selectively advantageous after the change in the selection regime at some time T. We compare two scenarios; A either already segregates in the population at time T and fixes from the standing genetic variation, or the population adapts from a new copy of the allele that only enters the population after the environmental shift.

Our results rely on two main assumptions. First, and most importantly, we assume that adaptation of the target allele does not interfere with positive or negative selection on other alleles, through either linkage or epistasis. This assumption is usually made in population genetic studies of selective sweeps. It is satisfied if the rate of selective substitutions is low and the time to fixation for each individual substitution is short, but is less plausible for weakly selected alleles with long average fixation times. In general, interference reduces fixation probabilities, with a stronger influence on weak substitutions (Barton 1995), although this does not translate into a large effect on the reduction of heterozygozity due to a selective sweep (Kim and Stephan 2003). In their study of fixation probabilities of alleles from the standing variation, Orr and Betancourt (2001) did not find a large effect of interference. This, however, may be a consequence of the neglect of new mutations and the restriction to a low initial frequency of the selected allele in their simulations. These assumptions make it unlikely that two or more beneficial alleles escape early stochastic loss and compete on their way to fixation. We therefore emphasize that our results are conditional on non-interference. Second, we assume that the variation at the locus under consideration is maintained in mutation-selection-drift balance prior to the environmental change. If selected alleles are maintained as a balanced polymorphism, or are not in equilibrium at all, this may clearly affect our conclusions.

Our results pertain to three main issues: The dependence of fixation probabilities on selection coefficients if alleles are taken from the standing genetic variation, the relative importance of the standing variation and new mutations as the origin of adaptive substitutions, and the expected impact of a selective sweep from the standing genetic variation on linked nucleotide variation. We will discuss them in turn.

Fixation probability from the standing variation

In a famous argument that helped to found the mirco-mutationist view of the adaptive process, Fisher (1930) showed that mutations with a small effect are much more likely to be beneficial than mutations with a large effect. Kimura (1983), however, pointed out a flaw in this argument: Even if a large majority of new beneficial mutations has a small effect, as Fisher argues, this may be offset by a much smaller fixation probability of weakly-selected alleles. An allele with (constant) heterozygote advantage hs_b that enters the population as a single new copy will escape stochastic loss and spread to fixation with probability $2hs_b$. One can think of stochastic loss as a sieve where small-effect alleles pass through the holes – and vanish from the population – much more often than alleles with a large selective advantage. A variant of this picture is known as *Haldane's sieve* and pertains to different levels of dominance: Substitutions are likely to be dominant since dominant alleles enjoy higher fixation rates.

This latter scenario is the subject of Orr and Betancourt (2001), who study Haldane's sieve if selected alleles are taken from the standing genetic variation. They conclude that the sieve is not active in this case. If the selected allele is deleterious under the original conditions (with heterozygote disadvantage $h's_d$), and if the level of dominance is maintained upon the environmental shift, $h = h'$, the net fixation probability is approximately independent of dominance. It is easy to understand why: The advantage of a higher fixation rate with larger h is compensated by the lower frequency of the initially deleterious allele in mutation-selection balance. Orr and Betancourt (2001) focus on a limited parameter range, where the selected allele is definitely deleterious under the original conditions and thus starts at a low frequency. In their figures, they also assume that the original deleterious effect is larger than the subsequent beneficial effect of the allele, meaning that the relative selective advantage $R_\alpha = 2h\alpha_b/(2h'\alpha_d + 1)$ is smaller than 1. Our study extends their analysis to arbitrary values of R_α. The simple analytical approximation for the probability of a substitution from the standing variation (Eq. 1.10 above, resp. Eq. 3 in ORR and BETANCOURT 2001), which uses the deterministic value for the initial frequency of A in mutation-selection balance, is no longer valid in the general case. Nevertheless, there is an equally simple expression, Eq. (1.8), that serves as an approximation for the entire parameter range.

Our results corroborate and extend the findings of ORR and BETANCOURT (2001). To the order of our approximation, the fixation probability from the standing genetic variation depends on selection only through R_α. If selection is strong in both environments, and $h' = h$, it is independent of dominance. More generally, if beneficial and deleterious effects of alleles in different environments were strictly proportional, the distribution of the effects of adaptations from the standing variation would coincide with the distribution of the effects of new beneficial mutations, as implicitly assumed in FISHER'S (1930) argument. The reason is the same as in the case of dominance: Advantages in the fixation probability due to a larger α_b are compensated by disadvantages due to a smaller initial frequency with higher α_d.

Remarkably, we find that the stochastic sieve is substantially weakened even if alleles with a larger selective advantage do not have a larger disadvantage to compensate for it. If alleles are originally neutral or under relatively weak deleterious selection, such that $R_\alpha > 1$, there is only a very weak logarithmic dependence of the fixation probability on all parameters for selection or dominance. The reason is the high initial frequency of the *successful* alleles in this case, which may be much higher than the average frequency of all segregating alleles. At these high frequencies, the fixation probability is only weakly dependent on the selection coefficient of the allele. There is, however, a sieve acting against alleles under disproportionally large past selection, $R_\alpha < 1$. If the selected physiological function (with fixed $h\alpha_b$) is met by several alleles with different $h'\alpha_d$, alleles with a relatively mild deleterious effect in the past, $h'\alpha_d < h\alpha_b$, will be preferred. Note that this should confer a certain level of resilience to the population if the environmental conditions change back.

Empirical estimates of R_α, the relative selection strength, are difficult to obtain and generally not available. There is no *a priori* reason to assume that s_b is either larger or smaller than s_d ($s_b < s_d$ was assumed by ORR and BETANCOURT 2001). To see this, note that the role of the alleles A and a and the selection coefficients s_b and s_d are exchanged if the environment changes back to the old conditions at some later time. This argument does not pertain to the average selection coefficient of *any* deleterious allele (which is plausibly larger than the average beneficial effect), but only to the selection coefficients of deleterious alleles that are beneficial in the new environment. Several factors can cause an upward or downward bias of R_α. R_α is downward biased if there is a bottleneck at the time of the environmental change. In this case, the effective population size that enters α_b is reduced relative to the original N_e that enters α_d. An upward bias in R_α could result from a change in dominance following the environmental shift. To see this, assume that alleles a and A serve different functions that are only (or mostly) used in the old and new environment, respectively. The physiological theory of dominance claims that the common observation of dominant wild-type alleles is a natural consequence of multi-enzyme biochemistry (e.g. Kacser and Burns 1981; Orr 1991; Keightley 1996). If this holds true, it is natural to expect that there is at least partial dominance of the respective advantageous (wild-type) allele, hence of a (A) in the old (new) environment, and thus $h > h'$. Finally, if R_α is measured among successful substitutions from the standing genetic variation, a further upward bias results from the stochastic sieve against alleles with large $h'\alpha_d$.

Relative importance of adaptations from the standing variation and from new mutations

In order to estimate the importance of the standing genetic variation as a reservoir for adaptations, we compare a polymorphic population, in mutation-selection-drift balance, with a monomorphic one. We can measure the additional adaptive potential of the polymorphic population in number of generations G_{sgv} a monomorphic population must wait for sufficiently many new mutations to arrive to match the fixation probability from the standing variation. G_{sgv} can be very large for mutations with small effect (of the order $1/hs_b$ generations). However, for a population of constant size it is always smaller than the average fixation time of the allele. This means that there is no clear separation of adaptive phases: by the time most alleles from the standing genetic variation with a given selective advantage $h\alpha_b$ have reached fixation, substitutions from new mutations (with the same $h\alpha_b$) will also be found. Only if the environmental change is followed by a strong reduction in population size, the reservoir of the standing variation is exploited well before new mutations start to play a role.

We have also determined the probability that the standing variation contributes to an adaptive substitution that is observed some time G after an environmental change. Clearly, this probability generally declines with G. For fixed G there are two distinct parameter regions where the standing variation is most important.

1. Adaptations from the standing variation are favored for alleles with small effect that are under relatively weak past selection, $R_\alpha \geq 1$. This is a direct consequence of the stochastic sieve that eliminates weak alleles in a new mutation scenario. The effect is especially pronounced if the environmental shift is followed by a bottleneck with incomplete recovery. The percentage of substitutions that use alleles from the standing variation is then almost independent of the mutation rate since Θ_u affects the fixation probabilities from standing and new variation in the same way.

2. The standing variation is also important for alleles with a large relative selective advantage ($R_\alpha \gg 1$) if the mutation rate Θ_u is also high. In this case, fixation probabilities are high under both scenarios, new mutations and standing genetic variation. Since the standing variation, other then new mutations is immediately available, it will usually contribute a major share to the substitution. Note that $R_\alpha \gg 1$ is plausible in particular for "important" adaptations with large effect, such as insecticide resistence alleles. Whether such an adaptation likely originated from the standing genetic variation then depends mainly on Θ_u .

Selective footprints of soft sweeps

For a classical sweep from a single new mutation, which we call a *hard sweep*, ancestral variation can only be preserved if there is recombination between the polymorphic locus and the selection target during the selective phase. In a 'core' region around the selection center all ancestral variation is erased. In contrast, with a *soft sweep*, multiple copies of the selected allele contribute to the substitution. Depending on the history of these copies, part of the ancestral variation may then be maintained and appear as haplotype structure in the population. There are two types of soft sweeps. For the first type, multiple copies that contribute to the substitution derive from independent mutations. For the second type, multiple copies that existed at the time of the environmental change contribute to the substitutions, but these copies are identical by descent.

Soft sweeps of the first type (independent origins) are frequent if the mutation rate on the population level is sufficiently high ($\Theta_u \gtrsim 0.1$), see Fig. 1.6. Their probability relative to a sweep from a single origin also increases with the selection strength $h\alpha_b$, i.e. altogether for alleles with high adaptive rates.

Suprisingly, soft sweeps of this type are not exclusive to adaptations from the standing genetic variation, but occur with the same probability for adaptations that originate only from new mutations, which have entered the population after the environmental change. Even if material from the standing variation is used, most soft sweeps with copies from independent origins also involve new mutations that. Since surviving copies represent independent ancestral haplotypes, we expect characteristic differences in the selective footprint relative to classic pattern of a hard sweep, where only a single ancestral haplotype survives in the core region close to the selection site. A discussion of the effect of soft sweeps on the summary statistics for nucleotide variation will be given elsewhere.

Soft sweeps of the second type (copies with a common origin prior to the environmental change) can only occur for adaptations from the standing genetic variation. They are frequent even for very low mutation rate $\Theta_u \to 0$ if the allele has a high relative selective advantage $R_\alpha \gtrsim 4$, see Fig. 1.5. The sweep pattern depends on the strength of deleterious selection that the allele has experienced in the old environment. For $R_\alpha > 1$, we expect a weaker footprint with a narrower sweep region than predicted for a hard sweep with same selective advantage $h\alpha_b$. We predict, however, that differences in the sweep patterns are only visible for a minimum R_α of $20 - 100$. For $\alpha_d = 0$, where the probability of multiple fixations and the resulting effect on the sweep pattern are strongest, this has been studied in a recent publication by Innan and Kim (2004). Using computer simulations, these authors indeed find much weaker selective footprints if the alleles are taken from the standing genetic variation. Since their minimum value of R_α is 1000, their results fit our predictions.

We can summarize our results on soft sweeps in three observations. First, evidence of a soft sweep does not result in an easy criterion to distinguish adaptive substitutions from the standing variation and recurrent new mutations. For a large parameter space we will not be able to detect any difference between these adaptive scenarios. This confirms the conclusion of Orr and Betancourt (2001), although partly for different reasons. For high $\Theta_u \gtrsim 0.1$, soft sweeps are frequent in both cases; for low Θ_u and $R_\alpha \lesssim 20$ they are either rare in both cases or do not lead to significant differences in the selective footprints. For a range of "interesting" substitutions, namely alleles with a large effect but a low mutation rate, however, the linked nucleotide pattern could be informative.

Second, soft sweeps are frequent in a limited but relevant parameter space. We expect soft sweeps with characteristic patterns on the selective footprints for high Θ_u, i.e. either if the population size is large, or if the allelic mutation rate is high, such as at mutational hotspots or if the adaptation corresponds to a loss-of-function mutation of the gene. We also expect soft sweeps for large adaptations with $h\alpha_b \gg h'\alpha_d$ (thus $R_\alpha \gg 1$) from the standing variation, even if the mutation rate is small. The effect of a soft sweep in this last case is a reduction in the width of the sweep region relative to a hard sweep. A possible candidate for a soft sweep of this type is the evolution of DDT resistance in non-African populations of *D. melanogaster*. In recent studies of nucleotide and microsatellite variability in the region around an *Accord* insertion that is associated with DDT resistance, Schlenke and Begun (2004) and Catania *et al.* (2004) found evidence for a selective sweep. The width of the sweep region, however, was much narrower in *D. melanogaster* than expected under putatively very strong selection (Catania *et al.* 2004) and as observed, for the "same" adaptation (with a *Doc* insertion) in *D. simulans* (Schlenke and Begun 2004).

Third, while hard sweeps from single mutations produce the strongest footprint for strongly selected alleles with short fixation times, the possibility of fixation of multiple alleles leads to an opposite trend: soft sweeps with a weaker footprints are more frequent for high α_b. Since the increase is only logarithmic, this trend is not very strong. Nevertheless, it could be visible for nucleotides that are tightly linked to the selected allele in regions of low recombination or in sufficiently small windows around the selection target. A genome-wide study of the small scale reduction of heterozygosity in narrow windows of 200 base pairs around replacement or silent fixations has recently been performed for *D. simulans* by Kern *et al.* (2002). We note that their counterintuitive finding of a sweep signature for preferred codon substitutions, but not for replacement substitutions, matches our prediction of a stronger sweep signal for weakly selected alleles close to the selection center. However, a quantitative analysis of soft sweeps that also accounts for other factors like population substructure is needed before any conclusions can be drawn.

1.5 Acknowledgements

We thank Sylvain Mousset and Wolfgang Stephan for fruitful discussions and John Parsch for helpful comments on the manuscript. The careful comments by Sally Otto and an anonymous reviewer led to many clarifying changes. We also thank Pieter van Beek for help with the computer simulations. This work was supported by an Emmy Noether grant by the Deutsche Forschungsgemeinschaft (DFG) to J.H.

1.6 Appendix

Fixation probability for mutation segregating at neutrality

In this appendix, we calculate the average fixation probability of an allele that is derived from a single mutation and segregates in the population under neutrality at the time T of the environmental change. The probability that there are exactly k copies at time T is distributed as $\rho(k) = a_N k^{-1}$, where $a_N = \sum_{k=1}^{2N_e-1}(1/k)$. Assuming a selection coefficient s_b for $t > T$ and no dominance ($h = 0.5$), the average fixation probability is given by

$$\begin{aligned}\Pi_{\text{seg}}(N_e, s_b) &= \frac{1}{a_N}\sum_{k=1}^{2N_e-1}\frac{1-\exp(-ks_b)}{k(1-\exp(-2N_e s_b))}\\ &= \frac{1}{1-\exp(-2N_e s)}\left(1-\frac{1}{a_N}\sum_{k=1}^{2N_e-1}\frac{\exp(-ks_b)}{k}\right). \end{aligned} \tag{1.21}$$

We derive the sum in Eq. (1.21) as

$$\begin{aligned}&\sum_{k=1}^{2N_e-1}\frac{e^{-ks_b}}{k} = \int_{s_b}^{\infty} d\tilde{s}_b \sum_{k=1}^{2N_e-1} e^{-k\tilde{s}_b} = \int_{s_b}^{\infty} d\tilde{s}_b \left[\frac{e^{-\tilde{s}_b} - e^{-2N_e\tilde{s}_b}}{1-e^{\tilde{-}s_b}}\right]\\ &= \int_{s_b}^{\infty} d\tilde{s}_b \left[\frac{1}{e^{\tilde{s}_b}-1}\right] + \int_{-s_b}^{-\infty} d\tilde{s}_b \left[\frac{e^{2N_e\tilde{s}_b}}{1-e^{\tilde{s}_b}}\right] = -\ln(1-e^{-s_b}) + \frac{{}_2F_1(1, 2N_e, 2N_e+1, e^{-s_b})}{2N_e e^{2N_e s_b}}\end{aligned} \tag{1.22}$$

where ${}_2F_1$ denotes the hypergeometric function. For $N_e s_b \gg 1$, this second term can be neglected and we obtain

$$\Pi_{\text{seg}}(N_e, s_b) \approx 1 + \frac{1}{a_N}\ln(1-e^{-s_b})\,. \tag{1.23}$$

In the limit of small s_b and large N_e this reduces to

$$\Pi_{\text{seg}}(N_e, s_b) \approx 1 + \frac{\ln(s_b)}{\ln(2N_e)+\gamma} \tag{1.24}$$

where $\gamma = 0.577\ldots$ is Euler's constant. For weak recessivity, this result holds if we replace s_b by $2hs_b$.

Fixation probability for allele in mutation-selection-drift balance

In order to calculate the frequency distribution of a derived allele, we start out with the Kolmogorov forward equation that describes the Wright-Fisher model in the diffusion limit (Ewens 2004),

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(a(x)f(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(b(x)f(x,t)\right) \tag{1.25}$$

where

$$a(x) = \frac{1}{2}\Big(-\alpha_d x(1-x)\big(2x+2h'(1-2x)\big)-\Theta_v x+\Theta_u(1-x)\Big) \quad \text{and} \quad b(x) = x(1-x) \tag{1.26}$$

are the drift and diffusion terms. Since the diffusion process is ergodic, the probability that the frequency of an allele falls into a certain interval $[x_1, x_2]$ is proportional to the average time T that an allele that starts out as a single copy spends in this frequency range before it is either lost or fixed. The frequency distribution therefore directly follows from the well-known transient behavior of the process, e.g. Ewens (2004), chapter 4. From equations (4.23) and (4.16) in Ewens (2004), we obtain

$$\rho(x) = C\,\frac{\exp[-\alpha_d(2h'x+(1-2h')x^2)]}{x^{1-\Theta_u}(1-x)^{1-\Theta_v}}\int_x^1 \frac{\exp[\alpha_d(2h'y+(1-2h')y^2)]}{y^{\Theta_u}(1-y)^{\Theta_v}}\,dy \tag{1.27}$$

where C is a normalization constant. Note that this expression deviates from Wright's stationary distribution of an allele in mutation-selection-drift balance since we condition on the case that A is derived.

Simple approximate relations for Eq. (1.27) are readily obtained in various limiting cases. First, direct numerical integration shows that back mutations can safely be ignored even in the neutral case $\alpha_d = 0$ because most alleles segregate at low frequencies (this is a consequence of conditioning on derived alleles). In the neutral case, this approximation directly leads to Eq. (1.5). If there is deleterious selection, we need to distinguish cases of weak and strong recessivity of the allele A. We will mostly concentrate on the case where deleterious selection on the heterozygote is sufficiently strong, $2h'\alpha_d \gg (1-2h')/2h'$ (i.e. weak recessivity). Under these conditions, we can ignore the quadratic terms in the exponentials and express $\rho(x)$ in terms of incomplete

Gamma functions,

$$\rho(x) = \\ C' \exp(-2h'\alpha_d x) x^{\Theta_u - 1} \frac{(-2h'\alpha)^{\Theta_u - 1}(\Gamma(1-\Theta_u, -2h'\alpha_d x) - \Gamma(1-\Theta_u, -2h'\alpha_d))}{1-x}, \tag{1.28}$$

with normalization constant C'. For definitely deleterious A ($2h'\alpha_d \geq 10$ is sufficient), the integrand in Eq. (1.27) is concentrated near $y = 1$. We can then expand y^{Θ_u} in the denominator to leading order around $y = 1$ (i.e. $y^{\Theta_u} \approx 1$) and obtain $\rho(x)$ in terms of simple functions, which leads to Eq. (1.6).

In order to obtain an analytical expression for the probability of fixation P_{sgv} or multiple fixation P_{mult}, we need to approximate $\rho(x)$ further. If the allele A is neutral prior to the environmental change, and $\Theta_u \ll 1$, $\rho(x)$ in Eq. (1.5) is approximately $\rho(x) \approx \Theta_u x^{\Theta_u - 1}$. Using this in Eq. (1.4)

$$\begin{aligned} P_{\text{sgv}}(\Theta_u, h\alpha_b) &\approx \Theta_u \int_0^1 \Big[x^{\Theta_u - 1}(1 - \exp[-2h\alpha_b x])\Big] dx \\ &\approx 1 - \frac{\Gamma(\Theta_u + 1)}{(2h\alpha_b + 1)^{\Theta_u}} \approx 1 - (2h\alpha_b + 1)^{-\Theta_u}, \end{aligned} \tag{1.29}$$

where we extend the integral over $\exp(-2h\alpha_b x)$ to ∞ after increasing $2h\alpha_b$ by 1 in order to avoid a singularity near $\alpha_b = 0$. We also use $\Gamma(\Theta_u + 1) \approx 1$ for $0 \leq \Theta_u \leq 1$.

For the deleterious case ($2h'\alpha_d \gg 1$), note that the allele frequency distribution is significantly larger than zero only for $x \leq 1/2h'\alpha_d$. Expanding around $x = 0$ we can approximate $\rho(x)$ in Eq. (1.6) as $\rho(x) \approx C'' x^{\Theta_u - 1} \exp(-2h'\alpha_d x)$ and obtain

$$\begin{aligned} P_{\text{sgv}}(\Theta_u, h'\alpha_d, h\alpha_b) &\approx 1 - \int_0^1 \frac{x^{\Theta_u - 1}}{\exp[(2h'\alpha_d + 2h\alpha_b)x]} dx \Big/ \int_0^1 \frac{x^{\Theta_u - 1}}{\exp[2h'\alpha_d x]} dx \\ &\approx 1 - \Big(\frac{1 + 2h\alpha_b + 2h'\alpha_d}{1 + 2h'\alpha_d}\Big)^{-\Theta_u} \end{aligned} \tag{1.30}$$

which gives Eq. (1.8). In Eq. (1.30), we have again extended integral limits after adding 1 to $2h'\alpha_d$, resp. $2h\alpha_b + 2h'\alpha_d$. We now see that the approximation for $2h'\alpha_d \gg 1$ reproduces the approximation for $\alpha_d = 0$ in the limit $\alpha_d \to 0$. We can therefore use it in the entire parameter range. For $\Theta_u < 1$, the probability that the allele A is not contained in the standing variation at time T can be approximated by the integral over $\rho(x)$ from 0 to $1/2N_e$ (confirmed by simulations, see also EWENS 2004, Chap. 5.7). With the above approximations for $\rho(x)$ this results in Eq. (1.7). Finally, also P_{mult} is obtained by an analogous calculation.

If the allele A is completely recessive prior to the environmental change, $h' = 0$, we again obtain an expression in incomplete Gamma functions for $\rho(x)$ similar to Eq. (1.28). For large α_d, this reduces to

$$\rho(x) \approx \frac{\alpha_d^{\Theta_u/2} \exp[-\alpha_d x^2]}{\Gamma(\Theta_u/2) x^{1-\Theta_u}} . \tag{1.31}$$

Using this expression in Eq. (1.4), we see that the term $\exp[-\alpha_d x^2]$ can be ignored as long as $2h\alpha_b > \sqrt{\alpha_d}$ since the integral is cut off by $\exp[-2h\alpha_b x]$. For $2h\alpha_b < \sqrt{\alpha_d}$, both selection coefficients are important. We can obtain a simple, yet compared to simulation data (not shown) reasonable, analytic approximation that captures this crossover behavior by formally replacing $2h'\alpha_d + 1$ by $\sqrt{\alpha_d} + 1$ in Eqns. (1.8), (1.7), and (1.18) if $h' = 0$.

The average frequency of the allele A at time T conditioned on later fixation, $\bar{x}_{\text{fix}}$, is calculated from the distribution $Pr(x|\text{fix}) = C\rho(x)\Pi_x(h\alpha_b)$. With the above approximations for $\rho(x)$, we obtain

$$\bar{x}_{\text{fix}} \approx \frac{\Theta_u}{2h'\alpha_d + 1} \frac{1 - (1 + R_\alpha)^{-(\Theta_u+1)}}{1 - (1 + R_\alpha)^{-\Theta_u}} \tag{1.32}$$

For $\Theta_u \to 0$, this gives

$$\bar{x}_{\text{fix}} \approx \frac{R_\alpha}{(2h'\alpha_d + 1)(1 + R_\alpha) \ln[1 + R_\alpha]} . \tag{1.33}$$

Finally, if also $\alpha_d = 0$, and $2h\alpha_b \gg 1$

$$\bar{x}_{\text{fix}} \approx \frac{2h\alpha_b}{(2h\alpha_b + 1) \ln(2h\alpha_b + 1)} \approx \frac{1}{\ln(2h\alpha_b)} . \tag{1.34}$$

For the calculation of the average increase in the age of a selected allele for a soft sweep with a weak trade-off, we use the frequency distribution of the allele at time T conditioned on *multiple* fixation, $\text{Pr}(x|\text{mfix}) \approx C\rho(x)(\Pi_x(h\alpha_b))^2$. [We use the Poisson approximation Eq. (1.16) and $2h\alpha_b x \approx 1 - \exp(-2h\alpha_b x)$ for small x where $\rho(x)$ is large.] We only consider the case $\Theta_u \to 0$ and $h = h' = 0.5$. For a given allele frequency x at time T, we determine the average age $t_a(\alpha_d, x)$ of the allele using Eq. (5.113) in EWENS (2004) (see also

Kimura and Ohta 1969),

$$\begin{aligned} t_a(\alpha_d, x) =& \frac{2}{\alpha_d(e^{\alpha_d}-1)} \int_0^x \frac{(e^{\alpha_d y}-1)(e^{\alpha_d(1-y)}-1)}{y(1-y)}\, dy + \\ & \frac{2(1-e^{-\alpha_d x})}{\alpha_d(1-e^{-\alpha_d})(e^{\alpha_d(1-x)})-1} \int_0^1 \frac{e^{-\alpha_d(1-y)}(e^{\alpha_d(1-y)}-1)^2}{y(1-y)}\, dy\,. \end{aligned} \tag{1.35}$$

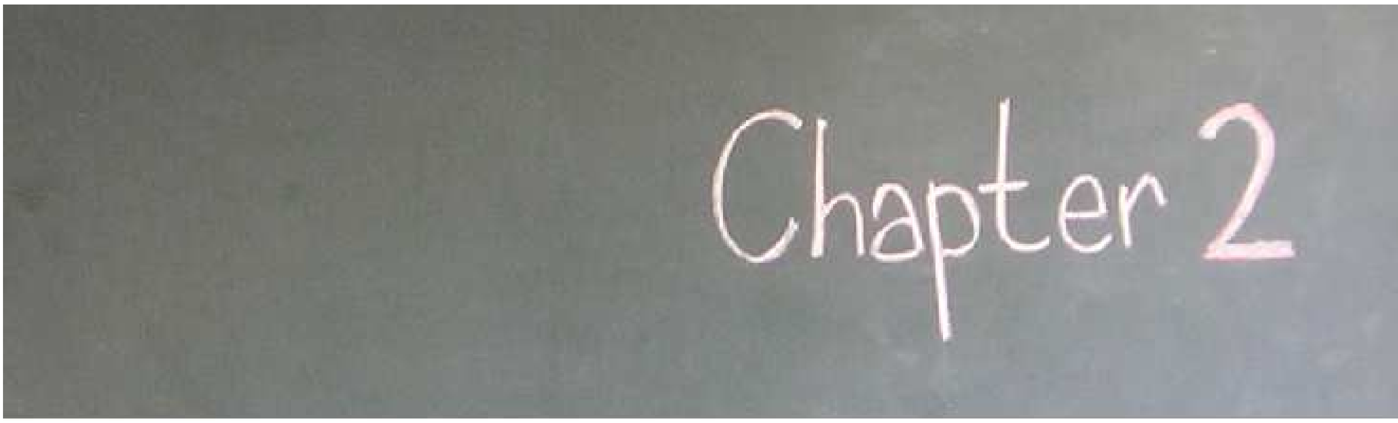
The increase in the age of the allele due to the change of the selection regime then is obtained by numerical integration as $t_\Delta = \int (t_a(\alpha_d, x) - t_a(\alpha_b, x))\Pr(x|\text{mfix})dx$. Choosing $x = 1$, Eq. (1.35) allows for a simple approximation for the fixation time of a new allele with selective advantage α_b. We derive

$$\begin{aligned} t_{\text{fix}}(\alpha_b) &= \frac{2}{\alpha_b(\exp[\alpha_b]-1)} \int_0^1 \frac{(\exp[\alpha_b y]-1)(\exp[\alpha_b(1-y)]-1)}{y(1-y)}\, dy \\ &= \frac{4}{\alpha_b(\exp[\alpha_b]-1)} \int_0^1 \frac{(\exp[\alpha_b y]-1)(\exp[\alpha_b(1-y)]-1)}{y}\, dy \end{aligned} \tag{1.36}$$

For $\alpha_b \geq 3$, this may be approximated as

$$\begin{aligned} t_{\text{fix}}(\alpha_b) &\approx \frac{4}{\alpha_b} \int_0^1 \frac{1-\exp[-\alpha_b y]-\exp[\alpha_b(y-1)]+\exp[-\alpha_b]}{y}\, dy \\ &\approx \frac{4}{\alpha_b}\big(\ln[\alpha_b] + \gamma - \alpha_b^{-1}\big) \end{aligned} \tag{1.37}$$

where $\gamma \approx 0.577$ is Euler's Gamma. The error term is of order α_b^{-3}. To the best of our knowledge, this simple result has not yet been used in the literature. Simulation results of our own (not included) and in Kimura and Ohta (1969) show that the estimate is very accurate. For $h \neq 0.5$, we can replace α_b by $2h\alpha_b$ in Eq. (1.37). The approximation then holds as a lower bound for t_{fix}, since the fixation time increases if h deviates from 0.5 in either direction.



Chapter 2

Soft Sweeps II – Molecular population genetics of adaptation from recurrent mutation or migration

Pleuni S. Pennings and Joachim Hermisson

Published in Molecular Biology and Evolution (2006) 23:1076-1084

In the classical model of molecular adaptation, a favored allele derives from a single mutational origin. This ignores that beneficial alleles can enter a population recurrently, either by mutation or migration, during the selective phase. In this case, descendents of several of these independent origins may contribute to the fixation. As a consequence, all ancestral haplotypes that are linked to any of these copies will be retained in the population, affecting the pattern of a selective sweep on linked neutral variation. In this study, we use analytical calculations based on coalescent theory and computer simulations to analyze molecular adaptation from recurrent mutation or migration. Under the assumption of complete linkage, we derive a robust analytical approximation for the number of ancestral haplotypes and their distribution in a sample from the population. We find that so-called "soft sweeps", where multiple ancestral haplotypes appear in a sample, are likely for biologically realistic values of mutation or migration rates.

2.1 Introduction

When a beneficial allele rises to fixation in a population, it erases genetic variation in a stretch of DNA that is linked to it. This phenomenon is called "genetic hitch-hiking" or a "selective sweep", and was first described by Maynard Smith and Haigh (1974). In the classical scenario for such an adaptive substitution, the beneficial allele arises in the population as a single new mutation and then increases to fixation under a constant selection pressure. Under this scenario, genetic variation in parts of the genome that are tightly linked to the selected site is lost and will only be recovered by new mutation. Ancestral variation, i.e. genetic variation that has been present in the population prior to the selective phase, is only maintained if recombination during the selective phase breaks the association between the study locus and the selected site. The resulting pattern of a selective sweep, a valley of reduced variation around the target of selection, has been described in some detail and is well understood (e.g. Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 1995; Durett and Schweinsberg 2004; Etheridge *et al.* 2005).

There is, however, a second scenario how ancestral variation can be maintained in the face of positive selection. Namely, if an adaptive substitution involves multiple copies of the same beneficial allele. This can happen in the following two ways. If adaptation occurs from the standing genetic variation, a large number of copies of the beneficial allele may be initially present. Fixation of the allele may then involve descendents of more than one of these copies. Alternatively, a beneficial allele can enter the population recurrently by mutation or migration during the selective phase. Again, descendents of several of these independent origins may contribute to the fixation of the allele. In both cases, ancestral haplotypes that are linked to any of these copies will be retained in the population. Clearly, this would affect the pattern of a selective sweep on linked DNA variation. We call selective sweeps that involve (descendents of) more than one copy of the selected allele, "soft sweeps". They are distinguished from the classical "hard sweeps" where ancestral variation is maintained only through recombination.

Selective sweeps from the standing genetic variation have been described in three recent publications. Hermisson and Pennings (2005) derive the probability for a soft sweep for adaptation from the standing genetic variation. Innan and Kim (2004) and Przeworski *et al.* (2005) describe the effect of an adaptive substitution from the standing variation on summary statistics for DNA variation, assuming that the allele had been neutral prior to the onset of positive selection. There is then the chance that ancestral variation – due to mutation during this first time period – is retained in the population even without recombination. However, as long as there is only a single origin of the beneficial allele (as assumed by Innan and Kim and Przeworski *et al.*), the effect is necessarily limited. Other than in the case of recombination, the surviving ancestral haplotypes are not independent, but identical by descent.

In this study, we focus on selective sweeps from a beneficial allele that enters the population recurrently by mutation or migration. We derive the probability for a soft sweep, given the mutation/migration rate and the selection coefficient of the beneficial allele. More generally, we determine the expected number of independent ancestral haplotypes and their frequency distribution in a sample from a locus that is tightly linked to the selected site. Our results show that soft sweeps are likely under biologically realistic conditions.

2.2 Model and Methods

Model and definitions

We study a single locus under selection in a haploid population of effective size N_e. For most of this study, only two alleles (or classes of alleles) at this locus are considered, an ancestral allele b and a new beneficial variant B with fitness advantage s. In general, we will allow s to depend on time and/or on the frequency of the beneficial allele. The B allele enters the population through either recurrent mutation at rate u or migration at rate m (where m is the per generation probability for an individual to be replaced by a migrant). We consider mutation and migration separately. Back mutation or migration are ignored. We define population level parameters for selection, mutation, and migration as $\alpha = 2N_es$, $\Theta = 2N_eu$, and $M = 2N_em$. Every generation consists of reproduction (including fertility selection), followed by mutation or migration, see figure 2.1.

Assume that the population is originally monomorphic for the ancestral allele b. After successful substitution, all individuals carry the B allele. Because mutation or migration are recurrent, this substitution may involve several copies of the B allele with independent origins in the sense that they do not trace back to a single ancestor in the study population. Independent copies are linked to independent genetic backgrounds that are randomly drawn either from the study population prior to the substitution or from the source population of migrants. We call these independent genetic backgrounds at the selected locus "independent ancestral haplotypes", or "ancestral haplotypes" for short. Note that with this definition differences due to new mutations or recombination events (i.e. events after the first beneficial mutation or migration event) are not considered. Note also that "independent" does not necessarily mean "different", since it includes the possibility that the same haplotype is drawn multiple times.

Suppose that we take a sample from the selected locus or from a tightly linked fragment (so that no recombination has taken place between this fragment and the selected locus) some time after fixation of the B allele. If there is more than one ancestral haplotype in the sample, we call this a soft selective sweep from recurrent mutation or migration. The opposite case (only one ancestral haplotype) is called a hard sweep. Note that soft and hard sweeps can be defined either with respect to a sample or with respect to the population. A soft sweep in a population means that there are several ancestral halpotypes at the selected locus in the population. In this paper we usually consider samples.

Simulations

We checked all analytical results by forward-in-time computer simulations. For this a Wright-Fisher model with $N_e = 500,000$ haploid individuals is simulated. Each run starts with a population that is monomorphic for the ancestral b allele. Reproduction is simulated by fitness-weighted multinomial sampling. After reproduction, every b-individual has probability u to mutate to B. In the migration model, every individual, independent of its genotype, is replaced by a migrand with probability m. Descendents of mutants and migrants are followed separately; at the observation time their frequencies are determined in a population sample. Data points are averages over $100,000$ runs ($10,000$ runs for $\alpha = 100$). The code is available on request.

2.3 Results

This section is organized in four parts. The first two consider recurrent mutation. We start with a detailed derivation for a sample of two, which is the simplest case. We then use intuitive arguments to motivate our main result, which is the frequency distribution of ancestral haplotypes for a sample of size n. All formal derivations for this general case are given in the supporting online material. In the third part we show how these results apply to the recurrent migration case. Finally, we briefly discuss several generalizations of the model.

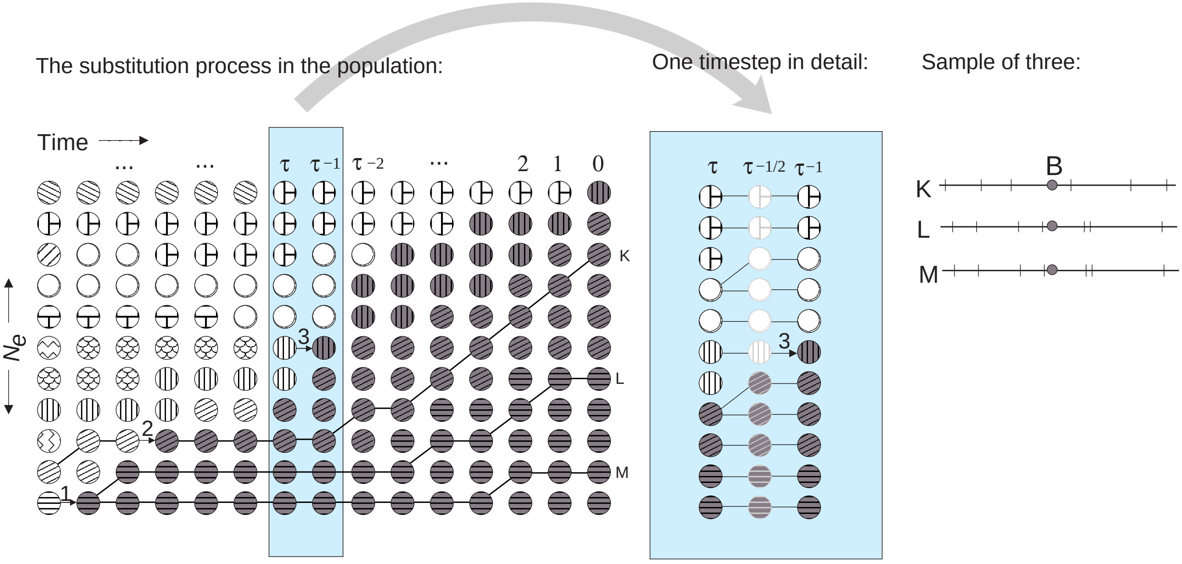


Figure 2.1: Soft selective sweep from recurrent mutation in a schematic Wright-Fisher model. Circles represent individuals, the different patterns indicate independent ancestral haplotypes. The beneficial allele B (dark grey individuals) substitutes the ancestral b allele (white). The B allele arises three times by independent mutation; individuals then change their color from white to grey, but keep their haplotype pattern. The "zoom" into a single time step shows how reproduction and mutation are separated. Directly after fixation (time 0), we take a sample of size three (K,L,M) that contains descendents from the first (L,M) and the second (K) mutational origin of B. The right panel shows DNA fragments of the sampled individuals. The vertical ticks represent neutral polymorphisms. Individuals L and M share a recent ancestor and are identical in this region of the genome. Individual K carries a different ancestral haplotype.

Soft sweeps from recurrent mutation in a sample of size two

Consider a sample of size two that is taken from a population at some time t_{obs}, measured from the time of fixation of the beneficial B allele. Initially, we will assume that sampling occurs directly at fixation, i.e. $t_{\text{obs}} = 0$. We want to derive the probability $P_{\text{soft},2}$ that the two copies of the B allele in the sample are not identical by descent, i.e. the probability of a soft selective sweep. We use a coalescent framework and define τ as the time in the past before the sample was taken, i.e. $\tau = 0$ for $t = t_{\text{obs}}$ and if $\tau_2 > \tau_1$, then τ_2 is further back in the past than τ_1.

Let x_τ be the fraction of the population that carries the beneficial allele B at time τ. We follow the fate of the two lineages backward in time until they either coalesce or one of the two mutates. $P_{\text{soft},2}$ is the probability that mutation happens before coalescence; we denote the alternative possibility that the lines coalesce before one of the two mutates as $P_{\text{hard},2} = 1 - P_{\text{soft},2}$.

Let $P_{\text{coal},2}(\tau)$ be the coalescence probability in generation τ and $P_{\text{mut},2}(\tau)$ the probability that one of the two lineages has mutated and had a b ancestor in generation τ. We can then express $P_{\text{hard},2}$ as

$$P_{\text{hard},2} = \left\langle \sum_{\tau=1}^{\infty} \Big(P_{\text{coal},2}(\tau) \cdot \prod_{i=1}^{\tau-1} \big(1 - P_{\text{mut},2}(i) - P_{\text{coal},2}(i)\big) \Big) \right\rangle_x \tag{2.1}$$

where the empty product, $\prod_{i=1}^{0}$, is defined to be 1. The product is the probability that neither mutation nor coalescence has happened until generation τ. $\langle \ldots \rangle_x$ denotes the expectation over the stochastic path $\{x_\tau\}_\tau$ of the frequency x_τ of B.

To calculate $P_{\text{mut},2}$, it is convenient to separate reproduction (and therefore coalescence) from mutation by introducing an artificial intermediate generation after reproduction but before mutation: Using the backward-in-time notation, individuals of generation τ reproduce to form generation $\tau - \frac{1}{2}$ and the individuals in this intermediate generation can mutate or not to form $\tau - 1$ (see figure 2.1). Ignoring back-mutation from B to b the number of B alleles in the $(\tau - 1)$th generation is given by

$$x_{\tau-1} = (1 - x_{\tau-\frac{1}{2}}) \cdot u + x_{\tau-\frac{1}{2}}. \tag{2.2}$$

in which the first term on the right-hand side is the new mutants and the second term is the B's that were already there. For a single B lineage, the probability that it is a mutant is

$$P_{mut,1}(\tau) = \frac{\big(1 - x_{\tau-\frac{1}{2}}\big)u}{\big(1 - x_{\tau-\frac{1}{2}}\big)u + x_{\tau-\frac{1}{2}}} = \frac{\big(1 - x_{\tau-1}\big)u}{\big(1 - u\big)x_{\tau-1}} \tag{2.3}$$

where $x_{\tau-\frac{1}{2}} = \frac{x_{\tau-1}-u}{1-u}$ (from equation 2.2). Thus, the probability for (at least) one mutation in a sample of two is $P_{\text{mut},2} = 2P_{\text{mut},1} - P_{\text{mut},1}^2$. If no mutation has happened, coalescence happens with rate $1/(N_e x_\tau)$. The exact coalescence probability in generation τ therefore is

$$P_{\text{coal},2}(\tau) = \frac{1 - P_{\text{mut},2}(\tau)}{N_e x_\tau}\,. \tag{2.4}$$

In a sufficiently large population, and for small values of u, we can safely ignore the occurrence of several events in a single generation. Formally, this is done by ignoring terms of order of $u/(N_e x)$ and u^2. If we can also ignore terms of order $s/(N_e x)$, we can further set $x_{\tau-1} \approx x_\tau$. We then obtain

$$P_{\mathrm{mut},2}(\tau) \approx \frac{\Theta(1-x_\tau)}{N_e x_\tau} \quad ; \quad P_{\mathrm{coal},2}(\tau) \approx \frac{1}{N_e x_\tau} \tag{2.5}$$

for the probability of mutation and coalescence. Using equation 2.5 in equation 2.1, $P_{\mathrm{hard},2}$ can now be expressed as

$$\begin{aligned} P_{\mathrm{hard},2} &= \left\langle \sum_{\tau=1}^{\infty} \left\{ \frac{1}{N_e x_\tau} \prod_{i=1}^{\tau-1} \Big(1 - \frac{1+\Theta(1-x_i)}{N_e x_i}\Big) \right\} \right\rangle_x \\ &= \frac{1}{1+\Theta} \left[\left\langle \sum_{\tau=1}^{\infty} \left\{ \frac{1+\Theta(1-x_\tau)}{N_e x_\tau} \prod_{i=1}^{\tau-1} \Big(1 - \frac{1+\Theta(1-x_i)}{N_e x_i}\Big) \right\} \right\rangle_x + \right. \\ &\qquad \left. \left\langle \sum_{\tau=1}^{\infty} \left\{ \frac{\Theta x_\tau}{N_e x_\tau} \prod_{i=1}^{\tau-1} \Big(1 - \frac{1+\Theta(1-x_i)}{N_e x_i}\Big) \right\} \right\rangle_x \right] \\ &= \frac{1}{1+\Theta} \left[1 + \frac{\Theta}{N_e} \left\langle \sum_{\tau=1}^{\infty} \prod_{i=1}^{\tau-1} \Big(1 - \frac{1+\Theta(1-x_i)}{N_e x_i}\Big) \right\rangle_x \right] \end{aligned} \tag{2.6}$$

where the sum in the second line is the probability that the two lineages eventually either coalesce or mutate, which is 1 for every realization of the path $\{x_\tau\}_\tau$. The expectation in the last line of equation 2.6 has a simple interpretation. It is the average time, T_1, in generations until either coalescence or mutation happens. T_1 certainly lies between 0 and T_{fix}, the average fixation time for the beneficial allele in the population. This gives an upper and lower bound for $P_{\mathrm{hard},2}$ as

$$\frac{1}{1+\Theta} \le P_{\mathrm{hard},2} \le \frac{1}{1+\Theta}\Big(1 + \frac{\Theta\, T_{\mathrm{fix}}}{N_e}\Big)\,. \tag{2.7}$$

Equivalently,

$$\frac{\Theta}{1+\Theta} \ge P_{\mathrm{soft},2} \ge \frac{\Theta}{1+\Theta}\Big(1 - \frac{T_{\mathrm{fix}}}{N_e}\Big)\,. \tag{2.8}$$

This result has several important implications. First, none of the details of the stochastic process that underlies the path $\{x_\tau\}_\tau$ enters into the estimate for $P_{\mathrm{hard},2}$ or $P_{\mathrm{soft},2}$. In fact, the value of T_{fix} in one of the bounds is the only quantity that depends on this process – and thus on the selection coefficient. Second, we see that the estimate gets very precise (upper and lower bounds converge) if $T_{\text{fix}}/N_e \ll 1$. This is easily fulfilled for strong selection. In this case, $P_{\text{hard},2}$ and $P_{\text{soft},2}$ depend only on Θ, but are entirely independent of all selection parameters.

Finally, one should note that the derivation does not depend on the assumption that the sample is taken directly after fixation. Assume, instead, that the population is sampled some time t_{obs} after fixation. In that case T_{fix} in equations 2.7 and 2.8 has to be replaced by the expected age of the oldest B allele that is found in the population at the time of observation. The approximation will be good as long as $(t_{\text{obs}} + T_{\text{fix}})/N_e \ll 1$. If the sample is taken before full fixation, the above estimates 2.7 and 2.8 hold if we condition on a sample that is monomorphic for B.

To assess the quality of the bounds for $P_{\text{soft},2}$ in equation 2.8 we need an estimate of T_{fix}. For a single copy of a beneficial allele that rises to fixation under a constant selection pressure $\alpha = 2N_e s$, a precise estimate of the fixation time is $T_{\text{fix}}/N_e \approx 4 \log \alpha/\alpha$ (Hermisson and Pennings 2005). For $\Theta \ll \alpha$, the same approximation holds also for fixation under recurrent mutation. With this estimate for T_{fix} both bounds deviate by less than $< 5\%$ for $\alpha > 500$. Figure 2.2 confirms that simulation data falls between the predicted bounds. Only for extremely strong selection ($s \approx 1$), some deviations appear (data not shown). The reason is that the approximation $x_{\tau-1} \approx x_\tau$ that we have used in the derivation is no longer accurate in this case.

Soft sweeps from recurrent mutation in larger samples

Consider now a sample of size n taken from the population at some time t_{obs}. If sampling occurs before fixation, we condition on samples that are monomorphic for the B allele. We are interested in the number and the frequency distribution of ancestral haplotypes in the sample.

If there are k ancestors of the sample that are associated with a B allele at time τ, the probability for mutation and coalescence at this time is approximately

$$P_{\text{mut},k} \approx \frac{k\Theta(1-x_\tau)}{2N_e x_\tau} \quad ; \quad P_{\text{coal},k} \approx \frac{k(k-1)}{2N_e x_\tau} \tag{2.9}$$

where x_τ is the frequency of the beneficial allele. Using these relations, we can extend the above approach and calculate upper and lower bounds for the probability of a soft selective sweep. These derivations are given in the online material. Below, we focus on just one of the bounds where a more intuitive derivation is possible.

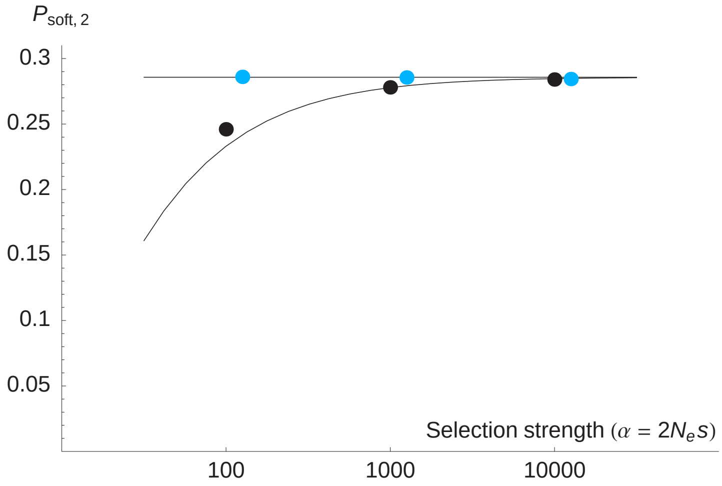


Figure 2.2: The probability of a soft selective sweep in a sample of size two, taken directly after fixation. The horizontal line represents the first order approximation (upper bound, equation 2.8), the curved line the second order approximation (lower bound, equation 2.8). Dots are simulation results; black dots are for mutation ($\Theta = 0.4$), the grey dots are for migration ($M = 0.4$).

We need two steps for our argument. First, note that the leading order approximation for a sample of size two (i.e. the lower bound for in eq. 2.7 and the upper bound in eq. 2.8) is equivalent to an approximation of the mutation probability $P_{\mathrm{mut},2}$. In fact, equation 2.6 reduces to $1/(1+\Theta)$ if we ignore the factor $(1-x_\tau)$ in the numerator of $P_{\mathrm{mut},2}$ in equation 2.5. We can apply the same approximation to $P_{\mathrm{mut},k}$ in eq. 2.9 and justify this step as follows: The denominator of $P_{\mathrm{mut},k}$ guarantees that mutation is only likely if x_τ is small. In this case, however, $(1-x_\tau) \approx 1$.

Secondly, without the $(1-x_\tau)$ term, we see that the coalescence and mutation rates in eq. 2.9 are both proportional to $(1/x_\tau)$. If we are only interested in the order of events in the genealogy of the sample (and not in the exact times at which coalescence and mutation happen) only the relative rates matter and we can ignore the x_τ dependence altogether (see the online material for a formal derivation). The result is that the problem is now equivalent to a standard neutral coalescent in a population of constant size where lines are stopped at mutations (also called “coalescent with killings” Durett 2002). This problem is long known and can be exactly solved (e.g. Ewens 2004, p. 335ff). In particular, the expected number of haplotypes and their frequency distribution are given by the Ewens sampling formula: Given the mutation rate

Θ for the B allele, the probability to find k haplotypes, occurring $n_1, \ldots, n_k$ times in a sample of size $n = \sum_i n_i$ is

$$\Pr(n_1 \ldots n_k | n, \Theta) = \frac{n!}{k! n_1 \cdots n_k} \frac{\Theta^k}{\Theta(\Theta+1)\cdots(\Theta+n-1)} \, . \tag{2.10}$$

Using this result for $k = 1$ and $n_1 = n$, we obtain an upper bound for the probability of a soft sweep as

$$P_{\mathrm{soft},n} \leq 1 - \Pr(n|n, \Theta) = 1 - \prod_{i=1}^{n-1} \frac{i}{i+\Theta} = a_{n-1}\Theta + \mathcal{O}(\Theta^2) \, , \tag{2.11}$$

where $a_n = \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{n}$. Eq. 2.11 reduces to (2.8) in the case of $n = 2$. The '$\leq$' expresses the fact that we have overestimated the mutation probability by ignoring the factor $(1 - x_\tau)$ in $P_{\mathrm{mut},k}$. The marginal distributions for the number of haplotypes k and the distribution for fixed k can also be given

$$\Pr(k|n, \Theta) = \frac{\Theta^k S_n^{(k)}}{\Theta(\Theta+1)\cdots(\Theta+n-1)} \tag{2.12}$$

where $S_n^{(k)}$ is Stirling's number of the first kind, and

$$\Pr(n_1 \ldots n_k | k, n, \Theta) = \frac{n!}{k! n_1 \cdots n_k S_n^{(k)}} \, . \tag{2.13}$$

In figures 2.3–2.5 we compare the estimates from equations (2.11)–(2.13) with simulation data for samples that are drawn at the time of fixation of the B allele. As can be seen from figures 2.3 and 2.5, the predictions are good for strong selection. For $\alpha = 100$, the simulation data deviate more strongly. The same effect is seen if the sample is taken a long time after fixation. The reason is the same as for a sample of size two: If the time from the first origin of the allele to the observation of the sample is very long, the small error that we have made by ignoring the factor $(1 - x_\tau)$ in the mutation probability accumulates over many generations. In a time-forward picture, this corresponds to the fact that ancestral haplotypes with a low frequency will slowly drift out of the population. Figure 2.5 shows that the distribution of the remaining haplotypes then becomes more uniform, as is predicted by Kimura (1955). Figure 2.3 also shows that the approximation works best for small samples sizes (see also the supplementary material).

Equation 2.11 and figure 2.4 show that the probability of a soft sweep depends strongly on Θ, the recurrent mutation rate of the beneficial allele on

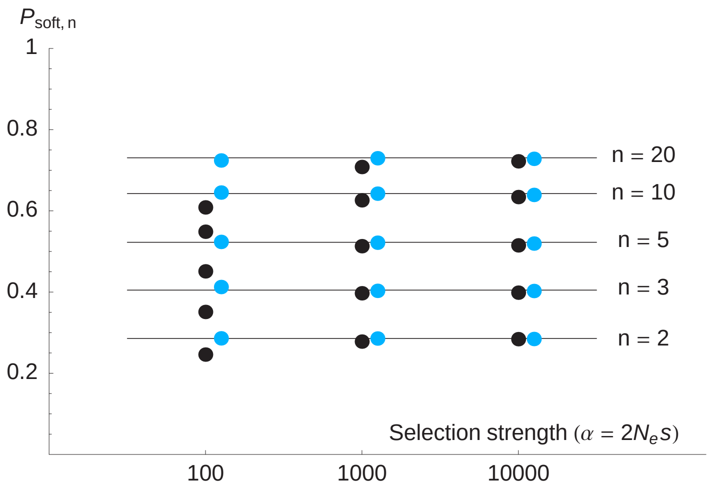


Figure 2.3: The probability of a soft sweep in samples of varying size n, taken directly after fixation. The horizontal lines represent the first order approximation (upper bound, equation 2.11). The dots are simulation results ($N_e = 500{,}000$, 100,000 runs). Black dots are for mutation ($\Theta = 0.4$), grey for migration ($M = 0.4$).

the population level. For low $\Theta < 0.01$, soft selective sweeps from recurrent mutation are rare. For Θ between 0.01 and 0.02 (depending on sample size), they will appear in about 5% of all cases. $0.01 < \Theta < 1$ is the transitional range where both, soft and hard sweeps will be found. For high mutation rates with $\Theta > 1$ almost all selective sweeps will be soft (see figure 2.4). Equation 2.11 shows that there is a logarithmic dependence on the sample size: $P_{\text{soft},n} \approx a_{n-1}\Theta \approx (\gamma + \log(n-1))$ (with Euler's $\gamma = 0.577...$), which can also be seen in figure 2.3.

To leading order, the probability of a soft selective sweep is independent of the selection strength. However, to second order, and as can be seen in figure 2.3, $P_{\text{soft},n}$ increases with selection strength. In other words, the tendency to maintain ancestral genetic variation in the face of positive selection increases with stronger selection. This is in strong contrast to the maintenance of variation due to recombination. As explained above, the reason for this effect is the longer fixation time of weakly beneficial alleles. If we sample at a fixed time after the start of the substitution process, the increase disappears (see also the online material).

Finally, we note that the results are slightly different when we consider the entire population instead of a sample. As we reported previously, the probability for a soft sweep on the whole population level increases with selection

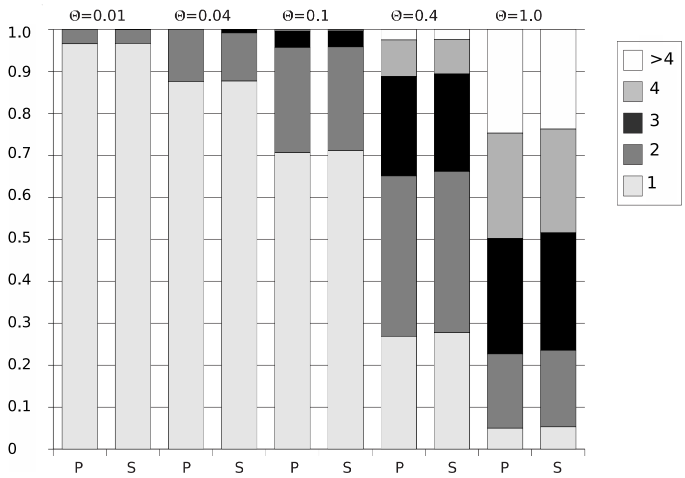


Figure 2.4: The probability of finding 1, 2, 3, 4, or more than 4 ancestral haplotypes (different mutational origins of the B allele) in a sample of 20 for different Θ values. For each Θ value we show the simulation results right (marked S) and the prediction left (according to equation 2.12, marked P). The simulations use $\alpha = 10,000$, the population is sampled directly after fixation.

strength (see Hermisson and Pennings 2005, figure 6). This holds true even if the sample is taken at a fixed time after the start of the substitution process (results not shown). This indicates that under strong selection more alleles are maintained in a population, which afterwards could be picked up by a new selection pressure. Note that our analytical results cannot be extended to the entire population, because the approach depends on the assumption that multiple events in a single generation and coalescent events with multiple mergers can be ignored.

Migration

Instead of new mutation, a beneficial allele can also enter a population through recurrent migration. We consider the following scenario. A population is split into two subpopulations. At the B locus, the subpopulation in the first deme is fixed for the B allele since a long time ago; the second subpopulation is initially fixed for the b allele. We assume that gene flow at the B locus into the second subpopulation was inhibited for a long time, either because of geographical isolation or because of selection against the B allele in the second deme. Now,

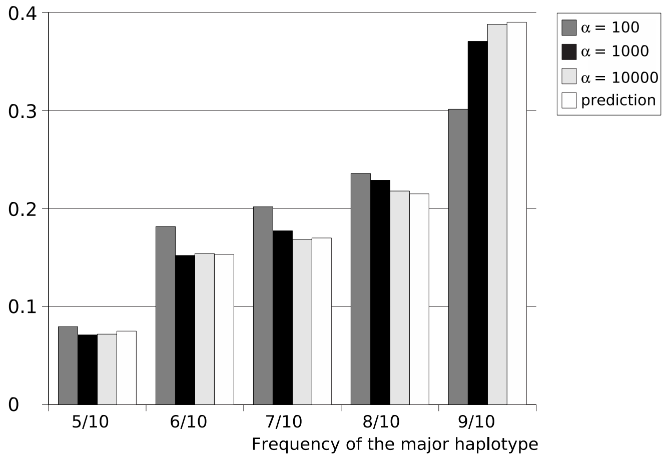


Figure 2.5: Haplotype frequency spectrum: The probability for a major ancestral haplotype with frequency 5 out of 10, 6 out of 10 etc, given that there are two haplotypes in the sample of 10. We show from left to right: $\alpha = 100$; $\alpha = 1,000$; $\alpha = 10,000$; prediction according to equation 2.13. $\Theta = 0.1$

however, both populations are linked through weak migration and the B allele is beneficial in both demes. We assume that a mutational origin of the B allele in the second subpopulation is unlikely and can be ignored. Thus, adaptation in the second deme will only occur from migrants.

Migration is modelled by a fixed probability m for every individual to be replaced by a migrant. For a (fixed) population size of N_e in the second deme, $N_e m$ is then the average number of successful migrants that arrive in that deme per generation. We ignore the possibility that over the relevant time scale (i.e. the typical fixation time of B) a lineage sampled in the second deme migrates to the first deme and back to the second deme. As a consequence, we can entirely focus on the evolutionary process in the second deme and treat the first deme as a reservoir of independent B haplotypes that enter the second deme at a constant rate.

We are interested in the expected number of different B haplotypes and their frequency distribution in a sample from the second deme. As in the mutation model we separate the two stages that produce a new generation and introduce an intermediate step after reproduction but before migration. Using the backward-in-time notation, individuals of generation τ reproduce to form generation $\tau - \frac{1}{2}$ and the individuals in this intermediate generation can be replaced or not by migrants to form $\tau - 1$. A migrant replaces a random resident individual, independent of the resident's genotype. We can thus write $x_{\tau-1}$ in terms of $x_{\tau-\frac{1}{2}}$ as

$$x_{\tau-1} = m + x_{\tau-\frac{1}{2}} \cdot (1-m). \tag{2.14}$$

where the last term represents the resident B's that are not replaced by migrants. For a B lineage, the probability that it has migrated and has an ancestor in deme one in generation τ is

$$P_{\text{mig},1}(\tau) = \frac{m}{m + x_{\tau-\frac{1}{2}} \cdot (1-m)} = \frac{m}{x_{\tau-1}} \ . \tag{2.15}$$

Ignoring the probability that multiple events happen in one generation, and using $x_{\tau-1} \approx x_\tau$, the probability for migration, backwards in time, for k ancestors at time τ is

$$P_{\text{mig},k} \approx \frac{kM}{2N_e x_\tau} \tag{2.16}$$

where $M = 2N_e m$. The probability for migration lacks the $(1 - x_\tau)$-factor of the mutation probability (equation 2.9). While only mutations from b to B introduce a new ancestral haplotype associated with the B allele, every migration, replacing either a b individual or a B individual in the subpopulation in deme two will add a new B haplotype.

We thus see that the migration and coalescence probabilities are strictly proportional, their relative rates do not depend on the frequency x_τ of the B allele. We can therefore directly map the coalescent to a neutral coalescent. The problem is fully solved by the Ewens sampling formula (equations 2.10 - 2.13), with Θ replaced by M, for arbitrary values of the selection coefficient. The simulation data in figures 2.2 and 2.3 show that this estimate is highly accurate. Our results can also be applied if the origin of the B allele in the first deme is more recent (less than $2N_e$ generations ago). In this case, however, there is a higher chance that different B haplotypes have a common origin and are thus similar or even identical.

Generalizations of the model

We have derived our results under a number of simplifying assumptions mainly for the clarity of the presentation. As it turns out, several of these assumptions can be significantly relaxed without changing our results. In this section, we show that the sampling distribution of ancestral haplotypes follows the Ewens sampling scheme as a good first order approximation under a wide range of biological scenarios.

Back mutations Inclusion of back mutations at rate v into the model brings three small changes. First, there is a small additional term proportional to uv added to the mutation probability $P_{\text{mut},n}$ in the coalescent. Second, there is a slight chance that multiple mutations from B to b and back occur on a single line of descent. On the time scales considered here, both these effects can be safely ignored even for high back-mutation rates. Third, back mutation also changes the expected frequency x_τ of the beneficial allele B. In particular, with high v, B may never reach full fixation. However, as long as we condition on samples that are monomorphic for B, this does not affect our results, which do not depend on the stochastic path $\{x_\tau\}_\tau$.

Changing population size In the migration model, we can allow arbitrary changes in the effective population size N_e of the population in the second deme. To maintain our results, we only need to keep the average number of successful immigrants, $N_e m$, (and thus $M = 2N_e m$) fixed. In generations with small N_e, this is compensated by a higher probability m for each individual to be replaced by a migrant. (For recurrent mutation a similar assumption of a constant Θ despite changing N_e does not seem to be meaningful). An important limiting case is that the second deme is initially altogether empty and only colonized by descendents of immigrants that appear at a constant rate. We stress that this is a purely demographic scenario without any positive selection that leads to the same expected pattern of ancestral haplotypes at the study locus. In contrast to selection, however, the pattern should be genome-wide in this case.

Mutation and migration Without any additional problem, we can combine mutation and migration into a single model. To leading order, the sampling distribution of ancestral haplotypes is then still given by the Ewens equations 2.10–2.13, with Θ replaced by $\Theta + M$. The leading correction terms are the same as above and depend on Θ only.

Adaptation from standing genetic variation Since our approximations do not depend on the path, $\{x_\tau\}_\tau$, they are not affected by changes of the selection pressure s as a function of time or of frequency, as long as the fixation time does not become too long. In particular, s may also change its sign during the course of the substitution process. This will be the case if the allele adapts from the standing genetic variation. As in the purely beneficial case, the Ewens approximation will be accurate as long as $(T_{fix} + t_{\mathrm{obs}}) \ll N_e$. This is always the case if selection (either positive or negative) is strong enough. Note that the sampling distribution counts the numbers of *independent* haplotypes (independent origins in HERMISSON and PENNINGS 2005). It does not count the number of descendents of all B copies that segregated in the population at the start of positive selection since the latter may still be identical by descent.

Diploidy and dominance Formally, our derivations above apply for a haploid model or for a diploid model with complete dominance. In these two cases, every B allele in a parent generation has the same expected number of offspring. In a diploid model with dominance coefficient $h < 1$, the expected number of offspring of a B allele depends on whether it comes from a homozygote BB or a heterozygote Bb individual. This increases the variance in offspring number relative to the haploid case, and therefore also the coalescence rate. As shown in the online material, however, the effect is very small, of the order s^2, and can usually be ignored. Dominance further changes the expected frequency path $\{x_\tau\}_\tau$ of the beneficial allele. Since this does not affect our results, they also apply to randomly mating diploids with an arbitrary level of dominance.

Variance in the fitness effects Until now, we have assumed that all beneficial B alleles are of a single type and have the same fitness advantage. If B corresponds to a class of (more of less) physiologically equivalent alleles rather than to a unique molecular genotype, this may not be realistic. It is therefore important to check the stability of our results under variations in fitness among the beneficial alleles. With this aim, we ran additional simulations where we split the B alleles into two classes B_+ and B_-. New mutations are assigned with equal probability to either of these classes. $B_\pm$ alleles have scaled selection coefficients $\alpha_\pm = \bar{\alpha}(1 \pm D)$. With this definition, D is the coefficient of variation of the distribution of α-values.

Figure 2.6 shows that for low Θ there is no visible difference in the number of ancestral haplotypes relative to the homogeneous case ($D = 0$), even for a large variance among the selection coefficients. For higher Θ the probability of a soft sweep is significantly reduced if D gets large. Note, however, that soft sweeps are very likely in this parameter range anyway. For the frequency spectrum, the predictions from the homogeneous case are even more stable. We find no visible deviation from the values predicted by equation 2.13 even

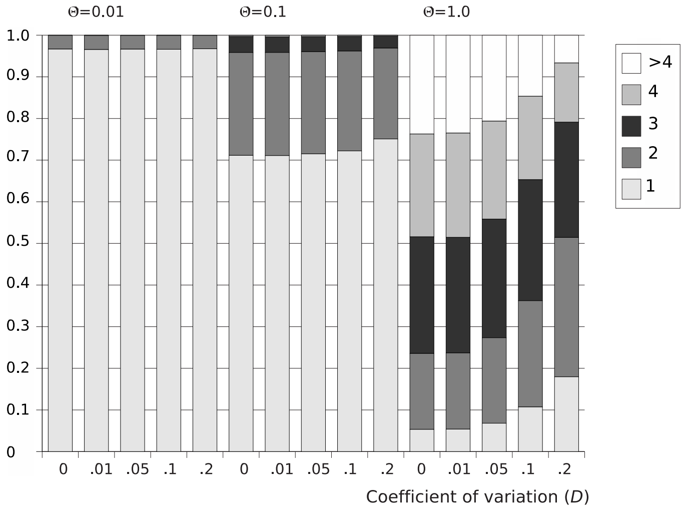


Figure 2.6: Effect of variance in the fitness of B alleles on the number of ancestral haplotypes in a sample. Beneficial mutation produces two kinds of B alleles, $B_{\pm}$, with equal probability. The scaled selection coefficients are $\alpha_{\pm} = (1 \pm D) \cdot \bar{\alpha}$, where the mean selection strength is $\bar{\alpha} = 10,000$. A sample of size 20 is taken at fixation of the $B_{\pm}$ alleles (i.e. when the ancestral b allele is lost). Simulation results are shown for three different Θ values, and values of D ranging from 0 (homogeneous fitness) to 0.2 (corresponding to a 50% larger α_{+} relative to α_{-}).

for $D = 0.2$ (figure S1 in the online material).

2.4 Discussion

How much genetic variation can be maintained in a population in the face of positive selection? Ever since the work of Maynard Smith and Haigh (1974), we know that positive selection removes genetic variation from a population. This has important consequences. First, the characteristic valleys of reduced variation around a selected site can be used to detect loci that underlie adaptation (e.g. Harr *et al.* 2002; Storz *et al.* 2004; Ometto *et al.* 2005; Haddrill *et al.* 2005). Second, if positive selection acts recurrently along the chromosome, it may be selection rather than genetic drift that controls the level of genetic variation in a population. This was formalized in the theory of genetic draft by Gillespie (1991). Positive selection and linkage may also limit the rate of the future adaptive process (Barton 1995).

The classical view is that selection erases all ancestral variation (variation that existed before the onset of selection) unless recombination during the substitution process breaks the linkage between the selected site and its genetic background. The point of this paper is that ancestral variation can also be retained if the favourable allele occurs recurrently and if several independent origins contribute to the adaptive substitution. Positive selection then results in what we call a soft selective sweep. Since every beneficial mutation is eventually recurrent, the crucial question is: for which mutation rate will recurrent mutation result in soft sweeps and thus affect the standard results of genetic hitch-hiking? From our results, we can answer this question as follows. If $\Theta = 2N_e u$ is the population-level mutation rate of the beneficial allele (or allelic class), then:

- For $\Theta < 0.01$, soft sweeps are rare (less than 5%) even in a large sample. In this parameter range, the classical results on hitch-hiking and selective sweeps hold as a good approximation.
- For $\Theta > 0.01$ soft sweeps start to play a role and will be observable for recent substitutions. In a transitional range, $0.01 < \Theta < 1$, soft sweeps coexists with classical hard sweeps. For $\Theta > 1$ almost all adaptive substitutions will result in soft sweeps.
- Analogous results hold if beneficial alleles are introduced by recurrent migration instead of mutation. Other parameters such as selection strength, dominance, etc., play only a minor role.
- Our results show much more than the probability of a soft sweep: For a given Θ, the expected number and distribution of ancestral haplotypes in a sample follows approximately the Ewens sampling formula.

The relatively low values for Θ that are necessary to obtain soft sweeps and the independence of the selection strength may come as a surprise. After all, if selection is strong adaptation is fast and the time for recurrent mutation limited. In fact, input of neutral mutations during the selective phase can often be neglected, even if their combined mutation rate on a DNA fragment is high ($\Theta \approx 10$ typical for *Drosophila* species). So why is the same not true for beneficial mutations that are much rarer? Here, it is important to note that the neutral mutations can be ignored because they are unlikely to be seen in a sample, not because they are unlikely to happen in the population during the selective phase. Also for beneficial mutations, multiple origins during the substitution process are likely, even for quite low values of Θ. And because of their positive fitness, they have a much higher probability to survive stochastic loss and to make it into the sample.

In a forward-in-time picture, this can be estimated as follows. For a beneficial allele with selective advantage $\alpha = 2N_e s$, the average fixation time is $T_{\text{fix}} \approx 4N_e \log(\alpha)/\alpha$. The average number of mutations that occur in this time is $2\Theta N_e \log(\alpha)/\alpha$. To get an idea of this quantity: if $\Theta = 0.01$, $N_e = 2 \cdot 10^6$, and $\alpha = 1,000$, then T_{fix} is about $55,000$ generations and the mutation will occur about 276 times during the fixation process of the first mutation. For neutral mutations, the probability for a given mutation to occur in a sample of size n is about n/N_e (for a starlike phylogeny). We thus obtain a probability of $2n\Theta \log(\alpha)/\alpha$ for recurrent neutral mutations to enter the sample, which strongly decreases with α. In contrast, the probability for beneficial mutations to escape stochastic loss and to appear in the sample is proportional to the selection coefficient s (approximately $2s(1-x)$ if x is the frequency of beneficial alleles that already segregate in the population). As a result, the dependence on s of the probability $P_{\text{soft},n}$ to observe recurrent beneficial mutations in a sample will largely cancel.

The fact that $P_{\text{soft},n}$ and, more generally, the number and distribution of ancestral haplotypes are independent of α, is only one aspect of the remarkable robustness of these estimates. Under the sole assumption that the substitution was relatively fast and recent, the approximations are independent of most details of the adaptive process. They are valid whether beneficial mutations arise through mutation or migration or both, in haploids or diploids, for arbitrary patterns of time dependent or frequency dependent selection, any level of dominance, and even for moderate variance in the selection coefficient among the beneficial alleles. Because of this generality, there should be a realistic chance that patterns associated with soft sweeps can be found in data.

Where should we expect soft selective sweeps due to multiple origins of the beneficial allele in nature? Two factors contribute to Θ, which is the crucial parameter: Soft selective sweeps should be expected if either the effective population size N_e, or the allelic mutation rate u is high. For example, in African *Drosophila melanogaster* with an estimated haploid size $N_e \approx 2 \cdot 10^6$, Watterson's estimator for Θ per site was measured to be $\Theta_W \approx 0.013$ (Ometto *et al.* 2005). This translates into a $P_{\text{soft},n}$ of $\sim 5\%$, if only mutation at a single site produces the beneficial allele. One should note, however, that Watterson's estimator is strongly affected by past demographic events. If the population has experienced recent strong growth, this estimator will severely underestimate the real Θ (which depends on the inbreeding effective population size at the time of the adaptation rather than on the variance effective size). For humans, in particular, it is questionable whether the often-cited low values for $\Theta_W \approx 0.001$ (or $N_e \approx 10,000$) are relevant for recent adaptations (e.g. to agriculture or diseases). A second scenario where soft selective sweeps from recurrent mutation are likely, are adaptations with a high allelic mutation rate, such as adaptive loss-of-function mutations. Finally, situations where beneficial alleles may have been introduced into a population by recurrent migration at a low, but steady rate are easy to imagine.

In human population genetic data, quite a few alleles are known that have risen in frequency due to positive selection and are associated with different haplotypes. These could be cases of soft sweeps from independent mutational origin. Some of these alleles are indeed produced by loss of function mutations (e.g. the FY-0 allele at the Duffy locus, Hamblin and Di Rienzo (2000), α and β thalassemia mutations, Flint *et al.* (1993)), but others are not (e.g. HbS, which causes sickle cell anemia, Flint *et al.* (1993), and HbE, which causes a mild variant of β thalassemia, Antonarakis *et al.* (1982)).

Schlenke and Begun (2005) found three immunity genes in *Drosophila simulans* that show clear signs of soft sweeps. The genes have extreme LD values, in each case caused by two distinct haplotypes at intermediate frequencies that have not recombined. In one case there is also a third invariant haplotype at low frequency. Each of the haplotypes has little or no polymorphism, ruling out the possibility of long-term balanced polymorphisms. The authors also did simulations to rule out the possibility that the patterns are caused by purely demographic scenarios such as bottlenecks. However, the pattern that is found in these three genes is perfectly compatible with soft sweeps.

Pathogens can have extremely high population sizes. It may, therefore, not be surprising that evidence for soft sweeps also comes from a recent study of *Plasmodium falciparum*, with an estimated population size of $10^{10} - 10^{12}$ per infected person (Roper *et al.* 2004). In this study microsatellite variation in both pyrimethamine-resistant and sensitive parasites was studied. The haplotype structure in the data clearly suggests that the double mutant *dhfr* allele (with longer clearance times than the sensitive parasites) in Africa has three independent mutational origins. The triple mutant allele (making the parasite almost resistant) seems, however, to have only one origin (Roper *et al.* 2003). In some cases, for example in viruses, Θ values may be so high that selective sweeps, at least for single mutants, can never be detected. All sweeps would involve alleles of many different origins and there will be no visible signature of selection.

An obvious next step to be taken is to add recombination to the model and study how soft sweeps affect patterns of nucleotide variation at linked neutral loci. Also, more realistic demographic scenarios still remain to be investigated. Aspects that we have not addressed in this paper include changes in population size for the mutation case, or more complex population structures. In general, population structure should make soft sweeps more likely. This is easy to see from the extreme case, where subpopulations are linked by very weak migration. If M is lower than Θ, it is more likely that adaptation in each population will be from its own mutational origin of the beneficial allele. On the meta-population level this would result in a soft sweep.

2.5 Acknowledgments

We thank Peter Pfaffelhuber for fruitful discussions and Sascha Glinka, Andreas Gros, John Parsch, Marcy Uyenoyama, and the referees for helpful comments on the manuscript. This work was supported by an Emmy Noether grant by the Deutsche Forschungsgemeinschaft (DFG) to J.H.

2.6 Supplemetary material

Supplementary figure S1

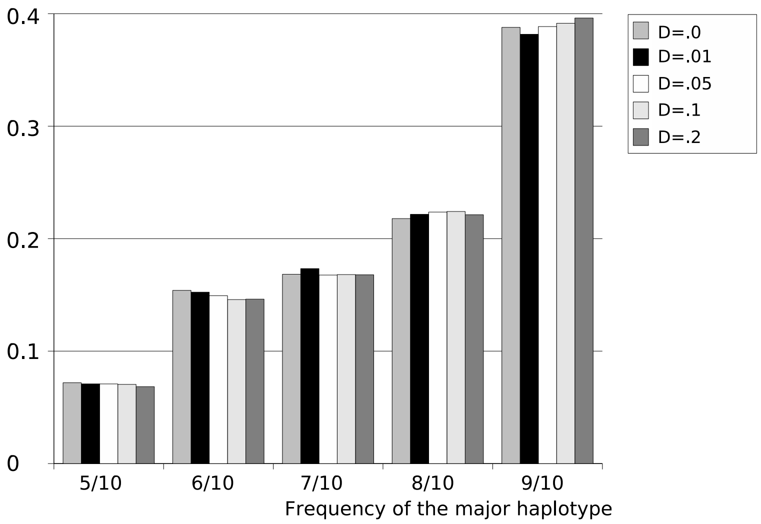


Figure S1: Effect of variance in the fitness of B alleles on the haplotype frequency spectrum: The probability for a major ancestral haplotype with frequency 5 out of 10, 6 out of 10 etc, given that there are two haplotypes in the sample of 10. Beneficial mutation produces two kinds of B alleles, B_+ and B_-, with equal probability. The scaled selection coefficients are $\alpha_\pm = (1 \pm D) \cdot \bar{\alpha}$, where the mean selection strength is $\bar{\alpha} = 10,000$. A sample of size 20 is taken at fixation of the $B_\pm$ alleles (i.e. when the ancestral b allele is lost). Simulation results are shown for three different Θ values, and values of D ranging from 0 (homogeneous fitness) to 0.2 (corresponding to a 50% larger α_+ relative to α_-).

Calculations for larger samples

In this appendix, we derive an approximation for the expected number and distribution of ancestral haplotypes in a sample from a selected locus. If there are k ancestors of the sample at time τ, the probability for mutation and coalescence at this time is approximately

$$P_{\text{mut},k} \approx \frac{k\Theta(1-x_\tau)}{2N_e x_\tau} \quad ; \quad P_{\text{coal},k} \approx \frac{k(k-1)}{2N_e x_\tau} \tag{2.17}$$

The probability that the sample fully coalesces before a beneficial mutation appears in its ancestry then is

$$P_{\text{hard},n} = \Bigg\langle \sum_{\substack{\tau_1,\tau_2,\ldots,\tau_{n-1}=0 \\ \tau_1<\tau_2<\cdots<\tau_{n-1}}}^{\infty} \prod_{k=1}^{n-1} \Bigg\{ \frac{(n-k+1)(n-k)}{2N_e x_{\tau_k}} \cdot \prod_{i=\tau_{k-1}+1}^{\tau_k-1} \left(1 - \frac{(n-k+1)(n-k+\Theta(1-x_i))}{2N_e x_i}\right) \Bigg\} \Bigg\rangle_x \tag{2.18}$$

where $\tau_1, \ldots, \tau_{n-1}$ are the generations where the coalescence events occur, and we define $\tau_0 \equiv 0$. As in the case of a sample of size two, we can proceed by rewriting this equation as

$$P_{\text{hard},n} = \prod_{j=1}^{n-1} \left(\frac{j}{j+\Theta}\right) \Bigg\langle \sum_{\substack{\tau_1,\tau_2,\ldots,\tau_{n-1}=0 \\ \tau_1<\tau_2<\cdots<\tau_{n-1}}}^{\infty} \prod_{k=1}^{n-1} \Bigg\{ \Bigg(\frac{(n-k+1)(n-k+\Theta(1-x_{\tau_k}))}{2N_e x_{\tau_k}} + \frac{(n-k+1)\Theta}{2N_e} \Bigg) \cdot \prod_{i=\tau_{k-1}+1}^{\tau_k-1} \left(1 - \frac{(n-k+1)(n-k+\Theta(1-x_i))}{2N_e x_i}\right) \Bigg\} \Bigg\rangle_x \tag{2.19}$$

Here, $(n-k+1)(n-k+\Theta(1-x_{\tau_k}))/(2N_e x_{\tau_k})$ is the probability that $n-k+1$ lines either mutate or coalesce at time τ_k in a coalescent where lines are stopped at mutational events (cf. Ewens, 2004, p. 335). We can now work with this full coalescent with mutations and treat the terms proportional to $(n-k+1)\Theta/(2N_e) = (n-k+1)u$ as perturbations. Note that these perturbations are a consequence of the factor $(1-x_\tau)$ that appears in the mutation probability, equation (2.17).

Expanding the k-product in the perturbation term, we obtain to leading order (u^0) an expression for the probability that all lines eventually either coalesce or mutate, which is 1, independently of the values of the x_τ. For the linear order, u^1, note that the following sum of products describes the average number of generations $T_k(\tau_{k-1})$ between the $(k-1)$th and the kth event (mutation or coalescence) in the coalescent, given that the $(k-1)$th event has occurred at time τ_{k-1} (and given the path $\{x_\tau\}_\tau$),

$$T_k(\tau_{k-1}) = \sum_{\tau_k=\tau_{k-1}+1}^{\infty} \prod_{i=\tau_{k-1}+1}^{\tau_k-1} \left(1 - \frac{(n-k+1)(n-k+\Theta(1-x_i))}{2N_e x_i}\right) . \quad (2.20)$$

Terms in linear order of u are averages of the $T_k(\tau_{k-1})$ over the realizations of the coalescent (and thus over τ_{k-1}) and over the paths $\{x_\tau\}_\tau$. We now obtain $P_{\mathrm{hard},n}$ up to linear order in u as

$$P_{\mathrm{hard},n} = \prod_{j=1}^{n-1} \left(\frac{j}{j+\Theta}\right) \left[1 + \frac{L_n(t_{\mathrm{obs}})\Theta}{2N_e} + \mathcal{O}(u^2)\right] . \quad (2.21)$$

where

$$L_n(t_{\mathrm{obs}}) = \sum_{k=1}^{n-1} (n-k+1) T_k(\tau_{k-1}) \quad (2.22)$$

is the average total treelength for a sample of size n at t_{obs}. Clearly, $L_n \leq nT_{\mathrm{fix}}$, if the population is observed at the time of fixation. As in the case of a sample of size 2, the population can be sampled at an arbitrary observation time t_{obs}. For $t_{\mathrm{obs}} > 0$ (observation after fixation), the leading order correction term in equation (2.21) will be small as long as $n(t_{\mathrm{obs}} + T_{\mathrm{fix}})/(2N_e) \ll 1$. Under the same sufficient condition, terms of second and higher order in u can be safely neglected relative to the leading order.

From equation (2.21) we can see why, to second order, the probability for a soft sweep increases with selection strength α. This is the case because the coalescence tree for a sample is shorter when selection is stronger if we sample at fixation. Thus $P_{\mathrm{hard},n}$ decreases with α and hence probability of a soft sweep $P_{\mathrm{soft},n}$ increases. However, if we sample at a fixed time after the start of the substitution process, this effect disappears. In that case, the frequency x_τ of the beneficial allele B at generation τ before the observation will be larger, on average, if α is larger. Therefore the coalescence and mutation rates are smaller for stronger selection (cf equation 2.17). As a consequence, coalescent trees will be, on average, slightly longer, which translates in a slight reduction of $P_{\mathrm{soft},n}$ with increasing α. This is also observed in simulations (results nor shown).

Time rescaling

In order to map the coalescence process under hitch-hiking onto a neutral coalescent we need to get rid of the dependence on x_τ. Formally this is done by a time rescaling. For this, we formulate the problem in continuous time. The probability for an event (coalescence or mutation) in the time interval $\tau, \tau + \delta\tau$ is then approximately

$$\frac{n(n-1)}{2N_e x_\tau}\delta\tau + \frac{n\Theta}{2N_e x_\tau}\delta\tau = \frac{n(n-1+\Theta)}{2N_e}\delta\tilde{\tau}(x) \tag{2.23}$$

where $\delta\tilde{\tau} = (1/x_\tau)\delta\tau$ defines a new time scale that depends on the random path $\{x_\tau\}_\tau$. We can now use the fact that the distribution of coalescent tree topologies is independent of the time scale. Since the distribution of haplotypes in a sample is only a function of the tree topology, the problem is equivalent to the neutral coalescent in the infinite sites model (Ewens, 2004, p. 335ff).

Coalescent probability for diploids with dominance

Consider a randomly mating diploid population of constant size N_e. There are two alleles and three genotypes at the study locus, bb, bB, and BB, with corresponding fitness values 1, $1+hs$ and $1+s$. Let x be the frequency of the B allele in generation t. We assume Hardy-Weinberg equilibrium in every generation, such that the genotype frequencies are $P_{bb} = (1-x)^2$, $P_{bB} = 2x(1-x)$ and $P_{BB} = x^2$, respectively.

Consider now all B alleles that are found in the population at generation $t+1$. The probability that a randomly drawn B allele from that generation derives from a bB parent is

$$\frac{2x(1-x)(1+hs)N_e}{2x(1-x)(1+hs)N_e + x^2(1+s)2N_e} = \frac{(1-x)(1+hs)}{1+hs+xs(1-h)}\,. \tag{2.24}$$

Similarly, the probability for a B allele to come from a BB parent is

$$\frac{x(1+s)}{1+hs+xs(1-h)}\,. \tag{2.25}$$

Two B alleles from generation $t+1$ can coalesce in generation t if either both have a bB parent or both have a BB parent. The coalescence probability for B alleles in generation t is therefore

$$\begin{aligned} P_{\text{coal},x} &= \frac{(1-x)^2(1+hs)^2}{(1+hs+xs(1-h))^2}\frac{1}{2x(1-x)N_e} + \frac{x^2(1+s)^2}{(1+hs+xs(1-h))^2}\frac{1}{x^2 2N_e} & (2.26)\\ &= \frac{1}{2N_e x}\left(1+\frac{x(1-x)s^2(1-h)^2}{(2+hs+xs(1-h))^2}\right). & (2.27)\end{aligned}$$

The result shows that the coalescence probability is slightly enhanced relative to a haploid population with size $2N_e$ if $h<1$ (no complete dominance). The reason is the larger variance in offspring number in the diploid population due to the different types of genotypes a B allele can come from. This variance is largest for intermediate frequency of the b allele, $x=0.5$. Note, however, that the correction term is generally small (proportional to s^2) and even smaller for small x, when coalescence during a substitution of a beneficial allele B is most important. For most practical purposes, the haploid formula should therefore be a good approximation.

Chapter 3

Soft Sweeps III – The signature of positive selection from recurrent mutation

Pleuni S. Pennings and Joachim Hermisson

Published in PLoS Genetics (2006) 12:e186

Polymorphism data can be used to identify loci at which a beneficial allele has recently gone to fixation, given that an accurate description of the signature of selection is available. In the classical model that is used, a favored allele derives from a single mutational origin. This ignores the fact that beneficial alleles can enter a population recurrently by mutation during the selective phase. In this study, we present a combination of analytical and simulation results to demonstrate the effect of adaptation from recurrent mutation on summary statistics for polymorphism data from a linked neutral locus. We also analyze the power of standard neutrality tests based on the frequency spectrum or on linkage disequilibrium (LD) under this scenario. For recurrent beneficial mutation at biologically realistic rates we find substantial deviations from the classical pattern of a selective sweep from a single new mutation. Deviations from neutrality in the level of polymorphism and in the frequency spectrum are much less pronounced than in the classical sweep pattern. In contrast, for levels of LD the signature is even stronger if recurrent beneficial mutation plays a role. We suggest a variant of existing LD tests that increases their power to detect this signature.

3.1 Introduction

Patterns of DNA polymorphism can be used to infer the processes that have played a role in the evolutionary history of a population. A process that is of primary interest to evolutionary biologists is directional selection, and the pattern that is left by it, a so-called selective sweep, has received a lot of attention in the literature since it was first described by Maynard Smith and Haigh (1974). By now, this pattern is well studied, at least for a simplified model, which assumes that a single adaptative mutation increases in frequency under constant selection pressure in a panmictic population of constant size (e.g. Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 1995; Durett and Schweinsberg 2004; Etheridge *et al.* 2005; Stephan *et al.* 2006). The signature that is created if these assumptions are met is characterized by (1) low polymorphism around the selected site, (2) an excess of low frequency variants both at the locus itself and in the flanking regions, (3) an excess of high frequency variants only in the flanking regions, and (4) strong linkage disequilibrium (LD) in the flanking regions, but no LD between mutations on opposite sides of the selected locus. There is a body of statistical tests based on these characteristics (e.g. Tajima 1989; Depaulis and Veuille 1998; Fay and Wu 2000; Kim and Stephan 2002; Nielsen *et al.* 2005), which have been used in a large number of studies seeking to identify loci that have undergone directional selection (e.g. Hamblin and Di Rienzo 2000; Storz *et al.* 2004; Akey *et al.* 2004; Catania *et al.* 2004; Ometto *et al.* 2005; Haddrill *et al.* 2005; Schlenke and Begun 2005).

One assumption of the simplified model is that only descendants of a single copy of the beneficial allele contribute to fixation. This may be different if (a) selection acts on the standing genetic variation or (b) adaptation occurs from recurrent mutation or migration. If (descendents of) multiple copies of a beneficial allele are involved in its fixation, this has consequences for the signature of selection. We therefore call such a signature a "soft selective sweep" and distinguish it form the classical pattern of a "hard sweep", where only a single copy is involved (Hermisson and Pennings 2005).

Adaptation from the standing genetic variation has been described in a series of recent articles (Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski *et al.* 2005; Teshima *et al.* 2006). Substantial changes to the classical hard sweep are observed, in particular, if the allele had been neutral prior to the onset of positive selection. The second scenario, adaptation from recurrent mutation or migration, was analyzed in (Pennings and Hermisson 2006). It turns out that soft selective sweeps from recurrent mutation are relevant if $\Theta_b > 0.01$ (where $\Theta_b = 2N_eu_b$ is the the population mutation parameter for the beneficial allele). Soft sweeps, under these conditions, are therefore likely if either the (inbreeding-)effective population size N_e is large or if the allelic mutation rate u_b is high. For example, LI and STEPHAN (2006), estimate that for African *Drosophila melanogaster*, which has high N_e, the mutation parameter per site is about 0.05. Since the allelic mutation rate Θ_b will usually be equal to or higher than the rate per site, soft sweeps from recurrent mutation should be frequent for this species. A large Θ_b is also expected, even for populations with moderate or small N_e, if adaptation involves a loss- or reduction-of-function mutation. Adaptive loss-of-function mutations have recently been identified in many species, such as humans (e.g., HAMBLIN and DI RIENZO 2000; WANG *et al.* 2006; XUE *et al.* 2006), *Drosophila melanogaster* (TAKAHASHI *et al.* 2001), *Arabidopsis thaliana* (SHIMIZU *et al.* 2004), and rice (OLSEN and PURUGGANAN 2002).

In this study, we describe how a soft sweep from recurrent mutation affects a neutral locus at some recombinational distance from the selected locus and which tests can be employed to detect soft sweeps. We will see that the deviation from the classical hard sweep pattern is even stronger than for adaptation from standing genetic variation. The reason is that haplotypes that are associated with different mutational origins of a beneficial allele are truly independent. In contrast, multiple copies of the beneficial allele that segregate in the standing genetic variation may still be identical by descent.

In the following, we first derive formulas for the site-frequency spectrum and the number of haplotypes at a locus tightly linked to the selected site. We compare the effects of recombination and recurrent beneficial mutation on the polymorphism pattern and explain the differences from the different timing of these events in the coalescent of a sample. In a second step, we describe the combined effect of recurrent mutation and recombination on summary statistics for DNA polymorphism at a linked neutral locus. Finally, we present a power analysis of various neutrality tests. Recent soft sweeps from recurrent mutation can be detected very well using LD based tests, but not using frequency spectrum based tests. We show that older sweeps can also be revealed by LD tests if information from a recently derived sister population is available.

3.2 Model and Methods

We consider a haploid population of constant effective size N_e. At a locus under selection there are two alleles, an ancestral allele b with fitness 1 and a beneficial variant B with fitness $1 + s$. The B allele may also correspond to a class of physiologically equivalent alleles, in which case we assume that these alleles are at the same locus and tightly linked. Mutation from b to B happens at rate u_b, back mutation is ignored. We study the polymorphism pattern at a neutral locus at a recombinational distance r from the selected site. The neutral mutation rate at the study locus is u_n and we assume an infinite sites model for this locus. Recombination within the neutral locus is denoted by r_n. We define population level parameters as $\Theta_b = 2N_e u_b$ (beneficial mutation rate for the allele), $\Theta_n = 2N_e u_n$ (neutral mutation rate), $R = N_e r$ (recombination rate between the selected and neutral locus), $R_n = N_e r_n$ (recombination rate between the two ends of the neutral locus) and $\alpha = N_e s$ (strength of selection). If we set $r = r_n = 0$, and assume that Θ_n is so high that two random halpotypes from the population are always different, the model is identical to the model from Pennings and Hermisson (2006).

We use a coalescent framework and define τ as the time in the past before fixation of the B allele. The frequency of the B allele is denoted by x_τ. The time from the first origin of a B allele that will contribute to fixation and $x_\tau = 1$ is referred to as the selective phase, and the length of the selective phase is T_{fix} generations. In the selective phase, the population can be separated in a growing B part and a shrinking b part (forwards in time). We can therefore use a structured coalescent to derive the sampling distributions at the neutral locus (Kaplan *et al.* 1989). If a $b \rightarrow B$ mutation happens during the selective phase, a new lineage enters the B part of the population. If this happens in the history of a sample, we call it a soft sweep. In a coalescent view, lineages in a sample at the selected locus can coalesce with each other or they can escape the B population by mutation. At the neutral locus, also a lineage can escape by recombination (see figure 3.1).

Simulations of positive selection. We used the program of Kim and Stephan (2002), to which we added the possibility of recurrent beneficial mutation. In the simulations, a neutral fragment is affected by the fixation of a beneficial allele at a nearby selected site. The fragment starts directly next to the selected site (at distance $R = 0$), or at one of five recombinational distances away from it ($R = 10; 20; 100; 200; 600$). Recombination and neutral mutation within the neutral fragment happens at rate $R_n = 10$ and $\Theta_n = 10$ (except

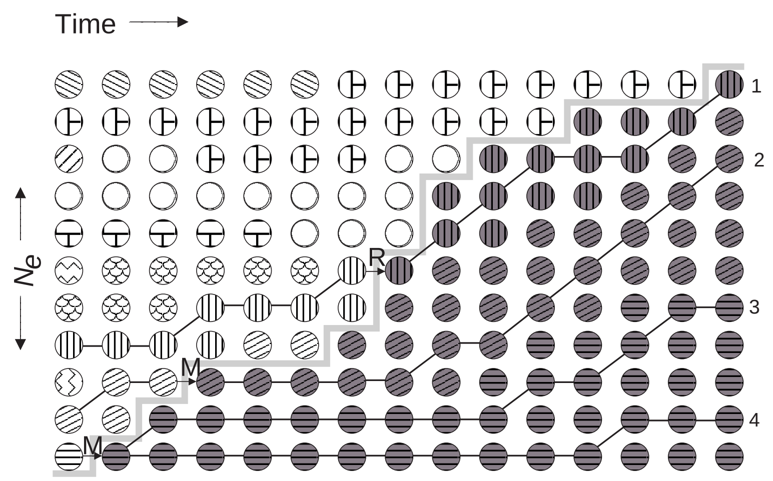


Figure 3.1: Selective sweep with recurrent mutation and recombination in a schematic Wright-Fisher model. Circles represent individuals in the population, the different patterns indicate independent haplotypes at the neutral locus. An individual is dark grey when it is associated with the beneficial allele B at the selected site and white when it is associated with the ancestral b allele. The B allele arises two times by independent mutations (indicated by M); individuals then change their color from white to grey, but keep their pattern. Similarly, a b lineage can recombine onto a B allele (indicated by R), in which case the individual also changes its color and keeps its pattern. Directly after fixation ($t = 0$), we take a sample of three individuals. If the sample would contain individuals $(2, 3, 4)$, it would have two ancestral haplotypes because it is a soft sweep. If the sample would be $(1, 3, 4)$ it would also contain two ancestral haplotypes, but this time because of recombination. In a coalescent view, both 1 and 2 escape the B part of the population.

for some additional simulations described in the text). This corresponds to a 500 bp long fragment if the per nucleotide mutation rate and recombination rate are both $1 \cdot 10^{-8}$, and the population size is $N_e = 1,000,000$. For all our figures, we assumed strong selection ($\alpha = 10,000$). Results from additional simulation runs with $\alpha = 1,000$ are described in the text. For all figures, we ran $10,000$ simulations per parameter combination.

A sample taken at tN_e generations after fixation of the beneficial B allele is simulated. For this, a coalescent graph with recombination is built backwards in time in three phases. The simulation starts with a standard ancestral recombination graph during the neutral phase from tN_e generations after fixation to fixation, followed by a structured coalescent during the selective phase, and finally a second neutral phase with an ancestral recombination graph before the origin of the B allele. This last phase lasts until all lineages have coalesced.

Backward in time, lineages can coalesce, they can recombine and they can mutate from B to b. During the neutral phases, coalescence can happen between all lineages and only recombination within the fragment is modelled. During the selective phase, coalescence can only happen between lineages in the same part (b or B) of the population. Recombination can happen either within the fragment or between the fragment and the selected site. In the latter case, it is only of interest whether the lineage changes the subpopulation that it belongs to, lineages can recombine from the B subpopulation into the b subpopulation and vice versa. When the breakpoint of the recombination event is within the fragment, the lineage splits in two and the part that is furthest from the selected site may change the subpopulation that it is in. Mutation from B to b (in the backward direction) can only happen during the selective phase, with the probability given in equation 3.4. Mutation from b to B is ignored.

The structured coalescent during the selective phase is conditioned on the frequency of the beneficial allele $x_\tau(0 < \tau < T_{fix})$, which is obtained by conducting for each replicate an independent forward in time simulation using a Wright-Fisher model with recurrent beneficial mutation. In the model without recurrent mutation (hard sweep model) we inserted a single beneficial mutant in the population. Conditioning on fixation was done by discarding all runs where the B allele did not go to fixation. Tajima's D is only defined if there is at least one polymorphic site and Kelly's ZnS is only defined if there are at least two polymorphic sites. For the means and standard deviations in figure 3.5, runs for which a statistic is not defined were taken out. The code was checked by comparing the probability of a soft sweep in backwards-in-time simulations against results from forward-in-time simulations.

Power analysis. Outcomes of the simulations with positive selection were compared with the critical values from neutral simulations with the same number of polymorphic sites (S), to check for significant deviations from the neutral expectation. Critical values were obtained from Hudsons ms (Hudson 2002) program, conditional on the number of polymorphic sites (as in, e.g., Depaulis *et al.* 2001; Przeworski 2002). Because we expect deviations of Tajima's D test in two directions, we used the test as a two-sided test, unlike Przeworski (2002) but like Depaulis *et al.* (2005). Using the neutral simulations we determined the D value at 2.5% and 97.5% of the distribution for each value of S. For the other tests (Fay and Wu's H, haplotype K and Kelly's ZnS test), we expect deviations due to positive selection in only in one direction. They were therefore implemented as one-sided tests. We assumed no recombination in the neutral simulations, which is the conservative choice because it will lead to stronger LD. For the power analysis we do not exclude runs in which there are no polymorphic sites, unlike PRZEWORSKI (2002) but like DEPAULIS *et al.* (2005). This is because we are interested in the probability that we can detect an episode of selection with a given neutrality test. If there are no polymorphic sites ($S = 0$), selection cannot be detected with a test that is conditioned on S.

For the tests for which we excluded new mutations we used KIM and STEPHAN (2002)'s program, conditional on S, to obtain the critical values for each t value (time after fixation). In these simulations, no mutations were allowed on in the last tN_e generations of the coalescent tree and we again assumed no recombination.

3.3 Results

Polymorphism pattern at a tightly linked locus

Approximate analytical results are possible for the expected polymorphism pattern at a locus that is tightly linked to the selected site ($r = r_n = 0$). In PENNINGS and HERMISSON (2006) we were interested in the number of ancestors a sample has at the beginning of the selective phase (forward in time). Each ancestor corresponds to an independent ancestral haplotype, i.e. an independent random pick from the ancestral population, before the onset of positive selection. Note that these draws do not necessarily result in *different* haplotypes. We showed that the distribution of independent ancestral haplotypes in a sample that is taken after fixation of the beneficial allele is approximately given by Ewens sampling formula. If we want to determine the frequency spectrum of polymorphic sites, we need to trace the history of the sample further back in time.

In the following, we assume that the population has been in neutral equilibrium prior to the single episode of positive selection that we consider (see the appendix for some added generality on this point). Because the relationship between the ancestral haplotypes is then given by a neutral coalescent, we need to combine Ewens' sampling formula for the distribution of ancestral haplotypes with a neutral coalescent for the history of these ancestral haplotypes. We find that the probability that a mutation is carried by ℓ individuals out of n is

$$P_{\text{anc}}[\ell|n] = \sum_{k=2}^{n} \frac{\Theta_b^k}{\Theta_{b(n)} - (n-1)!} \sum_{j=1}^{k-1} \frac{\binom{n}{\ell}}{j a_k \binom{k}{j}} \cdot S_\ell^{(j)} S_{n-\ell}^{(k-j)}. \tag{3.1}$$

(with $a_k := \sum_{i=1}^{k-1} \frac{1}{i}$; $\Theta_{b(m)} := \prod_{i=0}^{m-1}(\Theta_b + i)$ and $S_n^{(k)}$ is the nonnegative Stirling number of first kind). The derivation is in the appendix. In figure 3.2 we compare this prediction with simulation results. For the approximation, we have ignored neutral mutations during the selective phase, but they are included in the simulations. As can be seen in figure 3.2, the approximation holds very well for large α. For smaller α, an access of singletons becomes visible as neutral mutations accumulate during the selective phase (not shown).

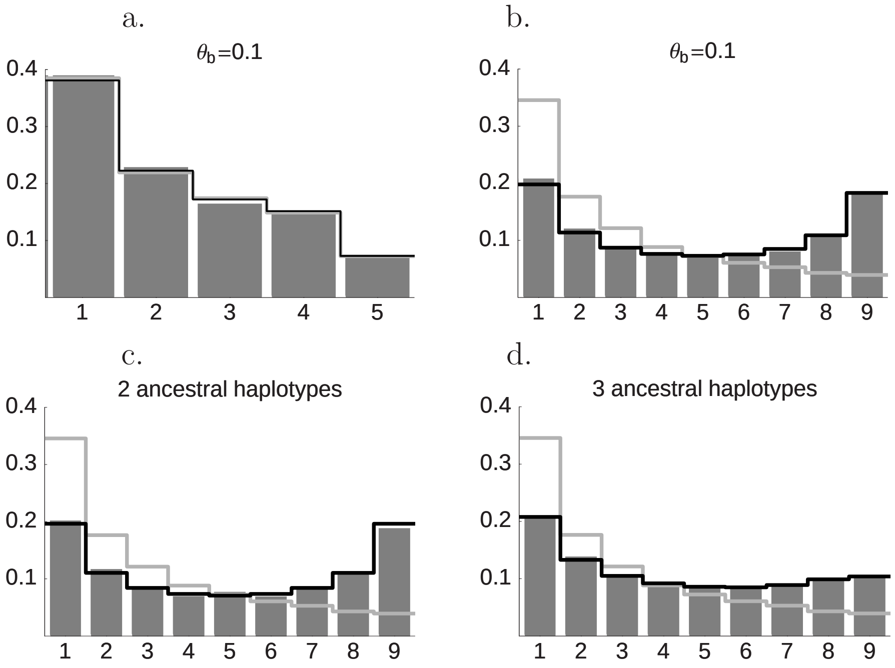


Figure 3.2: Frequency spectrum at fixation. Simulations are done without recombination, but with new mutations during the selective phase. The bars are simulation results, the black lines are the predictions from formula 3.1. The light grey line is the frequency spectrum under neutrality. a. Frequency spectrum at the time of fixation in a sample of 10, $\Theta_b = 0.1$. If there is only one ancestral haplotype (hard sweep) there will be no polymorphic sites, so conditioning on soft sweeps does not change the frequency spectrum. b. Same as a. but now polarized (see text). c. Same as b. but after a soft sweep with exactly two ancestral haplotypes (this frequency spectrum is symmetrical). d. Same as b. but after a soft sweep with three ancestral haplotypes.

A few things become clear from figure 3.2. First, panel (a.) shows that the folded frequency spectrum after a sweep for $\Theta_b = 0.1$ is virtually the same as the neutral expectation. In fact it is exactly the same if there are exactly two ancestral haplotypes (which is the most common outcome for $\Theta_b = 0.1$).

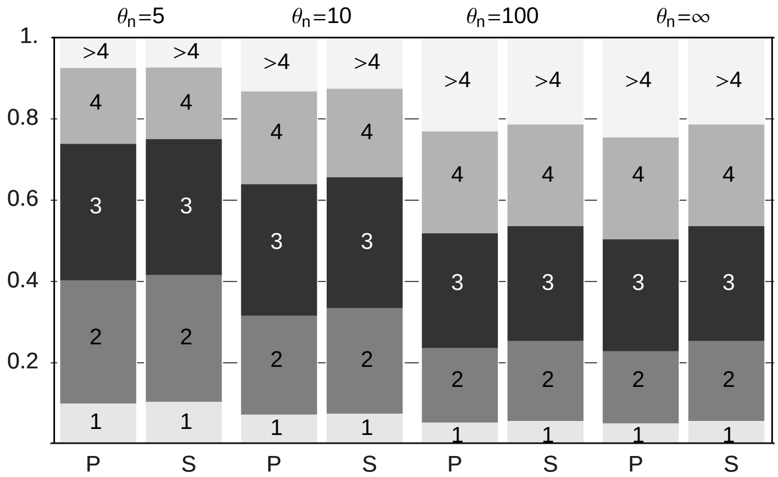


Figure 3.3: Probability of finding 1, 2, 3 etc. distinct haplotypes depending on the neutral mutation rate Θ_n, in a sample of 20 at the time of fixation, with $\Theta_b = 1.0$. Predictions from formula 3.2 are labeled P, simulation results are labeled S. Simulations are done without recombination and neutral mutations during the selective phase.

Second, the polarized (or unfolded) frequency spectrum is very different from the neutral expectation (see panel b.). There is a clear excess of high frequency variants when there are two or three ancestral haplotypes in the sample (panels c. and d.).

If there are two ancestral halpotypes, the polarized frequency spectrum is symmetrical. In this case, sites can only stay polymorphic if one variant is associated with the first beneficial mutation and the other is associated with the other beneficial mutation. The beneficial mutations, and therefore the neutral variants, must have complementary frequencies. They therefore have equal probability to end up in the major or in the minor haplotype, which results in the observed symmetry.

The number of distinct haplotypes can be lower than the number of independent ancestral haplotypes as defined in Pennings and Hermisson (2006), because there is a chance that independent ancestral haplotypes are identical. Whether ancestral haplotypes are the same or different is an infinite alleles problem and can therefore be described by a coalescent with killings (Ewens 2004). The number of distinct haplotypes, if the number of ancestral haplotypes is known, is given by Ewens sampling formula. To know the number of distinct haplotypes in a sample, we therefore need to combine two Ewens sampling formulas, one that tells us the number of ancestral haplotypes and one that tells us how many of these are distinct. The probability that there are ℓ distinct haplotypes in a sample is given by

$$\Pr[\ell|n, \Theta_b, \Theta_n] = \sum_{k=\ell}^{n} \frac{\Theta_n^\ell \Theta_b^k S_k^{(\ell)} S_n^{(k)}}{\Theta_{n(k)} \Theta_{b(n)}} \tag{3.2}$$

(the derivation is given in the appendix). In figure 3.3, the prediction from formula 3.2 is compared with simulation results. For $\Theta_n \to \infty$, the probability that two ancestral haplotypes are different is 1 and the number of distinct haplotypes is the same as the number of ancestral haplotypes. For lower values of Θ_n there may be fewer distinct than ancestral haplotypes. The difference is clearest for the categories with many haplotypes, because if many haplotypes are sampled from the population, it becomes less likely that they are all different. If there are only two ancestral haplotypes, they are distinct with probability $\frac{\Theta_n}{1+\Theta_n}$ (which is ≈ 0.91 for $\Theta_n = 10$). The expected number of haplotypes under neutrality, for $\Theta_n = 10$, is about 11 haplotypes. The number of distinct haplotypes in the sample after a soft sweep is therefore still much lower than the neutral expectation.

The footprint of selection at a linked locus

To describe the footprint of selection at a neutral locus at some distance from the selected locus, we need also to take recombination into account. When we trace the ancestry of a sample back in time, three things of interest can happen. (1.) Two lineages can coalesce when they find a common ancestor, (2.) one lineage can choose as its ancestor a b individual that has mutated into a B individual and thus escape the sweep (note that mutation happens at the associated selected locus and not at the neutral locus that we follow), or (3.) one lineage can recombine onto a b background. We assume the population is large and can therefore set $x_{\tau-1} \approx x_\tau$. The probabilities of coalescence, mutation and recombination in a generation τ, when there are k lineages left are given by:

$$\rho_{coal}(k, \tau) = \frac{k(k-1)}{2} \frac{1}{N_e x_\tau} \tag{3.3}$$

$$\rho_{mut}(k, \tau) = k \frac{\frac{1}{2}\Theta_b(1 - x_\tau)}{N_e x_\tau} \tag{3.4}$$

$$\rho_{reco}(k, \tau) = k \frac{\frac{1}{2}R(1 - x_\tau)}{N_e} \tag{3.5}$$

(e.g., Barton *et al.* 2004; Pennings and Hermisson 2006). Consider now a sample of size two. We are interested in the timing of the first event in the coalescence process of this sample and in the type of this event. The probability that the first event occurred τ generations ago and that this event was a beneficial mutation is

$$P_{\mathrm{mut},2}(\tau) \approx \rho_{mut}(2,\tau) \cdot \prod_{i=1}^{\tau-1} \big(1 - \rho_{\mathrm{coal}}(2,i) - \rho_{\mathrm{mut}}(2,i) - \rho_{\mathrm{reco}}(2,i)\big); \qquad (3.6)$$

where the product is the probability that no event has happened until $\tau - 1$. Equivalent equations hold for coalescence and recombination.

Figure 3.4 shows how the probabilities for each of these events and the frequency of the B allele change in time. It shows clearly that mutation events happen early in the selective phase, just like coalescence events. Recombination, on the other hand, happens later. We can see from formulas $1 - 3$ what causes this difference. Both the coalescence probability and the mutation probability have a $\frac{1}{x}$-term, but the recombination probability does not. This $\frac{1}{x}$-term causes the coalescence and mutation probabilities to rise steeply when the frequency of the B alleles goes down. The recombination probability has only a $(1 - x)$-term, which means it will go up when x goes down, but much less so.

Backwards in time, most recombination events happen at a time where it is unlikely that coalescence events have happened already. This separation in time of recombination and coalescence is used already in Maynard Smith and Haigh (1974). Durett and Schweinsberg (2004) and Etheridge *et al.* (2005) show that this is valid as a first order approximation in α. Recombination therefore tends to make single lineages escape and produces strongly unbalanced trees and polymorphism patterns with an excess of low frequency alleles. In contrast, the distributions for mutation and coalescence events fully overlap, which means that for larger samples it is likely that some coalescence events have happened before a mutation event and some after. As a consequence of this timing, family sizes of an escaping lineages can be anything from just one to almost all lineages. Mutation will therefore create very a different frequency spectrum (as seen in figure 3.2) than recombination.

That a recurrent beneficial mutation tends to happen early in the selective phase can also be understood in a forward in time picture. First, in order to appear in a sample, the mutation needs to reach a high frequency and this is more likely if it happens quickly after the first mutation. Second, early mutants have a higher probability to escape stochastic loss because the mean fitness in the population is still lower and therefore the relative fitness of a mutant higher. Third, simply more $b \rightarrow B$ mutations happen in the beginning of the selective phase because there are more b alleles in the population at this time.

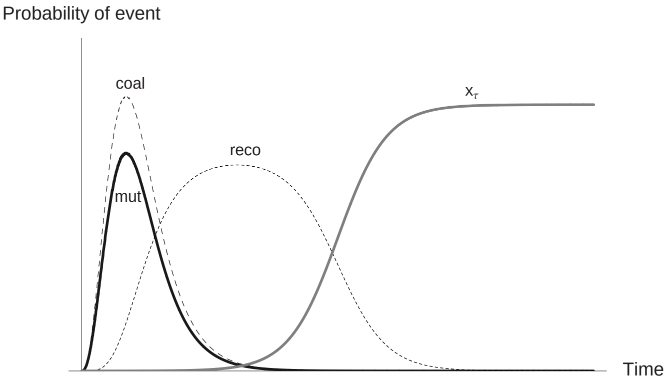


Figure 3.4: Timing of coalescence, recombination and mutation events during the selective phase in a sample of two. This plot shows the probability that recombination, mutation or coalescence happens during the selective phase when we trace the ancestry of a sample of size two back in time. The parameter values for this plot are chosen so that the timing of the three events is made clear, no importance should be given to the relative heights of the curves. The curve with label x_τ shows the frequency of the B allele in the population.

The effect of recurrent mutation on summary statistics

In order to describe the effect of positive selection under recurrent mutation on the polymorphism pattern, we consider a sample from a linked neutral locus that is taken at fixation of the beneficial allele. We derive analytical approximations for the number of pairwise differences π and the number of polymorphic sites S. These approximations are complemented by coalescent simulations for π, S, Tajima's D, Kelly's ZnS and the number of haplotypes K under neutrality and three scenarios for a selective sweep (figure 3.5): (a) a standard sweep model without recurrent mutation (hard sweep), (b) a sweep model with $\Theta_b = 0.1$ where we conditioned on soft sweeps (i.e. only those simulation runs were considered where a soft sweep had happened), (c) a sweep model with $\Theta_b = 1.0$. About 95% of all sweeps are soft in this case.

For our analytical approximations, we ignore neutral mutation during the selective phase. Following our results from the last section we also assume a complete separation in time between recombination on the one hand and coalescence and beneficial mutation on the other hand. This means that, in a coalescent framework, recombination during the selective phase is considered first, while coalescence and beneficial mutation events all occur right at the start of this phase. Finally, we ignore all events (recombination or coalescence) in the b part of the population. Coalescent simulations treat the full model, without any of these approximations.

Pairwise difference (π). Under the above assumptions, a pairwise difference can only occur if one of the two lineages escapes the B part of the population by recombination or mutation. If this happens, the probability that the site is polymorphic is the same as it was under neutrality. Recombination can happen anywhere during the selective phase, with rate $2r$ (for two lineages) per generation. We are, however, only interested in recombination events that involve b alleles, which will be the case for half of the events. Namely, averaged over the time of the selective phase the fraction of b alleles in the population is $\frac{1}{2}$. The number of relevant recombination events is therefore Poisson distributed with parameter $2rT_{\text{fix}}/2$, where T_{fix} is the fixation time. The probability that at least one recombination event happens is therefore

$$P_{reco} \approx 1 - \exp(-\frac{1}{2}rT_{fix}) \approx 1 - \exp(-R\frac{2\log[\alpha]}{\alpha}) \tag{3.7}$$

where we use $T_{fix} \approx N_e\frac{2\log[\alpha]}{\alpha}$ (Hermisson and Pennings 2005). This result coincides with Etheridge *et al.* (2005) and Nielsen *et al.* (2005). If no recombination with a b lineage has happened, there is a probability $\frac{1}{1+\Theta_b}$ that the lineages coalesce before a beneficial mutation happens (Pennings and Hermisson 2006). The probability that neither recombination nor mutation happens is then

$$\frac{1}{1+\Theta_b}\exp[-R\frac{2\log(\alpha)}{\alpha}]$$

And the expected π given the neutral π_n is

$$\pi = \pi_n \cdot \Big(1 - \frac{1}{1+\Theta_b}\exp[-R\frac{2\log(\alpha)}{\alpha}]\Big). \tag{3.8}$$

In figure 3.5, we compare this result with simulation data. The approximation works well as long as R is not too large (a1 and c1). For large R, lineages that have escaped from the B part of the population through recombination, may enter it again through another recombination event. This is ignored in the analytical approximation, which therefore overestimates π at large distances. The effect of recurrent mutation on the signature in π is straightforward: Since for a soft sweep, polymorphism is even maintained at $R = 0$, the depth of the reduction in π is reduced (b1 and c1).

The number of polymorphic sites (S). In our approximation, the number of polymorphic sites depends only on the number m of lineages at the start of the selective phase. These ancestral lineages are related by a neutral coalescent, and for m ancestors the expected number of polymorphic sites is $\Theta_n a_m$. For m we need to add up all lineages that escape the sweep by either recombination or beneficial mutation. The derivation and the result are given in the Appendix. The prediction is compared to simulation data in figure 3.5a2 and 3.5c2. For $R > 0$ the approximation is a bit worse than for π. The reason is that the separation in time of recombination and coalescence is less good for larger samples.

Just like for π, the footprint in S is weakened due to soft sweeps. When scanning for sweeps, low S or π is often the first indication that there may have been a sweep near the studied fragment. It is therefore important to realize that, contrary to a hard sweep, a soft sweep will usually not be characterized by a very low π or S.

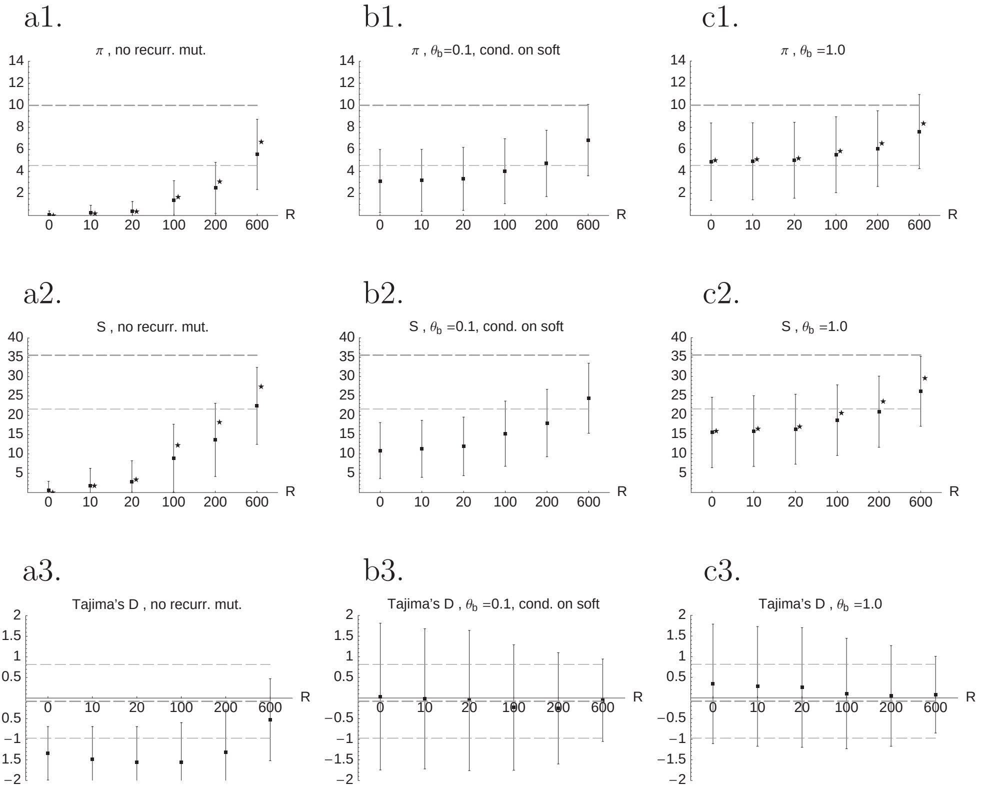
Tajima's D. Tajima's D is a frequency spectrum based test statistic (Tajima 1989). Roughly, it measures the contribution of intermediate frequency mutations to the total number of mutations. When this contribution is higher than expected, Tajima's D is positive, when lower, D is negative. After a hard sweep, Tajima's D tends to be very negative in the flanking regions, because recombination produces an excess of low frequency mutations. In contrast, this effect is almost not visible after a soft sweep. In fact, for soft sweeps, the mean D is not much different from 0. However, the standard deviation of D is greatly increased as compared to neutrality or the standard hard sweep. Both these phenomena can easily be understood. As we have already seen in our calculations for $R = 0$ above, the (average) folded frequency spectrum after a soft sweep is very similar to the neutral spectrum. As also predicted there, the average D close to the selected site is even positive for large Θ_b (c3). The large variance is a consequence of the timing of beneficial mutation events as shown in figure 3.4. Since mutation and coalescence can occur in any order there is a wide range of possible family sizes that can escape the sweep through mutation, which can result in either a very negative D (if a single lineage escapes) or a positive D (if a larger family escapes). As for a hard sweep, recombination reduces D in the flanking regions of a soft sweep. However, in the presence of polymorphism due to lineages that escape because of recurrent mutation, this effect is much reduced.

Kelly's ZnS. Both soft and hard sweeps affect the shape of the coalescent tree of a sample and thereby the associations (LD) between neutral mutations that fall on that tree. One way to measure LD is by using Kelly's ZnS statistic (KELLY 1997), which is based on pairwise LD. Mutations that happen on the same branch in a tree cause high ZnS values. The range of values that ZnS can take is from 0 to 1, with higher values denoting stronger LD. From the plots (figure 3.5 a4–c4), it looks as if both soft and hard sweeps show about the same result: ZnS is much higher than the neutral expectation. However, in this case the plots show only part of the story. ZnS is only defined if there are 2 or more polymorphic sites. In the case of a hard sweep, many runs (about 90% directly next to the selected site) produced fewer than 2 polymorphic sites. For those runs we could therefore not calculate ZnS. The runs with more than 1 polymorphism were mostly those where a recombination event had taken place and this leads to high ZnS values. In the soft sweep simulations there were only few runs with less than 2 polymorphic sites (8% of the runs next to the selected site). After a soft sweep, ZnS is therefore also high if no recombination has taken place yet.

The number of haplotypes. The number of haplotypes in a sample K, shown in figure 3.5 a5–c5, is simply a count of the number of different sequences that are found in a sample (DEPAULIS and VEUILLE 1998). Note that the number of haplotypes here is higher than in figure 3.3, because of recombination (both between the selected and the neutral locus and within the neutral locus) and new neutral mutations. K is much lower than the neutral expectation everywhere for both hard and soft sweeps. However, close to the selected site for the hard sweep, this is mainly due to a low number of polymorphic sites S, and not because of a strong haplotype structure. For example, if $S = 1$, there can only be two haplotypes, for $S = 2$, there can be either two or three haplotypes. In these cases, K is not a very informative statistic, at least if we already have the information about S. Away from the selected site, and everywhere for the higher Θ_b values, K is low because of haplotype structure. To capture this effect, we have made an attempt to standardize the K values. Using neutral simulations, we have estimated the expectation and standard deviation of K, given a fixed number of polymorphic sites. We have defined K' (standardized K) as

$$K' = \frac{K - E(K|S)}{sd(K|S)}$$

and we define $K' = 0$ if $S < 2$. The last row of panels of figure 3.5 shows that K' is lower than expected if K is low despite relatively high S. On the other hand, K' is not different from the neutral expectation if there are very few polymorphic sites.



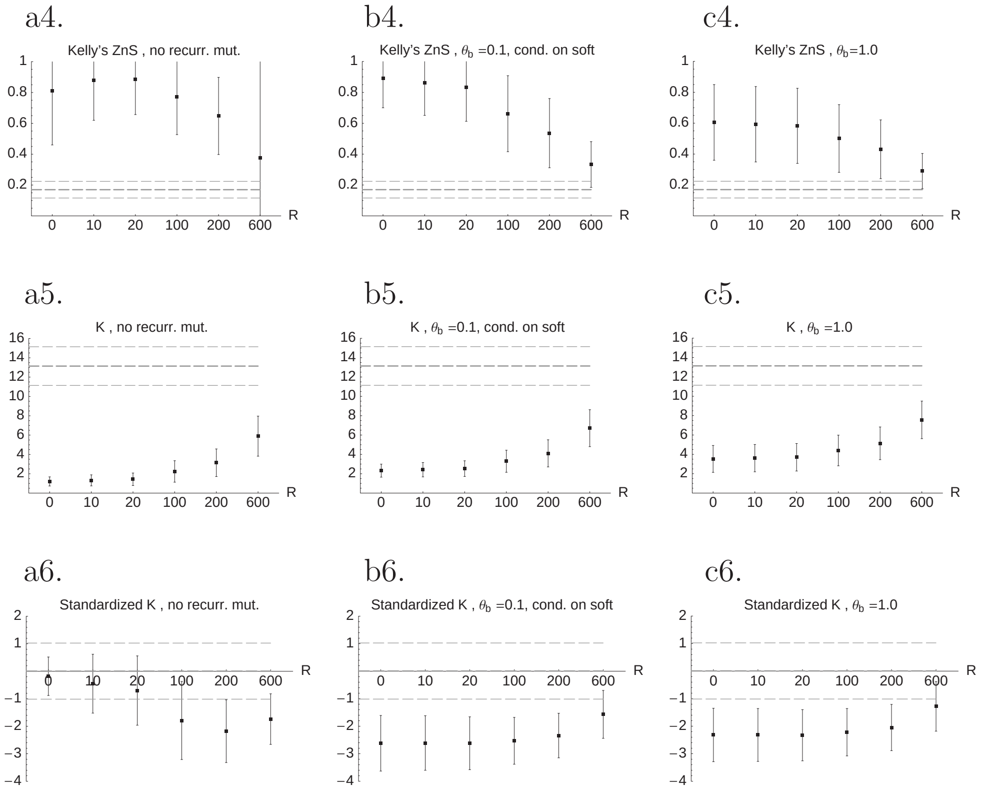


Figure 3.5: Means (± one standard deviation) of summary statistics in a sample taken at fixation of a beneficial allele. The x-axis shows the distance from the selected site in units of $R = N_e r$. The left column (a1–a6) shows hard sweeps (no recurrent mutation); the middle column (b1–b6) shows only soft sweeps for beneficial mutation rate $\Theta_b = 0.1$; and the right column (c1–c6) shows averages over all sweeps (hard or soft) for $\Theta_b = 1.0$. The statistics are from top to bottom are: (1) mean number of pairwise differences (π), (2) number of polymorphic sites (S), (3) Tajima's D, (4) Kelly's ZnS, (5) number of haplotypes K, (6) standardized K (see text). The grey lines indicate means (solid) ± one standard deviation (dashed) under neutrality. In the plots for π and S, asterisks depict predicted values based on formula 3.8 and 3.18. Parameters are as described in the Methods section.

Power analysis

Again using simulations, we have done a power analysis of two frequency spectrum based tests (Tajima's D and Fay and Wu's H) and two LD based tests (number of haplotypes K and Kelly's ZnS). For a given set of parameters, the probability is estimated that a simulation run results in a significantly positive or negative test statistic. Critical values for the tests are obtained using simulations of a neutral model without recombination (for details see the

Model and Methods section). We did simulations for six scenarios: without recurrent mutation, Θ_b values 0.1, 0.4, 1.0 and 4.0, and $\Theta_b = 0.1$ conditioned on a soft sweep. We have looked at these scenarios at seven different times after fixation of the beneficial allele: $t = 0$, 0.01, 0.05, 0.1, 0.2, 0.5, 1.0 (time is measured in N_e generations). We again looked at fragments at six recombinational distances from the selected site. The results are shown in figure 3.6 and 3.7.

Frequency spectrum based tests. We conducted Tajima's D as a two-sided test and Fay and Wu's H (Fay and Wu 2000) as a one-sided test. Results from Tajima's test are shown in figure 3.6 and 3.7; results for Fay and Wu's H are not shown, but only described below. For a classical hard sweep without recurrent mutation, frequency spectrum based tests have no power at the selected site, directly at fixation, simply because of lack of polymorphism. Tajima's D has moderately high power (up to 41%) to detect hard sweeps either at some distance ($R = 100$ or 200) from the selected site because of recombination, or at some time after fixation (t between 0.05 and 0.2) because of new mutations that have a low frequency. In figure 3.6a, the region of high power shows as a dark quarter ring. When we increase the beneficial mutation rate, and thereby allow for soft sweeps to happen, and also if we condition on soft sweeps (see figure 3.7), the power of Tajima's D goes down in the regions where the power was high before. At the same time, close to the selected site, where the power was low in the hard sweep case, the power goes up. Directly next to the selected locus, the power of D reaches 20% (see figure 3.7c). This is not surprising, even though the frequency spectrum after a soft sweep is expected to be similar to that under neutrality (see figure 3.2a). It is the large variance of D after a soft sweep (see figure 3.5) that causes these significantly negative D values. Since the mean D is not much different from 0, the large variance also causes significantly positive D values (19%), as is shown in figure 3.7d.

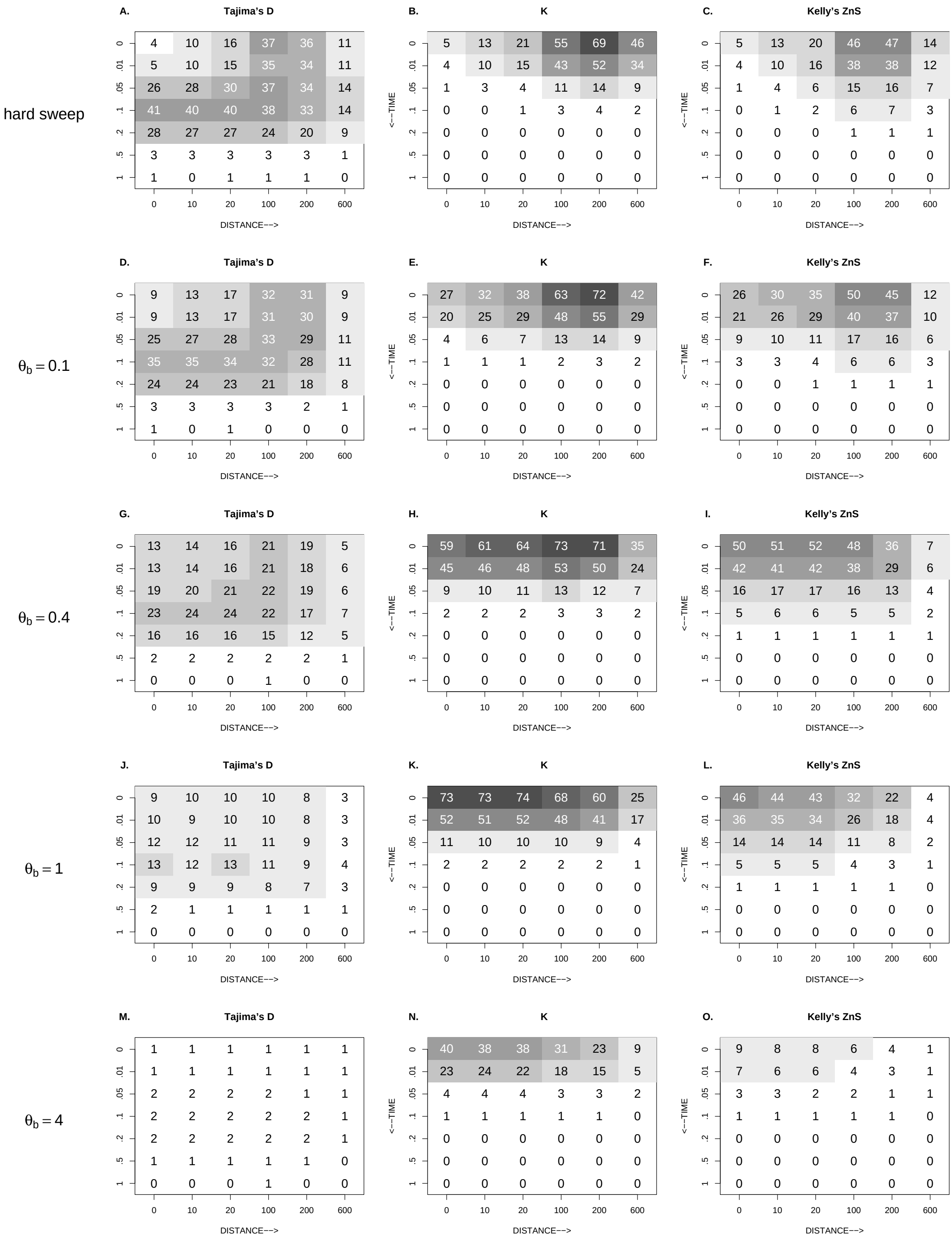
Fay and Wu's H is negative if there is an excess of high frequency derived mutations, which is expected in the flanking regions of a selected site after a hard sweep. Fay and Wu's H therefore has high power in the flanking regions (up to 63%). However, with time, the power reduces quickly, because new mutations that accumulate will have low frequencies and the high frequency variants may be lost by drift (Przeworski 2002). For higher Θ_b values, the power of H goes down in the flanking regions, but just as for Tajima's D, the power goes up (up to 34%) close to the selected site. In fact, we expect significant H values there, because the frequency spectrum close to the selected locus shows an excess of high frequency derived variants (see figure 3.2).

Linkage disequilibrium (LD) based tests. We used the K and ZnS tests as one-sided tests. We look for a lower than expected number of haplotypes (K test) or stronger than expected association between sites (ZnS test). Just like the frequency spectrum based tests, the LD tests do not have power to detect a hard sweep at the selected locus at the time of fixation, because there are no polymorphic sites. At some distance from the selected site, both LD tests have high power (up to 69% for K) to detect hard sweeps, especially at R from $100 - 600$. However, whereas Tajima's D performed best for hard sweeps, the LD tests perform better for soft sweeps. Their power increases if the beneficial mutation rate is increased, in particular close to the selected locus. This means, in particular, that recent soft selective sweeps from recurrent mutation, unlike hard sweeps, can be detected from polymorphism data (e.g. from introns) from a selected itself. Kelly's ZnS test shows roughly the same pattern as the K test. ZnS is somewhat less powerful at the time of fixation but its power lasts longer after the sweep. For both K and ZnS, it should be noted that their power reduces quickly after fixation and at $t = 0.1$ there is virtually no power left.

Effect of further parameters. We did additional simulation runs for weaker selection ($\alpha = 1,000$) and a different length of the neutral fragment ($\Theta_n = R_n$ from 2 to 40). None of these changes affects the qualitative results that we have reported above. For $\alpha = 1000$, the power of all tests are reduced by several percent, as already reported, e.g. by PRZEWORSKI (2002). Also, to compare results, the recombination distance R must be rescaled to $\approx R/10$ to account for the about 10 times longer fixation time. Importantly, however, the effect of the beneficial mutation rate and the change in the power of the tests from hard to soft sweeps stays the same.

The power of the frequency spectrum based tests generally slightly increases for longer fragments and more strongly decreases for shorter fragments, due to the larger number of polymorphic sites. For tests based on linkage disequilibrium (K and ZnS), there is a clear decrease of power in some cases for fragments with $R_n > 10 - 20$. This is expected since recombination within the fragment will reduce LD.

Power analysis, $\alpha = 10000$



hard sweep

A. Tajima's D

	0	10	20	100	200	600
0	4	10	16	37	36	11
.01	5	10	15	35	34	11
.05	26	28	30	37	34	14
.1	41	40	40	38	33	14
.2	28	27	27	24	20	9
.5	3	3	3	3	3	1
1	1	0	1	1	1	0

B. K

	0	10	20	100	200	600
0	5	13	21	55	69	46
.01	4	10	15	43	52	34
.05	1	3	4	11	14	9
.1	0	0	1	3	4	2
.2	0	0	0	0	0	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

C. Kelly's ZnS

	0	10	20	100	200	600
0	5	13	20	46	47	14
.01	4	10	16	38	38	12
.05	1	4	6	15	16	7
.1	0	1	2	6	7	3
.2	0	0	0	1	1	1
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

$\theta_b = 0.1$

D. Tajima's D

	0	10	20	100	200	600
0	9	13	17	32	31	9
.01	9	13	17	31	30	9
.05	25	27	28	33	29	11
.1	35	35	34	32	28	11
.2	24	24	23	21	18	8
.5	3	3	3	3	2	1
1	1	0	1	0	0	0

E. K

	0	10	20	100	200	600
0	27	32	38	63	72	42
.01	20	25	29	48	55	29
.05	4	6	7	13	14	9
.1	1	1	1	2	3	2
.2	0	0	0	0	0	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

F. Kelly's ZnS

	0	10	20	100	200	600
0	26	30	35	50	45	12
.01	21	26	29	40	37	10
.05	9	10	11	17	16	6
.1	3	3	4	6	6	3
.2	0	0	1	1	1	1
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

$\theta_b = 0.4$

G. Tajima's D

	0	10	20	100	200	600
0	13	14	16	21	19	5
.01	13	14	16	21	18	6
.05	19	20	21	22	19	6
.1	23	24	24	22	17	7
.2	16	16	16	15	12	5
.5	2	2	2	2	2	1
1	0	0	0	1	0	0

H. K

	0	10	20	100	200	600
0	59	61	64	73	71	35
.01	45	46	48	53	50	24
.05	9	10	11	13	12	7
.1	2	2	2	3	3	2
.2	0	0	0	0	0	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

I. Kelly's ZnS

	0	10	20	100	200	600
0	50	51	52	48	36	7
.01	42	41	42	38	29	6
.05	16	17	17	16	13	4
.1	5	6	6	5	5	2
.2	1	1	1	1	1	1
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

$\theta_b = 1$

J. Tajima's D

	0	10	20	100	200	600
0	9	10	10	10	8	3
.01	10	9	10	10	8	3
.05	12	12	11	11	9	3
.1	13	12	13	11	9	4
.2	9	9	9	8	7	3
.5	2	1	1	1	1	1
1	0	0	0	0	0	0

K. K

	0	10	20	100	200	600
0	73	73	74	68	60	25
.01	52	51	52	48	41	17
.05	11	10	10	10	9	4
.1	2	2	2	2	2	1
.2	0	0	0	0	0	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

L. Kelly's ZnS

	0	10	20	100	200	600
0	46	44	43	32	22	4
.01	36	35	34	26	18	4
.05	14	14	14	11	8	2
.1	5	5	5	4	3	1
.2	1	1	1	1	1	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

$\theta_b = 4$

M. Tajima's D

	0	10	20	100	200	600
0	1	1	1	1	1	1
.01	1	1	1	1	1	1
.05	2	2	2	2	1	1
.1	2	2	2	2	2	1
.2	2	2	2	2	2	1
.5	1	1	1	1	1	0
1	0	0	0	1	0	0

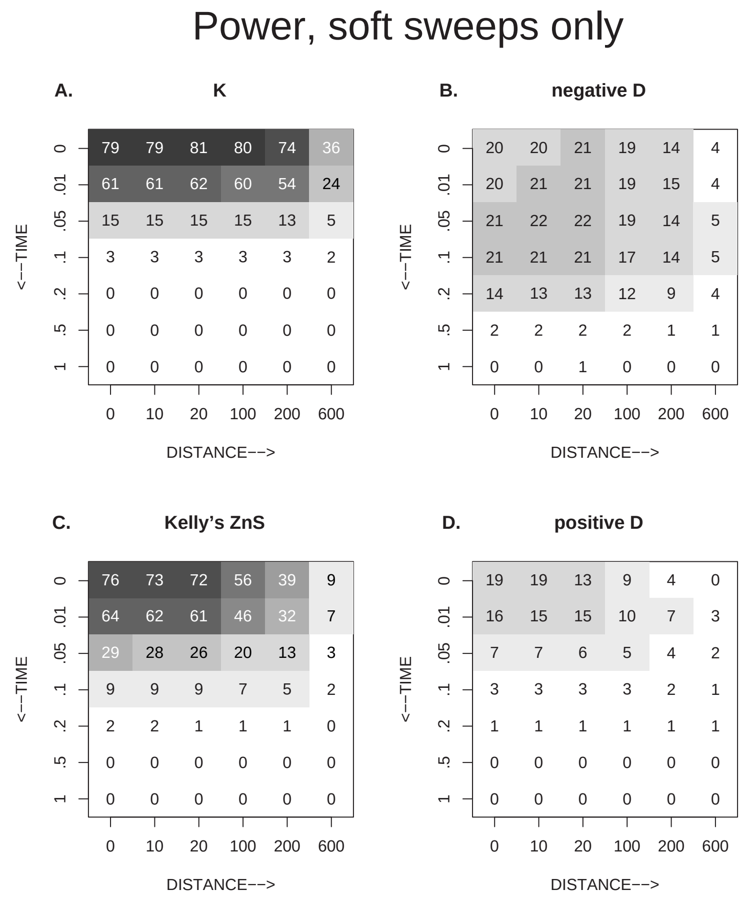
N. K

	0	10	20	100	200	600
0	40	38	38	31	23	9
.01	23	24	22	18	15	5
.05	4	4	4	3	3	2
.1	1	1	1	1	1	0
.2	0	0	0	0	0	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

O. Kelly's ZnS

	0	10	20	100	200	600
0	9	8	8	6	4	1
.01	7	6	6	4	3	1
.05	3	3	2	2	1	1
.1	1	1	1	1	1	0
.2	0	0	0	0	0	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

Figure 3.6: (last page) The percentage of simulation runs that yielded a significant test statistic depending on the value of Θ_b, other parameters as standard. The x-axis shows the distance from the selected site in units of $R = N_e r$. The y-axis shows the time since fixation of the B allele in units of N_e generations.



A. K

<--TIME	0	10	20	100	200	600
0	79	79	81	80	74	36
.01	61	61	62	60	54	24
.05	15	15	15	15	13	5
.1	3	3	3	3	3	2
.2	0	0	0	0	0	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

B. negative D

<--TIME	0	10	20	100	200	600
0	20	20	21	19	14	4
.01	20	21	21	19	15	4
.05	21	22	22	19	14	5
.1	21	21	21	17	14	5
.2	14	13	13	12	9	4
.5	2	2	2	2	1	1
1	0	0	1	0	0	0

C. Kelly's ZnS

<--TIME	0	10	20	100	200	600
0	76	73	72	56	39	9
.01	64	62	61	46	32	7
.05	29	28	26	20	13	3
.1	9	9	9	7	5	2
.2	2	2	1	1	1	0
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

D. positive D

<--TIME	0	10	20	100	200	600
0	19	19	13	9	4	0
.01	16	15	15	10	7	3
.05	7	7	6	5	4	2
.1	3	3	3	3	2	1
.2	1	1	1	1	1	1
.5	0	0	0	0	0	0
1	0	0	0	0	0	0

Figure 3.7: The percentage of simulation runs that yielded a significant test statistic if we condition on a soft sweep. $\Theta_b = 0.1$, other parameters as standard. The x-axis shows the distance from the selected site in units of $R = N_e r$. The y-axis shows the time since fixation of the B allele in units of N_e generations.

Improving the power of the LD based tests. The power of the LD based tests reduces very quickly with time. There are three reasons for this. First, ancestral variation disappears from the population through drift. Since it is ancestral variation that is in LD, tests will only detect significant deviations as long as there is sufficient ancestral variation in the sample. Second, new mutations accumulate and these mutations are not in LD, nor are they organized in clear haplotypes. Finally, recombination between the ancestral haplotypes can reduce LD and increase the number of haplotypes. In $0.1N_e$ generations, drift reduces the number of ancestral polymorphic sites by only about 15%, and it seems to be the other two factors that are most important for the reduction of power.

Note that our tests are very conservative in that they assume no recombination for the neutral simulations. If a reliable estimate of the recombination rate is available, neutral simulation with a (conservatively low) $R > 0$ can increase the power significantly (Wall and Hudson 2001). To account for the effect of new mutations, we suggest here a variant of the test that is possible in certain scenarios if data from a sister population is available.

Imagine that we are interested in local adaptations of a colony population to a new “island” habitat and that the “continental” founder population that continues to live under ancestral conditions is also known. Assume further that there is no recent gene flow between the two sister populations. We may then use data from the founder population to identify shared polymorphisms that predate the adaptation. Mutations that are only found in the island population may be new mutations and are taken out of the analysis. Glinka *et al.* (2003), for example, show that 65% of the mutations found in a European *Drosophila melanogaster* population (the “colony”) are also found in an African sister population, even though they have only a small sample from Africa. Under the assumption that there is no gene flow between the populations, we can consider mutations that are found in both populations as ancestral variation.

To see what would be the effect on the power of the different tests of using only ancestral variation, we have done simulations of positive selection where we have stopped neutral mutational input at the start of the selective phase. For neutral comparison, we stopped mutational input in the last tN_e generations of the tree. The result is promising: the power of the LD tests is much higher if we consider only ancestral variation (see figure 3.8). This is even though there are fewer mutations in the analysis (at $t = 0.1$, at the selected site, mean S is 16.4 with new mutations and only 9.6 without). The power also increases for hard sweeps in the flanking regions (results not shown). However, for Tajima’s D, the method does not work, the power stays low for soft sweeps,

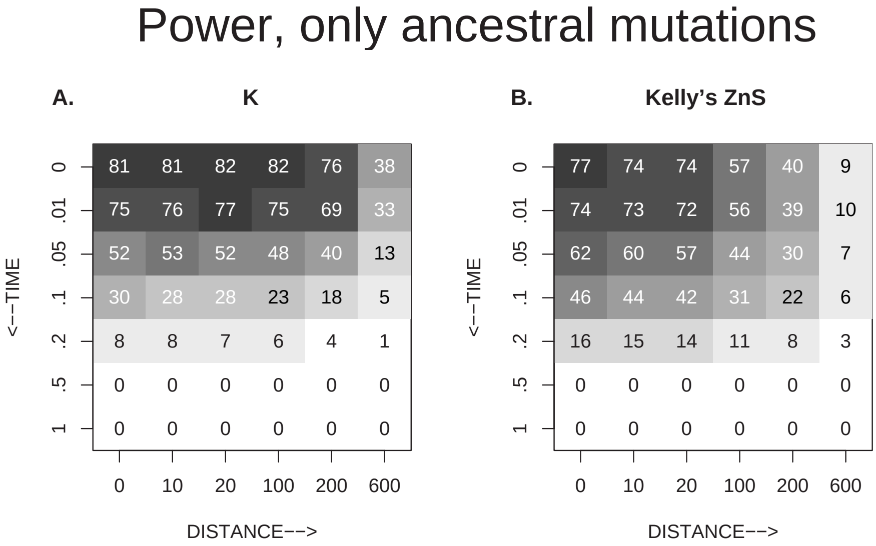


Figure 3.8: The percentage of simulation runs that yielded a significant test statistic if we condition on a soft sweep and ignore mutations during and after the sweep. $\Theta_b = 0.1$, other parameters as standard. The x-axis shows the distance from the selected site in units of $R = N_e r$. The y-axis shows the time since fixation of the B allele in units of N_e generations.

and for hard sweeps the power is higher if we allow for new mutations.

To apply this approach to data, the following steps should be taken. A not too divergent sister population is needed and an accurate estimate of the divergence time, d. To obtain critical values for the tests, neutral simulations should be done with no mutations in the last d generations. The data from the focus population should be compared with a large sample from a large sister population, so that as many mutations as possible can be identified as ancestral. If only a small sample is used, many mutations will have to be taken out of the analysis, making the tests less powerful. Similarly, power is lost if the sister population is a small or divergence time too long, such that many variants are lost due to drift.

Adaptation from recurrent migration

New beneficial alleles can enter a population also by recurrent migration, instead of mutation. In Pennings and Hermisson (2006), we have shown that the number and distribution of ancestral haplotypes directly at the selected site (at recombination distance $R = 0$) in this case is again given by the Ewens sampling formula, as in the recurrent mutation case. The mutation rate Θ_b is replaced by the number of migrants per generation M. If we assume that the adaptation in the source population is very old, such that migrants are related by a neutral coalescent, also the results on the polymorphism pattern at a tightly linked locus, as described above, carry over to the migration scenario (with Θ_n the mutation rate in the source population).

At a linked locus ($R > 0$) near the selected site, haplotypes from both populations may appear in a sample. Depending on the divergence time of the populations, these haplotypes may be much more divergent than haplotypes from a single population. As far as the LD pattern is concerned, the enhanced divergence among haplotypes leads to a clearer footprint of selection. Tests based on LD will therefore have a higher power if adaptation originates from migrants from a divergent source population. Divergence between both populations also has an effect that partly opposes the effect of the sweep. As Santiago and Caballero (2005) have shown for a sweep from a single migrational origin, heterozygosity may even be increased above the population average in the flanking regions of the selected site. The same effect will also be visible for a soft selective sweep from recurrent migration.

3.4 Discussion

Main results. The main result of our study is that soft sweeps from recurrent mutation leave a clear signature on the neutral DNA polymorphism pattern. For recent sweeps, this pattern may even be clearer than the classical signature of a hard sweep from a single new mutation. This may be surprising because (1) the variation is not as much reduced as in the hard sweep case (see figure 3.5) and (2) the folded frequency spectrum is not much different than the neutral expectation (see figure 3.2). In contrast, however, soft sweeps will typically lead to a stronger signal in LD as compared to the classical pattern. This is because a second beneficial mutant brings along with it a complete new haplotype. The presence of two (or more) independent haplotypes causes the polymorphic sites to be in complete LD.

After a recent hard sweep, polymorphism in the direct vicinity of the selected site is often almost completely erased. As a consequence, standard neutrality tests have very little power in this region. Recent positive selection can then only be detected from flanking regions of a selected gene, where ancestral polymorphism is maintained due to recombination. Positive LD, in particular, is also limited to these flanking regions and usually does not extend across the selected locus (Kim and Nielsen 2004; Stephan *et al.* 2006). In contrast, for a soft sweep form recurrent mutation, polymorphism in the shape of several ancestral haplotypes is maintained directly at the selected locus. This leads to strongly positive LD which extends to both sides of the selection center. Tests based on LD therefore have a high power over long stretches of DNA, including the selection locus. Since genes are a common selection target, and most available data are from genes, we expect that soft sweeps may indeed be easier to detect than hard sweeps.

For the classical signature of a hard sweep, Kim and Stephan (2002) and Kim and Nielsen (2004) have shown that most information is contained in the frequency spectrum. Adding LD to the analysis does not increase the power of a neutrality test much further (Kim and Nielsen 2004). We find that essentially the opposite is true for the pattern of a soft selective sweep from recurrent mutation. Soft sweeps are characterized by the LD pattern and not by the frequency spectrum. For the classical test based on the frequency spectrum, Tajima's D, we find that the mean hardly deviates from neutrality and the variance is much increased relative to both, neutrality and the classical hard sweep. The reason for the conspicuous difference to a hard sweep, where recombination leads to a negative D, lies in the timing of these events during the selective phase. While recombination typically happens later than coalescence (in a forward in time picture), and therefore produces low-frequency variants in a sample, recurrent beneficial mutation happens at the same time as coalescence. It can therefore either affect single branches (leading to a negative Tajima's D) or larger families of branches that have already coalesced (leading to positive Tajima's D values). The variance in D that results is even higher than for the case of a selective sweep from the standing genetic variation, where a similar phenomenon has been observed (Innan and Kim 2004; Przeworski *et al.* 2005). Indeed, as our figure 7 shows, we expect a significantly negative *or* positive Tajima's D, each in 20% of cases, for a recent soft sweep and data from the selected locus. Importantly, this demonstrates that significantly positive D is not incompatible with positive selection under this scenario.

The inverse roles of the frequency spectrum and the LD pattern for hard and soft selective sweeps suggest a dual approach to detect positive selection in genome scans. A standard frequency based test, such as Tajima's D, should be combined with a LD test like ZnS (given that the phase information is available), in particular if the effective population size and allelic mutation rates are likely to be large or if adaptation from recurrent migration could play a role. We note that an untypical signature of positive selection with strong positive LD across the selected site (as in the case of a soft sweep) could also result from hard sweep if there is gene conversion (see also Hamblin and Di Rienzo 2000). For this we need to assume that gene conversion happens during the selective phase and that the gene conversion tract includes the selected site.

While high levels of LD are a strong signal of a recent soft sweep from recurrent mutation, the pattern quickly fades for older sweeps due to new mutations and recombination (see figure 3.7). Here, we find that the power of LD based tests is greatly increased if new mutations can be taken out of the analysis. This is possible if polymorphism data from the same locus from a recently diverged sister population is available. One can then include only shared polymorphisms into the analysis, which effectively purges the study population of all mutations that occurred after the split. For practical use, the divergence time between the populations needs to be estimated and critical values for the test statistics need to be obtained from neutral simulations with no mutations since the divergence of the populations. The method works best if the sister population is large, if a large sample is available from the sister population, and if the divergence between the populations has occurred not too long before the start of positive selection in the study population. In this case, we obtain a high power of neutrality tests based on LD for about

$0.1 \times N_e$ generations, which is comparable to the values for Tajima's test for the classical sweep pattern (see figure 3.8).

Conditions and caveats. Throughout this study, we have assumed that the population in which we want to detect selection is panmictic with a constant size. It is well-known that population structure and demography can mimic the polymorphism patterns that are typical of positive selection. This is true for the classical sweep pattern, where population growth or bottlenecks are alternative mechanisms that can produce an excess of rare alleles. It also holds for the signature of a soft sweep from recurrent mutation. Strong positive LD can result, for example, from bottlenecks and from admixture (McVean 2002; Depaulis *et al.* 2003). Ignoring population demography can therefore lead to high rate of false positives in the neutrality tests. The general strategy to overcome this problem at least partly is to compare data from candidate loci with genome-wide data to account for demographic effects (cf. Ometto *et al.* 2005; Nielsen *et al.* 2005; Schlenke and Begun 2005; Schmid *et al.* 2005). Another scenario that is known to produce significantly positive LD is balancing selection. However, long-term balancing selection would lead to a haplotype structure where each of the haplotypes carries neutral variation. In contrast, the haplotypes after a soft sweep should contain only very little variation from new mutations, which should make it possible to distinguish these two scenarios.

One important assumption of our model is that the beneficial allele can only arise at a single locus. In some cases this may not be the case. For example, several mutations at different loci may affect the efficiency of a pathway in the same way. In the ancestral genetic background, all these mutations then have an equivalent effect on phenotype and fitness. In the presence of one of these mutations, a second mutation at a different locus may be neutral. If two of these mutations at different loci are picked up by selection and simultaneously increase in frequency, they will at some point start to interfere with each other. Fixation of the allele at one locus will stop the frequency increase at the other locus, leading to the pattern of a partial sweep.

We have also assumed that all variants of the beneficial allele have exactly the same fitness effect, which may be unrealistic. However, in Pennings and Hermisson (2006), we have looked at the effect of variable selection coefficients across the distribution of ancestral haplotypes and found that the effect is limited as long as this variation is not very strong. We therefore expect that also the results in this paper will remain robust under moderate differences in s. Similarly, we expect that all results that depend on the distribution of ancestral haplotypes due to recurrent mutation are robust to relaxations of various other model assumptions, which are all discussed in Pennings and Hermisson (2006). In particular, this holds for diploidy, frequency dependent selection or dominance, changing selection pressures and for adaptation from standing genetic variation.

Data. Patterns of soft selective sweeps from recurrent mutation have not been in the focus in genome scans for positive selection so far. Nevertheless, there are several examples in published data that are suggestive of soft sweeps. The clearest case comes from three immunity receptor genes in *Drosophila simulans* and was reported by Schlenke and Begun (2005). All three genes show extreme levels of LD due to two major haplotypes that have not recombined. In one case, there is a third haplotype at low frequency. While there are normal levels of variation among haplotypes, there is no variation within the haplotype classes, with the exception of a single singleton in one case. In accordance with our expectations for a soft sweeps from recurrent mutation, frequency spectrum based tests did not result in significant values. However, when the authors used the ZnS test, all three genes were highly significant and clear outliers relative to reference samples from other genes. The authors found that a bottleneck could not explain the high ZnS values. Since LD is maximal on the gene, but quickly decreases both upstream and downstream, the authors conclude that the gene itself has been the target of positive selection. As mentioned above, gene conversion during a hard sweep offers an alternative explanation for strongly positive LD that extends to both sides of a selection center. This seems possible in one of the genes (Tehao), where in the middle of the gene there is a stretch of 1300 bp without any polymorphism. However, no such stretch without polymorphism is visible for the other two genes. Together with the absence of a signal in the frequency spectrum this makes soft sweeps from recurrent origins the most plausible explanation.

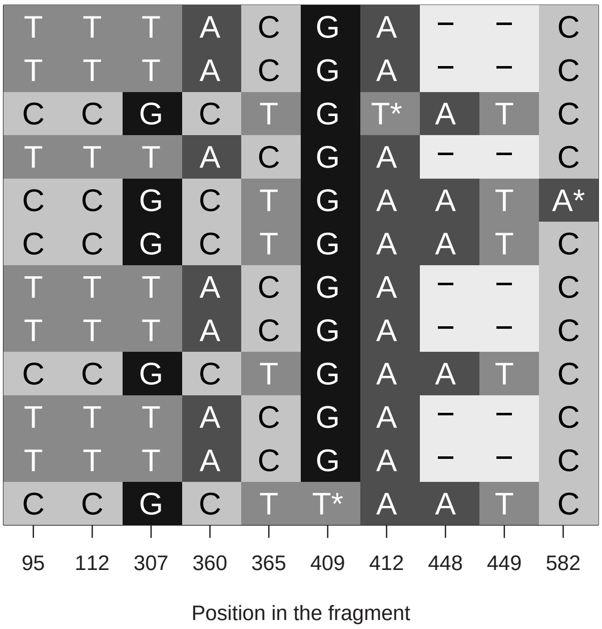
A second example is the Duffy locus in humans. The FY-0 allele at this locus confers resistance against malaria and is found at near fixation in sub-Saharan African populations, but is very rare everywhere else (Hamblin and Di Rienzo 2000). Also the responsible mutation is known. This mutation is found on two different haplotypes, which are characterized by a SNP and an indel on the 5' side of the beneficial mutation and a SNP on the 3' side. Because the haplotypes are characterized by few SNPs and because there are some singletons in the region as well, no test statistic is significant for this locus. However, other data, such as a very high F_{ST} value, strongly support the hypothesis that the FY-0 allele rose to fixation because of selection. This, combined with the two haplotypes that are seen, makes a soft sweep a plausible explanation, although a hard sweep with a gene conversion is an alternative scenario. In Hamblin *et al.* (2002), evidence was found for a hard sweep associated with the FY-0 allele in the Hausa population. However, this population was chosen for this study because it had only one of the two haplotypes.

As an illustration of the method that we suggest, we present data from a fragment on the X-chromosome from a European and an African sample of *Drosophila melanogaster* (figure 3.9). This fragment (fragment 163 from Ometto *et al.* 2005) has 9 polymorphic sites in the European fragment, and neither frequency spectrum based tests, nor LD tests show a deviation from neutrality. However, the 6 polymorphisms that are shared between Europe and Africa, are in perfect linkage disequilibrium in the European sample. When only considering the shared polymorphisms, there are two perfect haplotypes of which one is found 5 times and one is found 7 times in the sample. Both the ZnS test and the K test show significant deviation from neutrality.

LD or haplotype structure is used by many studies to find alleles that have recently increased in frequency. As long as the allele has not reached fixation, the region around the locus will show strong LD (Stephan *et al.* 2006). Sabeti *et al.* (2002) developed a method to use this pattern of strong LD to identify local or partial sweeps. A modified version of the Sabeti method was applied to HapMap data by Voight *et al.* (2006) to identify partial sweeps. Complete hard sweeps cannot be detected by this method, but with a slightly altered version of this method it should be possible to use HapMap data to detect soft sweeps.

3.5 Acknowledgments

We thank Yuseob Kim for help with his software, Andreas Lehnert and Andreas Gros for help with the figures and Haipeng Li, John Parsch, Peter Pfaffelhuber and Saskia Stehouwer for comments on the manuscript. This work was supported by an Emmy Noether grant by the Deutsche Forschungsgemeinschaft (DFG) to J.H.



95	112	307	360	365	409	412	448	449	582
T	T	T	A	C	G	A	−	−	C
T	T	T	A	C	G	A	−	−	C
C	C	G	C	T	G	T*	A	T	C
T	T	T	A	C	G	A	−	−	C
C	C	G	C	T	G	A	A	T	A*
C	C	G	C	T	G	A	A	T	C
T	T	T	A	C	G	A	−	−	C
T	T	T	A	C	G	A	−	−	C
C	C	G	C	T	G	A	A	T	C
T	T	T	A	C	G	A	−	−	C
T	T	T	A	C	G	A	−	−	C
C	C	G	C	T	T*	A	A	T	C

Figure 3.9: Polymorphic sites in a fragment on the X-chromosome of sample from *Drosophila melanogaster* in a sample from Europe. The polymorphic sites that are unique to the European sample are indicated by an asterisk. The indel of 2 bp is counted as one polymorphic site.

3.6 Appendix

Frequency distribution of ancestral variation

In this section, we derive the frequency distribution of ancestral neutral polymorphisms at a tightly linked neutral locus after a soft selective sweep from recurrent mutation. This means, we assume that no recombination during the selective phase has happened between the selected site and the locus studied. We focus on the contribution of ancestral variation to the frequency spectrum and thus ignore new mutations (neutral mutations that have occurred after the start of the selective phase).

Assume that we take a sample of size n directly (or sufficiently soon) after fixation of a beneficial allele that enters the population with a mutation parameter $\Theta_b = 2uN_e$. In Pennings and Hermisson (2006), we have shown that the distribution of ancestral haplotypes in such a sample follows the Ewens sampling formula. For the frequency spectrum of ancestral polymorphisms, we need to combine this result with a neutral coalescence process of the surviving ancestral haplotypes for the time prior to the selective phase. We need the following ingredients for a derivation:

First, according to the Ewens sampling formula, the probability for k ancestral haplotypes in the sample is

$$\Pr(k|n, \Theta_b) = \frac{\Theta_b^k}{\Theta_{b(n)}} S_n^{(k)} \tag{3.9}$$

where we define $\Theta_{b(m)} := \prod_{i=0}^{m-1}(\Theta_b + i)$ and $S_n^{(k)}$ is the nonnegative Stirling number of first kind

$$S_n^{(k)} = \sum_{n_1+\cdots+n_k=n} \frac{n!}{k! n_1 \cdots n_k} \tag{3.10}$$

which counts the number of permutations of n objects with k permutation cycles ($S_n^{(n)} = 1$; $S_n^{(k)} = 0$ for $k > n$). Since there are no ancestral polymorphisms if there is only a single ancestral haplotype, $k = 1$, we need to condition on $k > 1$,

$$\Pr(k|n, \Theta_b, k > 1) = \frac{\Pr(k|n, \Theta_b)}{1 - \Pr(1|n, \Theta_b)} = \frac{\Theta_b^k}{\Theta_{b(n)} - \Theta_b(n-1)!} S_n^{(k)}. \tag{3.11}$$

Second, the probability that the derived variant appears in j out of k haplotypes is

$$p(j|k) = \frac{1}{j a_k} \quad ; \quad a_k := \sum_{i=1}^{k-1} \frac{1}{i}. \tag{3.12}$$

given that the population is in neutral equilibrium. If this is not the case, an empirical frequency spectrum, estimated from genomewide data can be used instead (as in NIELSEN *et al.* 2005). And third, again according to the Ewens sampling formula, the probability for a haplotype distribution of $\{n_1, \ldots, n_k\}$, given that k haplotypes are found in a sample of size n is

$$\Pr(n_1, \ldots n_k|k, n) = \frac{n!}{k! n_1 \cdots n_k S_n^{(k)}} \tag{3.13}$$

Assume now that j out of k haplotypes carry the derived mutation. The probability that ℓ individuals out of n carry the derived mutation under this condition then gets

$$\Pr(\ell|j,k,n) = \sum_{\substack{n_1+\dots+n_j=\ell \\ n_{j+1}+\dots+n_k=n-\ell}} \Pr(n_1,\dots n_k|k,n) = \frac{\binom{n}{\ell}}{\binom{k}{j}} \cdot \frac{S_\ell^{(j)} S_{n-\ell}^{(k-j)}}{S_n^{(k)}} \tag{3.14}$$

We can now combine all these components to obtain the ancestral polymorphism spectrum as

$$\begin{aligned} P_{\text{anc}}[\ell|n] = \sum_{k=2}^{n} \Pr(k|n,\Theta_b,k>1) \sum_{j=1}^{k-1} p(j|k)\Pr(\ell|j,k,n) = \\ \sum_{k=2}^{n} \frac{\Theta_b^k}{\Theta_{b(n)}-(n-1)!} \sum_{j=1}^{k-1} \frac{\binom{n}{\ell}}{j a_k \binom{k}{j}} \cdot S_\ell^{(j)} S_{n-\ell}^{(k-j)}. \end{aligned} \tag{3.15}$$

where $\ell + k - n \leq j \leq \ell$. Conditioned on a soft sweep with k haplotypes we obtain:

$$P_{\text{anc}}[\ell|k,n] = \frac{\binom{n}{\ell}}{a_k S_n^{(k)}} \sum_{j=1}^{k-1} \frac{S_\ell^{(j)} S_{n-\ell}^{(k-j)}}{j\binom{k}{j}} \tag{3.16}$$

An interesting consequence is that the ratio of singletons to $(n-1)$-letons is $(k-1)$ to 1. So, if $k = 2$ the frequency spectrum is symmetrical.

Distribution of distinct ancestral haplotypes

Ancestral haplotypes are not necessarily distinct since they might be identical by descent. For the probability to obtain ℓ distinct ancestral haplotypes, given that there are k ancestral haplotypes, we need to follow these haplotypes in a neutral coalescent process with mutations prior to the selective phase. The number (and distribution) of distinct haplotypes is then again given by the Ewens sampling formula, this time on a sample of size k and with the neutral mutation rate Θ_n on the fragment, i.e. by $\Pr(\ell|k,\Theta_n)$ using equation (3.9). For the entire probability to obtain ℓ distinct ancestral haplotypes, we thus need to combine two Ewens sampling steps to obtain

$$\Pr[\ell|n,\Theta_b,\Theta_n] = \sum_{k=\ell}^{n} \Pr(\ell|k,\Theta_n)\Pr(k|n,\Theta_b) = \sum_{k=\ell}^{n} \frac{\Theta_n^{\ell-1}\Theta_b^{k-1} S_k^{(\ell)} S_n^{(k)}}{(\Theta_n+k-1)!(\Theta_b+n-1)!}. \tag{3.17}$$

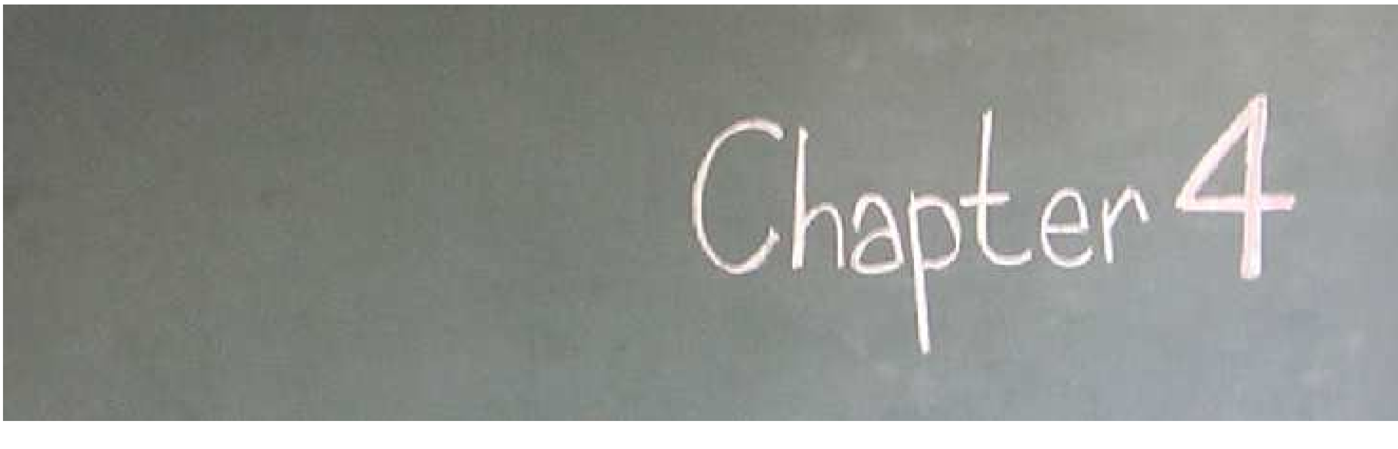
The expected number of polymorphic sites

We assume that lineages escape independently by recombination. Using equation (3.7), we thus obtain the probability that q lineages escape through recombination as a binomial

$$P_{reco}(q|n) = \binom{n}{q}(1 - \exp(-R\frac{2\log[\alpha]}{\alpha}))^q(\exp(-R\frac{2\log[\alpha]}{\alpha}))^{n-q}.$$

The probability that there are k ancestors for the $n - q$ lineages that have not escaped through recombination is given by $\Pr(k|n - q)$ (equation 3.9). The probability that there are m independent haplotypes in total is therefore given by

$$\Pr(m|n) = \sum_{q=0}^{m-1} P_{reco}(q|n) \cdot \Pr(m - q|n - q, \Theta_b) =$$
$$\sum_{q=0}^{m-1} \binom{n}{q}(1 - \exp(-R\frac{2\log[\alpha]}{\alpha}))^q(\exp(-R\frac{2\log[\alpha]}{\alpha}))^{n-q} \cdot \frac{\Theta_b^{m-q}}{\Theta_{b(n-q)}}\, S_{n-q}^{(m-q)} \tag{3.18}$$



Chapter 4

A one-locus model for sympatric speciation

Pleuni Pennings, Michael Kopp, Ulf Dieckmann and Joachim Hermisson

Manuscript in preparation

Models of competitive sympatric speciation have created much excitement, but they are also highly controversial. We present a thorough and largely analytical analysis of the evolution of assortative mating in a Roughgarden model, in which the ecological trait is determined by a single diallelic locus. The genetic architecture of this trait is given by a single parameter: the allelic effect x. A second parameter, σ_c, determines the individual niche width (or frequency-dependence of competition). Females are choosy and prefer mates with similar ecological phenotype. The degree of choosiness is determined by one locus with a continuum of alleles. We describe five possible regimes for the evolution of choosiness. In only one of them can complete reproductive isolation evolve from random mating in small mutational steps. In addition, we determine the regions where the ecological polymorphism is unstable, locally stable or globally stable. Our simple model allows us to investigate the roles of natural and sexual selection in sympatric speciation. We find that complete isolation may fail to evolve when natural selection favors heterozygotes, when sexual selection favors heterozygotes, or when sexual selection causes the ecological polymorphism to be unstable. Our findings are confirmed and extended by individual-based simulations.

4.1 Introduction

Interest in sympatric speciation has strongly increased in recent years. Empiricists have uncovered several likely examples of this mode of speciation in nature (Schliewen *et al.* 1994; Gíslason *et al.* 1999; Barluega *et al.* 2006; Savolainen *et al.* 2006). At the same time, theoreticians have made substantial progress in understanding the potential mechanisms leading to sympatric lineage splitting. One of these mechanisms is intraspecific competition. The idea of competitive speciation (Rosenzweig 1978) actually goes back to Darwin and has recently been studied in a series of models (e.g., Doebeli 1996; Dieckmann and Doebeli 1999; Kondrashov and Kondrashov 1999; Matessi *et al.* 2001; Bürger and Schneider 2006a). For example, Dieckmann and Doebeli (1999) used individual-based simulations of a competition model by Roughgarden (1972) to demonstrate that frequency-dependent disruptive selection on an ecological trait (i.e., a trait affecting resource competition) can promote the evolution of assortative mating (in a process similar to reinforcement). Strong enough assortative mating can lead to reproductive isolation and speciation.

The fact that competition leads to disruptive selection is not controversial. What is controversial, however, is under exactly what circumstances disruptive selection can lead to strong assortative mating. For example, it is unclear how much of the results from Dieckmann and Doebeli (1999) depend on the choice of initial conditions, parameter values, and the precise design of the simulations, and this question has lead to intense debate (Doebeli and Dieckmann 2005; Doebeli *et al.* 2005; Gavrilets 2005; Polechová and Barton 2005; Waxman and Gavrilets 2005). One reason for the continuing disagreement among evolutionary biologists is the complex nature of the Roughgarden-Dieckmann-Doebeli model, in which populations are subject to a variety of selective forces (stabilizing selection, frequency-dependent selection due to competition, sexual selection due to assortative mating) acting on a complex genetic architecture (multiple loci for both ecological and mating traits).

For this reason, several authors have attempted to gain a better understanding of sympatric speciation by studying simplified models that are more amenable to analytical or systematic treatment. Authors have used approximated fitness functions (Matessi *et al.* 2001; Bürger and Schneider 2006a), a simplified genetic architecture (Matessi *et al.* 2001; Schneider 2005), or a constant (non-evolving) level of assortativeness (Gourbiere 2004; Kirkpatrick and Nuismer 2004; Schneider 2005; Bürger and Schneider

2006a). For example, MATESSI *et al.* (2001) used a weak-selection approximation of the Roughgarden model and assumed that the ecological trait is determined by a single locus with two alleles. This approach allowed them to analytically show that evolution of assortative mating may come to a halt before inducing complete reproductive isolation.

In this paper, we use an approach similar to the one by MATESSI *et al.* (2001), but we do not make any assumptions about the strength of selection (i.e., our fitness function is not an approximation). Our model has two fixed parameters (the individual niche width, which determines the degree to which competition is frequency-dependent, and the allelic effect of the ecological locus) and one evolvable trait that determines female choosiness. We develop a simple invasion criterion, which enables us to study the behavior of the model in the entire parameter space. In addition, by comparing versions of the model with and without sexual selection, we can clarify the roles of sexual and natural selection on the evolution of assortative mating. Our simplified model shows surprisingly complex behaviour. We describe five qualitatively different regimes for the evolution of assortative mating (including two regimes previously described by MATESSI *et al.* 2001). If the population starts at random mating and mutational steps are small, complete isolation can evolve in only one these regime. In the other regimes, choosiness either does not evolve at all, or it stops at an intermediate level. In addition, there is a range of parameters for which the polymorphism at the ecological locus may be lost. Our results are confirmed and extended by individual-based simulations.

4.2 The model

Our model is a single-locus version of the speciation model by DIECKMANN and DOEBELI (1999), which in turn is based on ROUGHGARDEN's (1972) model of intraspecific competition.

Resource distribution, competition, and dynamics in an asexual population

To describe our assumptions about competition and population regulation it is useful to first focus on the dynamics of a polymorphic asexual population (i.e., a population composed of various clones differing in phenotype). Individuals are characterized by a quantitative trait, which we will refer to as the ecological trait. Let the number of individuals with phenotype X be denoted by $N(X)$.

We assume that the carrying capacity for individuals with phenotype X is a Gaussian function of X with mean 0 and variance σ_k^2:

$$K(X) = K_0 \exp\left(-\frac{X^2}{2\sigma_k^2}\right). \tag{4.1}$$

The scaling parameter K_0 is the carrying capacity for individuals with phenotype 0.

Competition between a pair of individuals with phenotypes X and Y depends on their phenotypic distance and is again assumed to be a Gaussian function with variance σ_c^2, that is

$$\alpha(X,Y) = \exp\left(-\frac{(X-Y)^2}{2\sigma_c^2}\right). \tag{4.2}$$

σ_c determines how slowly competition decreases with phenotypic distance. The total amount of competition experienced by an individual with phenotype X is

$$A(X) = \sum_Y \alpha(X,Y)N(Y). \tag{4.3}$$

$A(X)$ can be seen as the "ecologically effective population size" experienced by an individual with phenotype X.

We assume that generations are overlapping (i.e. time is continuous) and population sizes are large enough to ignore stochastic processes such as genetic drift. The dynamics of a subpopulation with phenotype X are described by

$$\dot{N}(X) = rN(X)\left(1 - \frac{A(X)}{K(X)}\right) \tag{4.4}$$

where r is the intrinsic population growth rate.

The above model is commonly interpreted in terms of competition among phenotypically variable consumers for an equally variable resource. The canonical example is birds with different beak sizes specializing on differently sized seeds. Then, $K(X)$ is the (initial) distribution of resources favored by consumers with phenotype X, and $\alpha(X-Y)$ describes the overlap in resource use between two individuals. σ_c determines the range of resources used by a single individual, that is the "individual niche width". If $\sigma_c < \sigma_k$, the phenotype $X = 0$ is an "evolutionary branching point" (Geritz *et al.* 1998), where the trait is under frequency-dependent disruptive selection.

Sexual reproduction and mate choice

We now describe the dynamics of a population with sexual reproduction. We assume a fixed sex ratio of 1/2. Mate choice is affected by phenotypic similarity with respect to the ecological trait. The ecological trait, therefore, acts as a "magic trait" (*sensu* GAVRILETS 2004), an assumption that is thought to be most conducive for sympatric speciation. The probability that an encounter between a male and a female with phenotypes X and Y leads to mating is proportional to

$$\mu(X,Y) = \exp\left(-\frac{(X-Y)^2}{2\sigma_m^2}\right), \tag{4.5}$$

where σ_m measures how slowly the mating probability declines with phenotypic distance. $\sigma_m \to \infty$ means random mating, whereas $\sigma_m = 0$ corresponds to completely assortative mating.

In order to describe different types of assortative mating, we introduce an additional factor $Q(X)$, which determines female mating activity. The idea is that females with different phenotype have different encounter rates with males. Encounters then result in mating with probability $\mu(X,Y)$). The activity factor $Q(X)$ does not depend on the mating pair, but on the female phenotype alone. We then obtain separate mating rates for females, $\phi_f(X)$, and males, $\phi_m(X)$, with phenotype X as

$$\phi_f(X) = \sum_Y N(Y)\mu(X,Y)Q(X) \tag{4.6a}$$

$$\phi_m(X) = \sum_Y N(Y)\mu(X,Y)Q(Y) \tag{4.6b}$$

and an effective mating rate of all X phenotypes

$$\phi(X) = \frac{1}{2}\Big(\phi_f(X) + \phi_m(X)\Big) = \sum_Y N(Y)\mu(X,Y)\frac{Q(X)+Q(Y)}{2} \tag{4.6c}$$

Note that MATESSI *et al.* (2001) use the term mating rate in a different sense.

Different choices for the female mating activity $Q(X)$ lead to different kinds of models. We focus on two possibilities. In our first model, $Q(X)$ is chosen such that all (ecological) phenotypes mate and reproduce at an identical rate (independent of their degree of choosiness). This is achieved by setting $\phi(X) = 1$ and solving the linear equation system (see Appendix 1). Note that male and female mating rates for a given phenotype can be different. In this model, assortative mating does not lead to sexual selection: all phenotypes contribute to the offspring pool according to their relative frequency in the population. There is also no cost of choosiness.

Our second model follows DIECKMANN and DOEBELI (1999). Here, $Q(X)$ is chosen such that the female mating rates are normalized, $\phi_f(X) = 1$ (see Appendix 2.1). This simply means that each female will look for a mate until she finds one that she does not reject. As a consequence, females do not pay a cost for being choosy. In contrast to females, however, males are subject to sexual selection imposed by female choosiness. This mode of sexual selection favors frequent phenotypes over rare phenotypes.

The birth rate for individuals with phenotype X is

$$\begin{aligned} B(X) &= r \sum_{Y,Z} N(Y)N(Z)\mu(Y,Z)\frac{Q(Y)+Q(Z)}{2} R_{YZ \to X} \\ &= r \sum_{Y,Z} N(Y)N(Z)\mu(Y,Z)Q(Z)R_{YZ \to X} \end{aligned} \tag{4.7}$$

where $R_{YZ \to X} = R_{ZY \to X}$ is the probability that a mating between individuals with phenotypes Y and Z produces offspring with phenotype X. The per capita death rate for phenotype X is

$$d(X) = r\frac{A(X)}{K(X)} \tag{4.8}$$

The dynamics of phenotype frequencies are given by

$$\dot{N}(X) = B(X) - N(X)d(X). \tag{4.9}$$

In contrast to the dynamics, phenotype fitnesses do not depend on the rate at which individuals are born, $B(X)$, but rather on the rate at which they give birth, which equals $r\phi(X)$. The Malthusian fitness of phenotype X is given by

$$W(X) = r\phi(X) - d(X). \tag{4.10}$$

In the rest of this paper, we will assume that the intrinsic population growth rate $r = 1$.

According to equation (4.10), fitness has two components, one related to mating success and one to survival. We will refer to these two fitness components as relating to sexual and natural selection, respectively. As we will show below, the interplay of these two selection components can explain many aspects of the model behavior.

The one-locus, two-allele case with constant choosiness

In the following, we assume that the ecological trait is determined by a single diploid locus with alleles '+' and '−'. Individuals with genotype +/+ have phenotype x, individuals with genotype +/− have phenotype 0, and individuals with genotype −/− have phenotype $-x$. Furthermore, we assume that the population is monomorphic with respect to the σ_m (i.e., all females have the same degree of choosiness). Therefore, at this point, we do not need to make specific assumptions about the genetics of σ_m.

As there are only three ecological phenotypes, we can use a simplified notation. We will denote the numbers of individuals carrying these genotypes by N^+_{hom}, N_{het}, and N^-_{hom}, respectively, where 'hom' and 'het' stand for homozygotes and heterozygotes. In symmetric cases, we will drop the upper indices '+' and '−' and simply write $N^+_{\text{hom}} = N^-_{\text{hom}} \equiv N_{\text{hom}}$. Analogous indices will be used for the other parameters. Furthermore, in the one-locus case, stabilizing selection can be described by the single variable

$$k \equiv \frac{K_{\text{hom}}}{K_{\text{het}}} = \frac{K(x)}{K_0}, \tag{4.11a}$$

and for the competition and mating functions, we only need two values each:

$$\begin{aligned} a &\equiv \alpha(-x, 0) = \alpha(0, x), &(4.11b)\\ a' &\equiv \alpha(-x, x), &(4.11c)\\ m &\equiv \mu(-x, 0) = \mu(0, x), &(4.11d)\\ m' &\equiv \mu(-x, x). &(4.11e) \end{aligned}$$

With the Gaussian functions (4.2) and (4.5), $a' = a^4$ and $m' = m^4$ (although, in general, other choices are possible).

In the following, we will frequently discuss our results in terms of the parameters k, a, and m (instead of x, σ_c, and σ_m). In particular, we will use the notion that $1-k$ measures the strength of stabilizing selection (i.e., the reduction in resource availability for homozygotes), $1-a$ measures the strength of negative frequency-dependent selection (the reduction in competition between homozygotes and heterozygotes), and $1-m$ measures the degree of assortativeness or female choosiness (e.g., the probability of a heterozygous female to reject a homozygous male), which in turn determines the strength of sexual selection on males (in the model with sexual selection). Note also that all three parameters depend on the allelic effect x.

Finally, if all females have the same level of choosiness, there are three possible equilibria regarding the genotypes at the ecological locus: two monomorphic equilibria ($N^+_{\text{hom}} = K(x)$, $N_{\text{het}} = N^-_{\text{hom}} = 0$ or $N^+_{\text{hom}} = N_{\text{het}} = 0$, $N^-_{\text{hom}} =$

$K(x)$) and one symmetric polymorphic equilibrium with $N^+_{\text{hom}} = N^-_{\text{hom}} = N_{\text{hom}}$. To characterize the polymorphic equilibrium we define

$$n \equiv \frac{N_{\text{het}}}{N_{\text{hom}}}, \tag{4.12}$$

where n can be between 0 (if speciation is completed) and 2 (at Hardy-Weinberg equilibrium).

Invasion analysis for female choosiness

Our principal interest is in the evolution of female choosiness. We will frame our discussion in terms of the parameter m (the probability of a heterozygous female to accept a homozygous male), which ranges from 0 to 1. $m = 1$ ($\sigma_m \to \infty$) corresponds to random mating and $m = 0$ ($\sigma_m \to 0$) to complete isolation. Our approach will be to focus on the invasion fitness of rare mutants. Assume that the population is monomorphic for m (and the ecological locus is at a polymorphic equilibrium, see above). Can a rare mutant with a different choosiness invade? The basic idea is as follows: As long as the mutant is sufficiently rare its mating strategy has no direct influence on its fitness. In the model without sexual selection, this is because the mating rate ϕ is identical for all phenotypes. In the model with sexual selection, the mating chances of all females are identical, and the mating chances of males (ϕ_m) depend only on the choosiness of the resident females. However, the mating strategy of the mutant females determines the distribution of mutant genotypes in the next generation. In particular, females with a lower m than the residents will have proportionally more homozygous offspring (with respect to the ecological genotype), and females with a higher m will have more heterozygous offspring. Therefore, a mutant with low m will be able to invade the population if and only if (at the polymorphic equilibrium of the resident population) $W_{\text{hom}} > W_{\text{het}}$, whereas a mutant with high m will be able to invade if $W_{\text{hom}} < W_{\text{het}}$. An evolutionary equilibrium is reached if $W_{\text{hom}} = W_{\text{het}} = 0$.

Simulations

So far, we have assumed that genetic drift is absent. To check whether the results of our model also hold in a finite population with drift we carried out simulations using an individual-based model. The model is very similar to the one used by Dieckmann and Doebeli (1999). Individuals are diploid hermaphrodites and have two traits: an ecological trait and a choosiness trait.

The ecological trait is determined by one diallelic additive locus. The choosiness trait is determined by one locus with a continuum of alleles. The total population size depends mainly on K_0, which is set to 500. Simulations start with a population of 100 individuals that are assigned random alleles for the ecological trait (which leads to Hardy-Weinberg proportions) and without variation for the choosiness trait. The population is then allowed to grow to an ecological equilibrium, before mutation for the choosiness trait starts. Mutational effects are defined on a scale of m (rather than σ_m). Mutation rate u and mutational stepsize are varied. Stepsizes are fixed or drawn from a (truncated) normal distribution. There is no mutation at the ecological locus. The simulations use continuous time. $2N$ events are considered the equivalent of one generation. An event can be a either a birth event or a death event. For each event, first an individual is chosen. If the event is a birth event, the chosen individual functions as a female. A mating partner is chosen depending on the ecological phenotype and the choosiness of the female and the distribution of ecological phenotypes in the population. Simulations are stopped after 100.000 generations or if the ecological trait is no longer polymorphic. C++ code is available on request.

4.3 Results

Evolution of female choosiness in the model without sexual selection

In the model without sexual selection (i.e., with $\phi(X) = 1$), the evolution of female choosiness is determined by natural selection alone. As shown in Appendix 1, for each parameter combination, natural selection favors a unique value $\hat{n}$ of the heterozygote-to-homozygote ratio n (eq. A5). (This is also the ratio that would be reached in an asexual population of three competing clones with phenotypes $-x$, 0, and x.) Female choosiness then evolves in such a way that $n = \hat{n}$, within the constraints that n must be between 0 and 2 (because we do not allow for dis-assortative mating). Depending on the degree of choosiness, this leads to three qualitatively different evolutionary regimes.

The random mating regime The population evolves toward random mating ($m = 1$) if $\hat{n} \geq 2$, that is if

$$k \leq \frac{1 + 2a + a'}{2 + 2a} \tag{4.13}$$

(red line in Figure 4.1). This is the case whenever either stabilizing selection is strong (k small, x large) or competition is weakly frequency-dependent (a large, σ_c large). In both cases, homozygous individuals do not gain a lot from being away from the optimum. Random mating is stable for any value of a if $k < 0.5$, that is if the resource density for homozygotes is less than half of that for heterozygotes.

The complete isolation regime The population evolves towards complete reproductive isolation (i.e., sympatric speciation) if $\hat{n} \leq 0$, that is if

$$k \geq \frac{1 + a'}{2a} \tag{4.14}$$

(blue line in Figure 4.1). This is the case if stabilizing selection is weak (k large, x small) and competition is moderately frequency-dependent (a and σ_c intermediate). If competition is too weakly frequency-dependent (i.e., everybody competes with everybody else) heterozygotes cannot be completely suppressed by the homozygotes. If competition is strongly frequency-dependent, competition between different phenotypes is weak, such that intermediate phenotypes can coexist with the extreme ones. Each phenotype can then be said to occupy its own ecological niche.

The partial isolation regime If neither condition (4.13) nor condition (4.14) is fulfilled, the population will evolve toward intermediate level of choosiness, leading to partial isolation. The resulting phenotypic distribution is bimodal (i.e., $n < 1$) if $k > (1 + a + a')(1 + 2a)$ (dotted line in Figure 4.1).

Evolution of female choosiness in the model with sexual selection against rare males

Natural versus sexual selection

In the model with sexual selection (i.e., $\phi_f(X) = 1$), all females have equal mating success, but rare males might be at a disadvantage. Therefore, evolution of female choosiness is determined by both natural and sexual selection. It is instructive to first consider the effects of sexual selection alone. Sexual selection is determined by $\phi_{\mathrm{m,het}}$, the mating success of heterozygous males, which can be written as

$$\phi_{\mathrm{m,het}} = \frac{2m}{1 + mn + m'} + \frac{n}{2m + n} \tag{4.15}$$

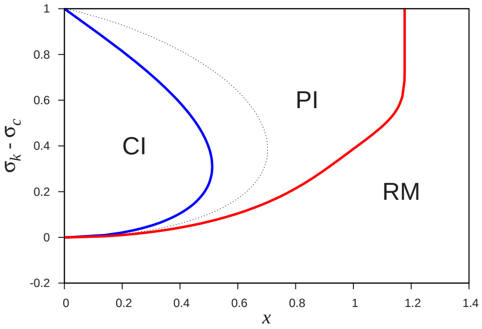


Figure 4.1: Evolutionary regimes for the evolution of female choosiness in the model without sexual selection: complete isolation (CI), partial isolation (PI), and random mating (RM). At the boundaries of the partial isolation regime, the heterozygote to homozygote ratio n (as defined in eq. A5) equals 0 (blue line) or 2 (red line). The dotted line marks parameter combinations where $n = 1$. Therefore, to the left of this line, the phenotype distribution is bimodal. Note that the axis are oriented in a way such that a high value corresponds to strong selection. σ_k is assumed to be equal to 1.

Sexual selection favors heterozygotes if $\phi_{\text{m,het}} > 1$, which is the case if $n > 1-m-m^2-m^3$. Thus, heterozygote males get more matings than homozygote males whenever their frequency exceeds a threshold, which decreases with m. The heterozygotes are always favoured by sexual selection if they are more common than either of the homozygotes ($n > 1$) or if $m > 0.54$. If $m < 0.54$, heterozygotes may get more matings even if they are less common than either of the homozygotes (i.e. the threshold $n < 1$). This is because they are in the middle of the phenotypic distribution and have a chance to mate with females from both homozygote classes.

For understanding the behavior of the model with sexual selection, it is essential to recognize that natural selection is negatively frequency-dependent (rare phenotypes experience less competition), whereas sexual selection is positively frequency-dependent (males with common phenotype are more likely to find a mating partner). Thus, when heterozygotes are rare, they *may be* favored by natural selection but disfavored by sexual selection, and when they are common, they are favored by sexual selection but may be disfavored by natural selection.

Stability of evolutionary equilibria

As in the model without sexual selection, evolutionary equilibria can be characterized by random mating, partial isolation, or complete isolation. We start

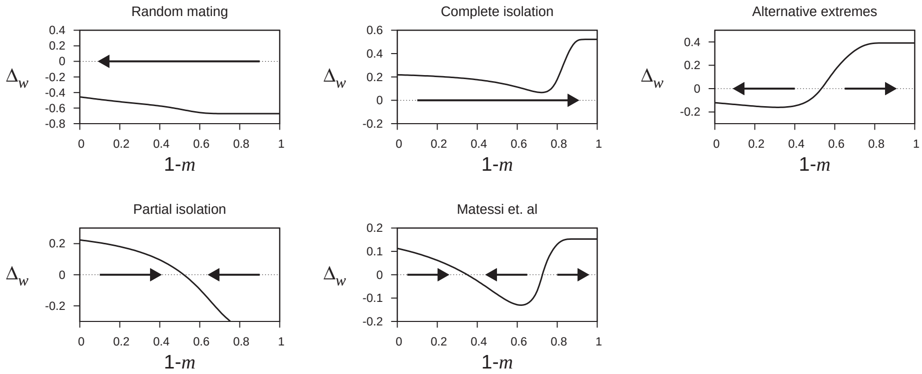


Figure 4.2: The relative fitness of homozygotes as a function of female choosiness in the five dynamic regimes. The plots show $\Delta_w = W_{\text{hom}} - W_{\text{het}}$ as a function of $1 - m$. Arrows indicate the direction of selection on $1 - m$. Parameters: $x = 1.2, \sigma_c = 0.6$ (random mating), $x = 0.3, \sigma_c = 0.3$ (complete isolation), $x = 0.7, \sigma_c = 0.9$ (alternative extremes), $x = 0.4, \sigma_c = 0.2$ (partial isolation), $x = 0.55, \sigma_c = 0.4$ (Matessi et al.).

by giving the stability conditions for these types of equilibria. For the moment, we assume that the ecological locus is at the polymorphic equilibrium described in Appendix 2.2.

Stability of random mating If mating is random ($m = 1$), sexual selection on males is absent (i.e. $\phi_i = 1$ for all i) and the evolutionary dynamics are determined by natural selection only. Therefore, the condition for stability of random mating is the same as in the model without sexual selection, that is, it is described by condition 4.13 (red line in Figure 4.4).

Stability of complete isolation In the model with sexual selection, complete isolation ($m = 0$) is evolutionarily stable if

$$k > \frac{1 + a'}{4a} \tag{4.16}$$

(Appendix 2.3; blue line in Figure 4.4), which is the case if stabilizing selection is weak (k large, x small) and competition is weakly frequency-dependent (a large, σ_c large). By comparing the blue lines in Figures 4.1 and 4.4, one can see that the parameter range where complete isolation is stable is considerably increased by sexual selection. This is because heterozygote males have a low mating rate when rare.

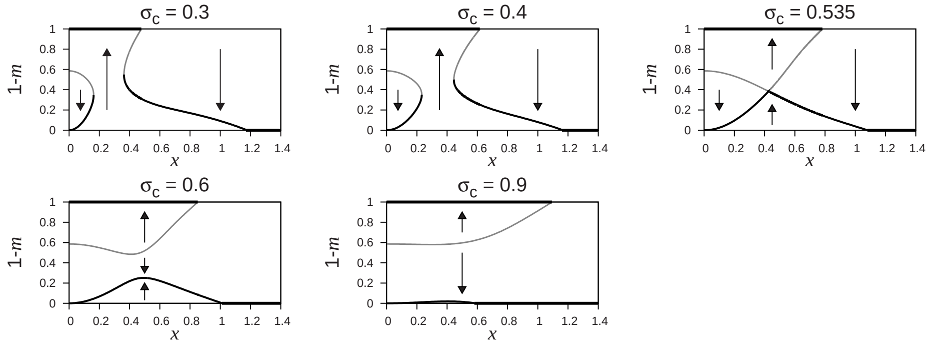


Figure 4.3: Equilibrium values of female choosiness $1-m$ (the probability of a heterozygote female to reject a homozygote male) as a function of x, for various values of σ_c. Black lines show stable equilibria and gray lines unstable equilibria. Arrows indicate the direction of selection. Note that m depends on x.

Stability of partial isolation Intermediate equilibria must be determined numerically by solving the condition $W_{\text{hom}} = W_{\text{het}}$ with respect to m. An intermediate equilibrium with $m = \hat{m}$ is stable if $W_{\text{hom}} < W_{\text{het}}$ for $m < \hat{m}$ and $W_{\text{hom}} > W_{\text{het}}$ for $m > \hat{m}$ (see Figure 4.2). Intermediate equilibria are maintained by a balance between negatively frequency-dependent natural selection, which tends to decrease the frequency of heterozygotes, and positively frequency-dependent sexual selection, which tends to increase the frequency of heterozygotes. Therefore, the heterozygote-to-homozygote ratio n is always larger than for the same parameters in the model without sexual selection (results not shown).

Additional regimes

Unlike in the model without sexual selection, the fact that an equilibrium is evolutionarily stable does not guarantee that it is actually reached. For some parameter combination, two equilibria can be stable simultaneously, and the outcome of evolution then depends on initial conditions. It is, therefore, necessary to study the invasion fitness gradient for m (as approximated by the difference $W_{\text{hom}}-W_{\text{het}}$) over all possible mating strategies (Figures 4.2 and 4.3). This analysis reveals that, in addition to the random mating, partial isolation, and complete isolation regimes, there are two additional regimes characterized by alternative equilibria (Figure 4.4).

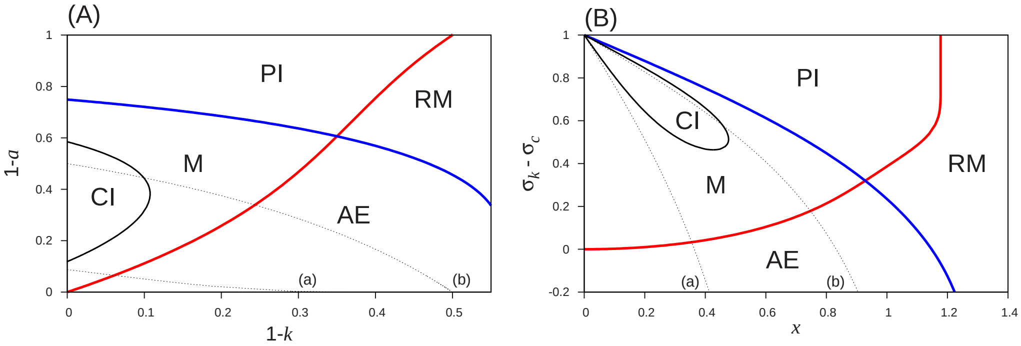


Figure 4.4: The five regimes for the evolution of female choosiness in the model with sexual selection: random mating (RM), alternative extremes (AE), partial isolation (PI), complete isolation (CI), and the Matessi et al. regime (M). In (A), the regimes are shown in the $1-a$ versus $1-k$ parameter space. In (B), the same results are presented in the x versus $\sigma_k - \sigma_c$ plane (assuming $\sigma_k = 1$). Note that the axis are oriented in a way such that a high value corresponds to strong selection. To the right of the red boundary line, random mating is evolutionarily stable (see inequality 4.13). To the left of the blue line, complete isolation is stable (see inequality 4.16). To the left of the dotted line marked (a), the polymorphic equilibrium is unstable for intermediate values of m (see Figure 4.5). To the left of the dotted line marked (b), the monomorphic equilibria are stable if m is sufficiently small (see inequality 4.17).

The alternative extremes regime If both random mating and complete isolation are stable, the outcome depends on the initial level of choosiness. If choosiness is already strong, sexual selection against heterozygotes will drive the population towards complete isolation, whereas, if choosiness is weak, both natural and sexual selection will drive it towards random mating.

The Matessi et al. regime In the area of parameter space where random mating is unstable and complete isolation is stable, there are two different regimes. For large k and intermediate a, complete isolation evolves from all initial conditions. This is the complete isolation regime, as described in the model without sexual selection. For smaller k or more extreme a, however, complete isolation is not the only stable equilibrium. Instead, there are two additional intermediate equilibria, one stable and one unstable. A population that starts at random mating and evolves in small steps will reach the stable intermediate equilibrium, and only a population that already starts with a high level of choosiness can evolve to complete isolation. This regime has first been described by MATESSI *et al.* (2001). Therefore, we call it the *Matessi et al. regime*. (Note that DOEBELI (1996) also briefly describes stable intermediate equilibria of assortativeness, and he also describes that the outcome of his simulations depends on initial conditions.)

Stability of the polymorphic equilibrium

So far, we have assumed that the ecological locus is always at a polymorphic symmetric equilibrium, with the proportion of heterozygotes determined by female choosiness (see Appendix 2.2). This equilibrium is, indeed, always favored by natural selection (for $\sigma_c < \sigma_k$ by frequency-dependent disruptive selection, and for $\sigma_c > \sigma_k$ due to heterozygote advantage). However, monomorphic equilibria (containing either only the + or the − allele) may become stable due to the positive frequency-dependence of sexual selection (acting against rare males). Obviously, speciation is only possible if the ecological locus is polymorphic.

In Appendix 2.4, we show that the monomorphic equilibria are locally stable if

$$m < 2ak - 1, \tag{4.17}$$

that is, if sexual selection deriving from female choosiness is strong enough relative to natural selection (line (b) in Figure 4.4).

Stability of the polymorphic equilibrium can be computed numerically by standard linear stability analysis (i.e., by numerically calculating the eigenvalues of system A10 at this equilibrium). In cases where both the polymorphic and the monomorphic equilibria are locally stable, their respective domains of attractions can be estimated by iterating system (A10) with different initial allele frequencies. As shown in Fig. 4.5, the domains of attraction of the monomorphic equilibria tend to be very small whenever the polymorphic equilibrium is locally stable. Therefore, loss of polymorphism mainly plays a role when the polymorphic equilibrium is unstable. This is most likely for intermediate m and small x. More precisely, the range of x where the polymorphic equilibrium can be unstable falls mainly into the domain of the Matessi et al. regime (line marked (a) in Figure 4.4).

Simulation results

For the complete isolation, partial isolation and random mating regimes, our simulations of the individual-based model with random mutational stepsize confirmed the results from the invasion analysis. The expected equilibrium is always reached, independent of the starting point of the simulations. Some

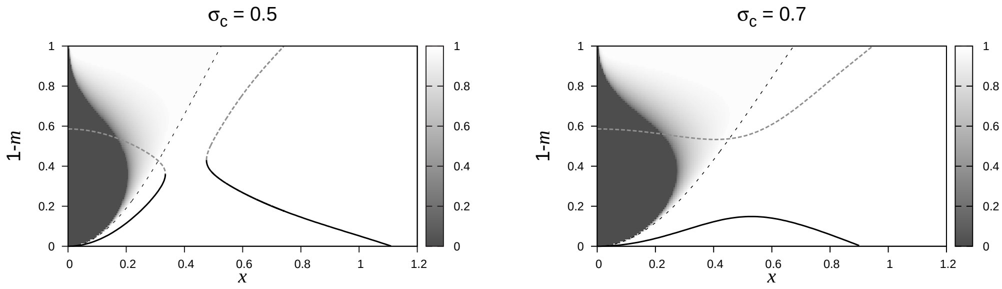


Figure 4.5: Stability of the polymorphic equilibrium of the ecological locus as a function of x and female choosiness $1-m$. The grey and black lines are the same as in Figure 4.3, that is they show unstable and stable equilibria for $1-m$. In the dark grey area, the polymorphic equilibrium is unstable. To the left of the dashed line, the monomorphic equilibrium is locally stable (see inequality 4.17. Shades of grey indicate the size of the domain of attraction of the polymorphic equilibrium in terms of the frequency of the + allele (white: polymorphic equilibrium is globally stalbe; dark grey: polymorphic equilibrium unstable). Note that a population where $1-m$ evolves in small steps, starting from random mating ($1-m=0$) will reach the stable intermediate equilibrium before the polymorphic equilibrium of the ecological locus looses stability.

example runs are shown in Figure 4.6. In the complete isolation regime, assortativeness goes up until m is about 0.2. At that point, there are practically no heterozygotes left in the population, the probability that two different homozygotes mate with each other is very low ($\approx 0.2^4 = 0.0016$) and selection for assortative mating is very weak. For the alternative extremes regime, the outcome of the simulations has some element of stochasticity. If the simulation is started with an m value close to the unstable equilibrium, then the population may cross this unstable equilibrium by drift and end up at the “wrong” phenotype (Figure 4.7). Finally, the behavior of the model in the Matessi et al. regime is more complicated, and we subjected it to a more extensive analysis.

Analysis of the Matessi et al. regime

In the Matessi et al. regime, a finite population that starts at random mating and that is polymorphic for the ecological locus can evolve towards three possible evolutionary equilibria. (1) It can lose the polymorphism at the ecological locus, (2) it can end up at the stable equilibrium with intermediate m and (3) it can end up at the stable equilibrium at complete isolation ($m = 0$), thereby “jumping” over the unstable intermediate equilibrium (see Figure 4.2). The outcome depends on the mutation rate and mutational stepsize, but also

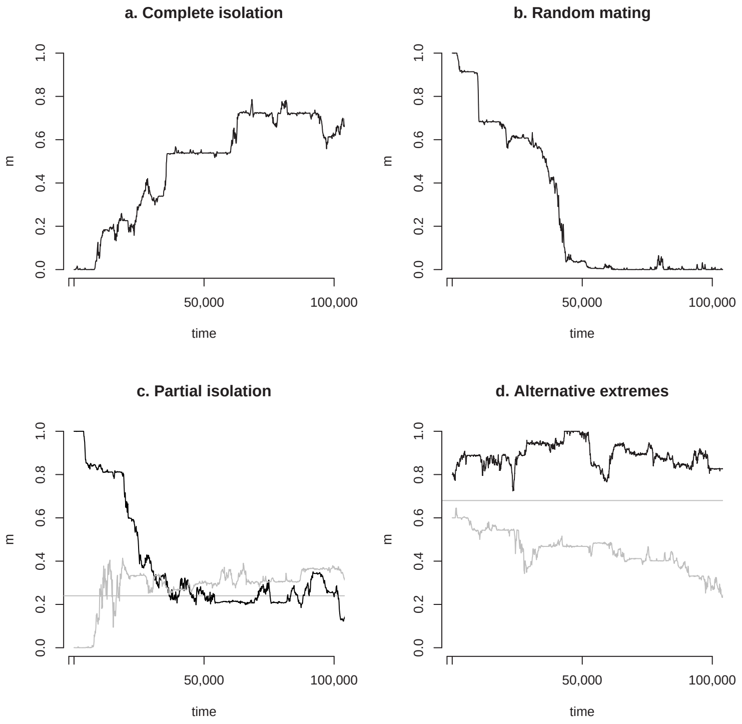


Figure 4.6: Outcomes of individual-based simulations for four regimes. Parameter values: (a) $\sigma_k = 1$, $\sigma_c = 0.3$, $x = 0.3$, $m(\text{start}) = 1$. (b) $\sigma_k = 1$, $\sigma_c = 0.6$, $x = 1.2$, $m(\text{start}) = 1$. (c) $\sigma_k = 1$, $\sigma_c = 0.3$, $x = 0.6$, $m(\text{start}) = 1$. (d) $\sigma_k = 1$, $\sigma_c = 0.9$, $x = 0.7$, $m(\text{start}) = 0.25$ or 0.4. Other parameters: $K_0 = 500$, mutation rate $=4.10^{-4}$, mutational stepsize is from normal distribution with mean 0.1.

on the values of σ_c and x, because these determine the strength of selection and therefore the stability of the equilibria in finite populations. We have done simulations to determine the impact of these factors. Unfortunately, our computer resources did not allow us to do a thorough analysis of the effect of population size as well. Population size certainly also plays a role, because it determines the importance of drift and the number of mutations entering the population.

Effect of σ_c and x. We were interested in the evolutionary outcome when the population starts at random mating. We therefore ran simulations with different values of σ_c and x, starting with $m = 1$ and the ecological locus polymorphic with allele frequencies 0.5. Mutation only happens at the choosiness locus. The results of 20 simulation runs per parameter combination are shown

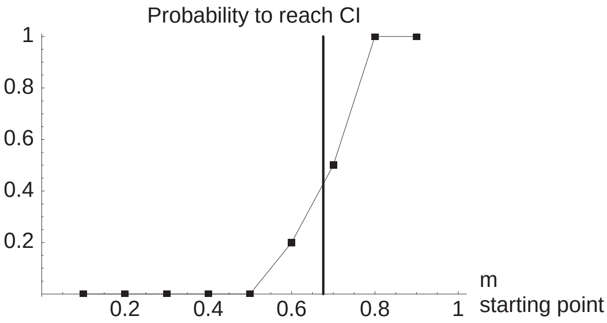


Figure 4.7: Probability that a simulation run ends with the population being in complete isolation, rather than random mating depending on the starting value of m, for one parameter combination from the alternative extremes regime. $\sigma_c = 0.9$ and $x = 0.7$. The vertical line shows the m value at the unstable intermediate equilibrium. Other parameters: $K_0 = 500$, mutation rate $u = 4 \cdot 10^{-4}$; the mutational stepsize is 0.1.

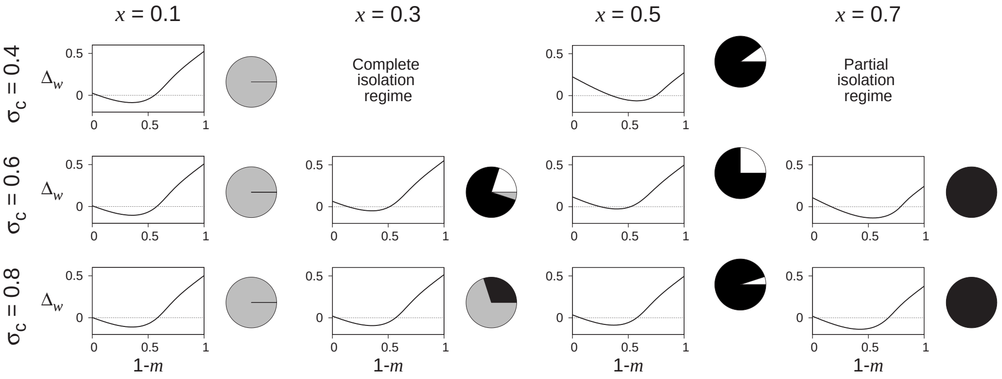


Figure 4.8: The relative fitness of homozygotes as a function of female choosiness and the distribution of simulation outcomes for several points in the Matessi et al. regime. The line plots show $\Delta_w = W_{\text{hom}} - W_{\text{het}}$ as a function of $1 - m$. The pie charts are based on 20 simulations each and show the probability that the polymorphism at the ecological locus is lost (denoted by grey), or complete isolation is reached within 100,000 generations (white). If neither was the case, we assume that the population is "stuck" at the stable equilibrium with intermediate m (black). The parameter combination that was analyzed in Figures 3 and 5 of Dieckmann and Doebeli (1999) corresponds to is $\sigma_c = 0.4$, $x = 0.5$. Other parameters: $K_0 = 500$, mutation rate $u = 4 \cdot 10^{-4}$, mutational stepsize equal to 0.1.

in Figure 4.8. The polymorphism is always lost when x is small and σ_c is large, which is the area where the polymorphic equilibrium is unstable for intermediate m. Close to this area, where the polymorphic equilibrium is stable, but its domain of attraction is small, the polymorphism is lost sometimes (compare to Figures 4.4 and 4.5). In an area close to the complete isolation regime, the population sometimes evolves to complete isolation. For this to happen, the population has to cross a range of m values where selection favours weaker assortativeness. To understand why this "jump" is possible in some cases but not in others, we looked at the difference between homozygote and heterozygote fitness in the neighborhood of the intermediate stable equilibrium (also shown in Figure 4.8). Selection can be strong only if this difference is large. We find that in the cases where the population makes the jump, there is only a small range of m values where selection is in the direction of less assortative mating and, in these cases, this selection is not very strong, i.e. the heterozygotes are only slightly more fit than the homozygotes.

If it is possible that the population jumps to the equilibrium at $m = 0$, it may also jump back. To check whether this happens, we did simulations over longer time spans. We find that the population may jump back and forth between the equilibria, but this happens regularly only for some parameter combinations. This result can be understood by looking at the homozygote advantage at the complete isolation equilibrium (see Figure 4.8). When this advantage is large, it means that selection to maintain complete isolation is strong. We find that in most cases jumping back is rare and selection to stay in complete isolation is strong. We therefore expect the equilibrium at $m = 0$ to be more stable than the equilibrium at intermediate m. To confirm this hypothesis, we did simulations in which the population starts exactly at one of the two equilibria. We then introduce a mutant that has the choosiness level of the other equilibrium and we wait to see if the mutant can invade. If the fixation probability of such a mutant is higher in one direction than in the other, this suggests that over long time spans, the population will spend more time at that equilibrium. This analysis of fixation probabilities confirms what we expected. With few exceptions, the complete isolation equilibrium was more stable than the stable intermediate equilibrium. One exception is the parameter combination that was used by Dieckmann and Doebeli (1999), ($\sigma_c = 0.4$ and $x = 0.5$) where the equilibria seem neutral with respect to each other. Another exception is found at points such as $\sigma_c = 0.4$ and $x = 0.6$, where the intermediate equilibrium is much more stable than complete isolation (data not shown).

Effect of mutation rate and mutation effect size. Next, we looked at the effect of the mutation rate and the size of the mutational effect for one combination of σ_c and x. The results are in Figure 4.9. When the stepsize is 0.25 (on the scale of m), this means there are 4 steps between random mating

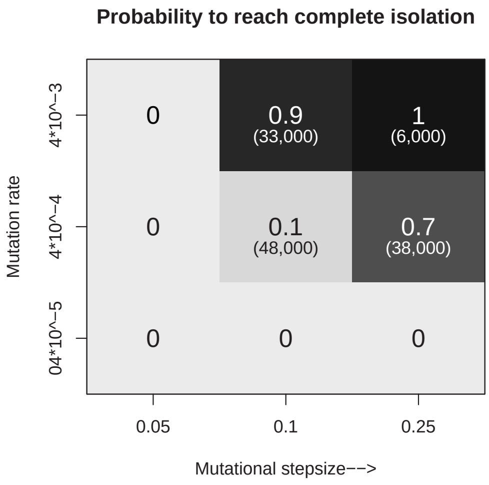


Figure 4.9: (Preliminary results) The probability to reach the "complete isolation" equilibrium within $100,000$ generations, depending on the mutation rate and the mutational stepsize. Mutations have a fixed stepsize. The number of steps between random mating and complete isolation is 1/stepsize, and ranges from 4 to 20. In brackets is the mean number of generations until the complete isolation equilibrium was reached. The population was assumed to be in complete isolation when there were no heterozygotes left. Other parameters: $\sigma_c = 0.4$, $x = 0.5$, $K_0 = 500$.

and complete isolation. We find, as expected, that with higher mutation rate and larger stepsize, the complete isolation equilibrium is reached more often and faster.

4.4 Discussion

Under what conditions can intraspecific competition lead to sympatric speciation? – This question has been at the focus of much recent debate (Doebeli and Dieckmann 2005; Doebeli *et al.* 2005; Gavrilets 2005; Polechová and Barton 2005; Waxman and Gavrilets 2005). Here, we have analyzed a one-locus version of the Roughgarden (1972) model as used by Dieckmann and Doebeli (1999), which gives us a detailed overview of what can happen in a model of competitive speciation.

We find that evolution of complete reproductive isolation – and, as a consequence, sympatric speciation – is possible in a relevant area of parameter space. More precisely, complete isolation is predicted for an intermediate niche width (as determined by the competition parameter σ_c) if stabilizing selection is not too strong (small x or $1 - k$ in Figures 4.1 and 4.4). A necessary condition can be seen from our model without sexual selection: Competition between homozygotes of opposite type must be substantially smaller than competition between homozygotes and heterozygotes, such that the net reduction in competition due to a split into isolated clusters outweighs stabilizing selection from the resource distribution ($a' \ll a$). If the niche width is too broad (weak frequency dependence of competition, σ_c large), the disruptive force due to competition is too weak to generate multiple niches. If the niche width is too narrow (frequency dependence of competition too strong), a third niche at an intermediate phenotype opens up that is filled by the heterozygotes. Finally, even for an optimal range of competition, that is, full competition between homozygotes and heterozygotes ($a = 1$), but no competition between homozygotes of opposite type ($a' = 0$), complete isolation does not evolve if stabilizing selection is too strong (from equation 4.16, we see that complete isolation is always unstable for $k < 1/4$). In our model without sexual selection, this necessary condition is also sufficient for speciation to happen. However, with sexual selection, complete isolation is not always reached in the described parameter range, for reasons which we explain in the next section.

In addition to sympatric speciation, we find a large parameter region in both our models where assortative mating always evolves to an intermediate equilibrium value (partial isolation regime in Figure 4.1 and 4.4). Partial reproductive isolation evolves, in particular, if competition is strongly frequency-dependent (short ranged), such that a third niche for heterozygotes emerges. The size of this niche determines the proportion of heterozygotes that is sustained and the appropriate level of choosiness. If the niche for heterozygotes gets sufficiently large so that the (optimal) ratio of heterozygotes and homozygotes n exceeds 2, random mating becomes stable against increased levels of assortativeness (for $n > 2$, dis-assortative mating would evolve if this was possible). Note that partial reproductive isolation in natural populations is frequently interpreted as indicating “incipient speciation”. Our model shows that this is not necessarily true: Frequency-dependent selection can allow coexistence of three phenotypes, but with $n < 2$, partial isolation can be a stable outcome of evolution.

Natural and sexual selection

Our results can be explained by the interplay of natural and sexual selection (see also Kirkpatrick and Nuismer 2004; Gourbiere 2004). We analyzed the roles of the two selective forces by comparing the behavior of the model with and without sexual selection. Without sexual selection, we find only three regimes: complete isolation, partial isolation and random mating. Each regime has only one stable equilibrium, and the ecological polymorphism is always stable.

Adding sexual selection against rare males makes the behavior of the model more complex. In addition to the three regimes mentioned above, there are now two regimes – the alternative extremes regime and the Matessi et al. regime – where the evolutionary outcome depends on initial conditions. In these regimes, sexual selection favors heterozygotes when they are common, and this can stop (further) increase of assortativeness. If elevated levels of assortative mating already exist in a population, sexual selection can promote the evolution of complete isolation. Therefore, sexual selection causes a big increase in the parameter range in which complete isolation is stable (shown by the area left of the blue line, in Figure 4.1 and 4.4). This effect is particularly striking in the alternative extremes regime, where natural selection alone favors random mating. In addition, in some parts of the parameter space, sexual selection can cause a loss of the ecological polymorphism. This is because males carrying a rare allele will have difficulties finding a mate.

The region where complete isolation can evolve in small steps in an infinite population (complete isolation regime in Figure 4.4) is considerably reduced relative to the model without sexual selection. In a larger range (corresponding to the Matessi et al. regime in Figure 4.4), complete isolation is stable, but there is another stable equilibrium at intermediate assortativeness. Evolution in small steps will always stop at the intermediate equilibrium with partial isolation. However, our simulations show that evolution of complete isolation is still possible by "jumping" the intermediate optimum if mutation rates and mutational effects are sufficiently high. In the extreme case that a single mutation results in complete assortativeness, simulations show that in about 80% of the parameter range corresponding to the Matessi et al. regime such a mutation has positive invasion fitness if the population is currently at the intermediate equilibrium.

The evolution of complete isolation can also be inhibited by the loss of the ecological polymorphism, as previously described by Bürger and Schneider (2006a). For weak frequency dependence of competition and weak stabilizing selection (the lower left corner of Figure 4.4) the polymorphic equilibrium becomes unstable for intermediate values of assortativeness. Our simulations show that in this case almost always a monomorphic equilibrium is reached. In Bürger and Schneider (2006a), the loss of the polymorphism was described as a consequence of the evolution to an intermediate optimum for assortativeness in the Matessi et al. regime. However, our model shows that these phenomena are two different mechanisms that can both prevent speciation. An infinite population that takes small mutational steps will always evolve to partial isolation without loosing the polymorphism (see Figure 4.3). A finite population, however, can loose the polymorphism if it moves past the stable intermediate equilibrium by drift (Figure 4.8).

Discussion of the modelling approach

Our approach in this study was to analyze a simplified version of the Dieckmann and Doebeli (1999) model. This approach allowed us to (1) analyze the model in the entire ecological parameter space, (2) to gain a detailed and intuitive understanding of the interaction between the various selective forces and (3) to unify, in a single model, a large number of phenomena which have previously been studied or described individually in separate models. The latter include (1) the role of natural versus sexual selection (see Gourbiere 2004; Kirkpatrick and Nuismer 2004), (2) conditions for the maintenance or loss of the ecological polymorphism (see Kirkpatrick and Nuismer 2004; Bürger and Schneider 2006a,b), (3) potential evolutionary stability of incomplete isolation (Doebeli 1996; Matessi *et al.* 2001), and (4) the importance of ecological niches and a resulting non-linear relationship between niche width and the likelihood of speciation (Gourbiere 2004; Bolnick 2006; Bürger and Schneider 2006b).

Our work can be seen as an extension of the study by Matessi *et al.* (2001). These authors used a quadratic approximation to our fitness function, which is valid if overall selection is weak. More precisely, the approximation guarantees that selection is always purely disruptive and heterozygotes always have lowest viability, as is the case to the left of the blue line in Figure 4.1. In accordance with our results for this area, Matessi *et al.* (2001) found two of or five regimes – the Matessi et al. regime and the complete isolation regime. It is in the regions where their approximation is not valid that we find the other three regimes.

The key simplification in our model is the assumption that the ecological trait is determined by a single locus with two alleles. How general are our results with regard to the genetic architecture of the trait? – On the one hand, it seems reasonable to expect that the five regimes described in this paper are generic also for other genetic architectures, because the interplay of natural and sexual selection, which we have described here, should be qualitatively independent of genetic details. This intuition is supported by the observation that a behavior similar to the Matessi et al. regime was also found in a multilocus model by Doebeli (1996), and complete isolation was, of course, reached in the initial multilocus model by Dieckmann and Doebeli (1999).

On the other hand, the one-locus assumption has the obvious consequence that intermediate phenotypes can only exist as heterozygotes. Therefore, whenever more than two phenotypes can potentially coexist, natural selection tends to move the population toward partial isolation or random mating. In a model with a different genetic architecture, the evolution of assortative mating might instead lead to more than two reproductively isolated species (Bolnick 2006; Bürger and Schneider 2006b). Thus, the behavior of the model in the partial isolation, random mating and alternative extremes regimes might be different in a model with more than one ecological locus. In particular, speciation may be possible for a larger range of parameters.

Another important assumption of our model is that the allelic effect of the ecological locus (x) is constant, whereas, in principle it might also be subject to selection (Geritz and Kisdi 2000; Kopp and Hermisson 2006; Kristan Schneider, unpublished manuscript). The coevolution of genetic architecture and assortative mating seems to be an interesting avenue for future studies.

4.5 Appendices

Appendix 1: Analysis of the model without sexual selection

We need to solve the linear equation system (4.6c) for the $Q(X)$. It can be shown that the ecological locus is always polymorphic. Concentrating, therefore, on the symmetric equilibrium with $N^+_{\text{hom}} = N^-_{\text{hom}}$, the mating rates for homozygotes and heterozygotes are given by

$$\phi_{\text{hom}} = (1+m')N_{\text{hom}}Q_{\text{hom}} + mN_{\text{het}}\frac{Q_{\text{hom}}+Q_{\text{het}}}{2} = 1 \tag{A1}$$

$$\phi_{\text{het}} = N_{\text{het}}Q_{\text{het}} + mN_{\text{hom}}(Q_{\text{hom}}+Q_{\text{het}}) = 1. \tag{A2}$$

This is solved by

$$Q_{\text{hom}} = \frac{n+(1-n/2)m}{N_{\text{hom}}[(1+m')(n+m)+n^2m/2]} \tag{A3}$$

$$Q_{\text{het}} = \frac{1+m'-(1-n/2)m}{N_{\text{hom}}[(1+m')(n+m)+n^2m/2]}. \tag{A4}$$

From the condition $W_{\text{hom}} = W_{\text{het}} = 0$ and equation (4.10), we see that $d_{\text{hom}} = d_{\text{het}} = 1$. Furthermore, using equation (4.8), the latter condition is fulfilled if

$$n = \hat{n} \equiv \frac{1-2ak+a'}{k-a}. \tag{A5}$$

In our model, this equation is biologically meaningful only for $0 \leq \hat{n} \leq 2$ (as we do not allow for dis-assortative mating). Outside of this range, n equals either 0 or 2 (at the evolutionary equilibrium). Finally, the birth rate for homozygotes is

$$B_{\text{hom}} = N^2_{\text{hom}}Q_{\text{hom}} + mN_{\text{hom}}N_{\text{het}}\frac{Q_{\text{hom}}+Q_{\text{het}}}{2} + \frac{1}{4}N^2_{\text{het}}Q_{\text{het}}. \tag{A6}$$

With the condition $B_{\text{hom}}/N_{\text{hom}} = 1$ (from equation 9) and with the above values for n and Q_{hom} and Q_{het}, we obtain a condition for m (and $m' = m^4$), which can be solved numerically. For $0 < n < 2$, there is always a unique positive solution. For $n = 0$, $m = 0$, and for $n > 2$, $m = 1$.

Appendix 2: Analysis of the model with sexual selection

Appendix 2.1: Genotype fitnesses and dynamics in the one-locus, two-allele model

Here, we spell out the equations for the one-locus, two-allele model with constant female choosiness. The effective population sizes with respect to competition (see eq. 4.3) are given by

$$
\begin{aligned}
A^+_{\text{hom}} &= N^+_{\text{hom}} + aN_{\text{het}} + a'N^-_{\text{hom}}, &\text{(A7a)}\\
A_{\text{het}} &= aN^+_{\text{hom}} + N_{\text{het}} + aN^-_{\text{hom}}, &\text{(A7b)}\\
A^-_{\text{hom}} &= a'N^+_{\text{hom}} + aN_{\text{het}} + N^-_{\text{hom}}. &\text{(A7c)}
\end{aligned}
$$

Similarly, for given m, the "female activity factors" are given by

$$
\begin{aligned}
Q^+_{\text{hom}} &= (N^+_{\text{hom}} + mN_{\text{het}} + m'N^-_{\text{hom}})^{-1}, &\text{(A8a)}\\
Q_{\text{het}} &= (mN^+_{\text{hom}} + N_{\text{het}} + mN^-_{\text{hom}})^{-1}, &\text{(A8b)}\\
Q^-_{\text{hom}} &= (m'N^+_{\text{hom}} + mN_{\text{het}} + N^-_{\text{hom}})^{-1}. &\text{(A8c)}
\end{aligned}
$$

With these definitions, the fitness functions of the three ecological genotypes (according to eq. 4.10) can be written as

$$
\begin{aligned}
W^+_{\text{hom}} &= \frac{1}{2}\Big(1 + N^+_{\text{hom}}Q^+_{\text{hom}} + mN_{\text{het}}Q_{\text{het}} + m'N^-_{\text{hom}}Q^-_{\text{hom}}\Big) - \frac{A^+_{\text{hom}}}{K^+_{\text{hom}}}, &\text{(A9a)}\\
W_{\text{het}} &= \frac{1}{2}\Big(1 + mN^+_{\text{hom}}Q^+_{\text{hom}} + N_{\text{het}}Q_{\text{het}} + mN^-_{\text{hom}}Q^-_{\text{hom}}\Big) - \frac{A_{\text{het}}}{K_{\text{het}}}, &\text{(A9b)}\\
W^-_{\text{hom}} &= \frac{1}{2}\Big(1 + m'N^+_{\text{hom}}Q^+_{\text{hom}} + mN_{\text{het}}Q_{\text{het}} + N^-_{\text{hom}}Q^-_{\text{hom}}\Big) - \frac{A^-_{\text{hom}}}{K^-_{\text{hom}}}. &\text{(A9c)}
\end{aligned}
$$

Finally, the dynamics of genotype frequencies (see eq. 4.9) are given by

$$\begin{aligned}\dot{N}^{+}_{\mathrm{hom}} =& N^{+}_{\mathrm{hom}}\left(N^{+}_{\mathrm{hom}} + \frac{mN_{\mathrm{het}}}{2}\right)Q^{+}_{\mathrm{hom}} + \\ & N_{\mathrm{het}}\left(\frac{mN^{+}_{\mathrm{hom}}}{2} + \frac{N_{\mathrm{het}}}{4}\right)Q_{\mathrm{het}} - \frac{N^{+}_{\mathrm{hom}}A^{+}_{\mathrm{hom}}}{K^{+}_{\mathrm{hom}}},\end{aligned} \tag{A10a}$$

$$\begin{aligned}\dot{N}_{\mathrm{het}} =& N^{+}_{\mathrm{hom}}\left(\frac{mN_{\mathrm{het}}}{2} + m'N^{-}_{\mathrm{hom}}\right)Q^{+}_{\mathrm{hom}} + \frac{N_{\mathrm{het}}}{2} + \\ & N^{-}_{\mathrm{hom}}\left(m'N^{+}_{\mathrm{hom}} + \frac{N_{\mathrm{het}}}{2}\right)Q^{-}_{\mathrm{hom}} - \frac{N_{\mathrm{het}}A_{\mathrm{het}}}{K_{\mathrm{het}}},\end{aligned} \tag{A10b}$$

$$\begin{aligned}\dot{N}^{-}_{\mathrm{hom}} =& N_{\mathrm{het}}\left(\frac{N_{\mathrm{het}}}{4} + \frac{mN^{-}_{\mathrm{hom}}}{2}\right)Q_{\mathrm{het}} + \\ & N^{-}_{\mathrm{hom}}\left(\frac{mN_{\mathrm{het}}}{2} + N^{-}_{\mathrm{hom}}\right)Q^{-}_{\mathrm{hom}} - \frac{N^{-}_{\mathrm{hom}}A^{-}_{\mathrm{hom}}}{K^{-}_{\mathrm{hom}}},\end{aligned} \tag{A10c}$$

Appendix 2.2: The symmetric polymorphic equilibrium

Here, we show how to calculate the polymorphic equilibrium of system (A10). Let $n \equiv \frac{N_{\mathrm{het}}}{N_{\mathrm{hom}}}$ denote the frequency of heterozygotes relative to the frequency of one the homozygotes. Because, at the polymorphic equilibrium, $B_{\mathrm{hom}} = N_{\mathrm{hom}}d_{\mathrm{hom}}$ and $B_{\mathrm{het}} = N_{\mathrm{het}}d_{\mathrm{het}}$, the equilibrium value $\hat{n}$ satisfies

$$\hat{n} = \frac{d_{\mathrm{hom}}B_{\mathrm{het}}}{d_{\mathrm{het}}B_{\mathrm{hom}}} = 2\left(\frac{2m' + (m + \frac{1}{2} + \frac{m'}{2})\hat{n} + \frac{m}{2}\hat{n}^2}{2 + (m + \frac{1}{2} + \frac{m'}{2})\hat{n} + \frac{m}{2}\hat{n}^2}\right)\left(\frac{1 + a' + a\hat{n}}{2a + \hat{n}}\right)\frac{1}{k}. \tag{A11}$$

This is a fourth-order equation that can be solved analytically (e.g., by using *Mathematica*) and has exactly one positive solution (proof?). Once $\hat{n}$ is known, it is straightforward to arrive at the equilibrium values for N_{hom} and N_{het}.

Appendix 2.3: Stability of complete isolation

Complete isolation ($m = 0$) is stable if $W_{\mathrm{het}} < 0$ given $N_{\mathrm{het}} = 0$ and $N^{+}_{\mathrm{hom}} = N^{-}_{\mathrm{hom}}$, that is if heterozygotes cannot invade a population of only homozygotes.

First, it must be noted, that completely assortative mating $m = 0$ does not ensure the absence of heterozygotes. However, the condition for the absence of heterozygotes is the same as for the stability of complete assortativeness (namely that heterozygotes have negative fitness when rare). Therefore, in the following, we can indeed set $N_{\mathrm{het}} = 0$.

Next, we have to look at mating success of heterozygote males in the limit of $m \to 0$. We have

$$\phi_{\mathrm{m,het}} = \frac{2mN_{\mathrm{hom}}}{N_{\mathrm{hom}} + mN_{\mathrm{het}} + m'N_{\mathrm{hom}}} + \frac{N_{\mathrm{het}}}{2mN_{\mathrm{hom}} + N_{\mathrm{het}}} \tag{A12}$$

While the first term on the right-hand side of this equation clearly goes to 0 for $m \to 0$, the limit of the second term depends on whether N_{het} approaches 0 slower or faster than m. By solving a first-order approximation of equation (A10c) for N_{het}, it can be shown that the equilibrium value of N_{het} is proportional to m', which in turn, approaches 0 faster than m. Therefore, $\lim_{m\to 0} \phi_{\mathrm{m,het}} = 0$.

With this information, it follows from (4.10) that $W_{\mathrm{het}} = 1/2 - d_{\mathrm{het}}$, which is negative if

$$d_{\mathrm{het}} = \frac{2aN_{\mathrm{hom}}}{K_{\mathrm{het}}} < \frac{1}{2}. \tag{A13}$$

It follows easily from (A10) that $N_{\mathrm{hom}} = K_{\mathrm{hom}}/(1 + a')$, yielding the stability condition (4.16).

Appendix 2.4: Stability of monomorphic equilibria

Local stability of a monomorphic equilibrium (say with the '+' allele fixed) can be calculated analytically by focusing on the fitness of an invading (mutant) '−' allele. As long as this allele is rare, it will occur almost exclusively in heterozygotes. The monomorphic equilibrium is stable if the mutant allele cannot invade, which is the case if $W_{\mathrm{het}} < 0$. It is easy to see (from equations A9 for $N_{\mathrm{het}} \to 0$) that $\phi_{\mathrm{m,hom}} = m$ and $d_{\mathrm{het}} = ak$. Together with equation (4.10), this leads to condition (4.17).

Summary

Chapter 1

There are two ways in which a population can adapt to a rapid environmental change or habitat expansion. It may either adapt through new beneficial mutations that subsequently sweep through the population or by using alleles from the standing genetic variation. We use diffusion theory to calculate the probabilities for selective adaptations and find a large increase in the fixation probability for weak substitutions, if alleles originate from the standing genetic variation. We then determine the parameter regions where each scenario – standing variation vs. new mutations – is more likely. Adaptations from the standing genetic variation are favored if either the selective advantage is weak or the selection coefficient and the mutation rate are both high. Finally, we analyze the probability of "soft sweeps", where multiple copies of the selected allele contribute to a substitution and discuss the consequences for the footprint of selection on linked neutral variation. We find that soft sweeps with weaker selective footprints are likely under both scenarios if the mutation rate and/or the selection coefficient is high.

Chapter 2

In the classical model of molecular adaptation, a favored allele derives from a single mutational origin. This ignores that beneficial alleles can enter a population recurrently, either by mutation or migration, during the selective phase. In this case, descendents of several of these independent origins may contribute to the fixation. As a consequence, all ancestral haplotypes that are linked to any of these copies will be retained in the population, affecting the

pattern of a selective sweep on linked neutral variation. In this study, we use analytical calculations based on coalescent theory and computer simulations to analyze molecular adaptation from recurrent mutation or migration. Under the assumption of complete linkage, we derive a robust analytical approximation for the number of ancestral haplotypes and their distribution in a sample from the population. We find that so-called “soft sweeps”, where multiple ancestral haplotypes appear in a sample, are likely for biologically realistic values of mutation or migration rates.

Chapter 3

Polymorphism data can be used to identify loci at which a beneficial allele has recently gone to fixation, given that an accurate description of the signature of selection is available. In the classical model that is used, a favored allele derives from a single mutational origin. This ignores the fact that beneficial alleles can enter a population recurrently by mutation during the selective phase. In this study, we present a combination of analytical and simulation results to demonstrate the effect of adaptation from recurrent mutation on summary statistics for polymorphism data from a linked neutral locus. We also analyze the power of standard neutrality tests based on the frequency spectrum or on linkage disequilibrium (LD) under this scenario. For recurrent beneficial mutation at biologically realistic rates we find substantial deviations from the classical pattern of a selective sweep from a single new mutation. Deviations from neutrality in the level of polymorphism and in the frequency spectrum are much less pronounced than in the classical sweep pattern. In contrast, for levels of LD the signature is even stronger if recurrent beneficial mutation plays a role. We suggest a variant of existing LD tests that increases their power to detect this signature.

Chapter 4

Models of competitive sympatric speciation have created much excitement, but they are also highly controversial. We present a thorough and largely analytical analysis of the evolution of assortative mating in a Roughgarden model, in which the ecological trait is determined by a single diallelic locus. The genetic architecture is then given by a single parameter: the allelic effect x. A second parameter, σ_c, determines the niche width or frequency-dependence of competition. Females are choosy and prefer mates with similar ecological phenotype. The degree of choosiness is determined by one locus with a continuum

of alleles. We describe five possible regimes for the evolution of choosiness. In only one of them can complete reproductive isolation evolve from random mating in small mutational steps. In addition, we determine the regions where the ecological polymorphism is unstable, locally stable or globally stable. Our simple model allows us to investigate the roles of natural and sexual selection in speciation. We find that complete isolation may fail to evolve when natural selection favors heterozygotes, when sexual selection favors heterozygotes or when sexual selection causes the ecological polymorphism to be unstable. Our findings are confirmed and extended by individual based simulations.

Bibliography

Bibliography

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson, and L. Kruglyak, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. **2:** 1591–1599.

Antonarakis, S. E., S. H. Orkin, and H. H. Kazazian Jr., 1982 Evidence for multiple origins of the (E)-globin gene in Southeast Asia. Proc. Nat. Acad. Sci. USA **79:** 9–117.

Barluega, M., K. N. Stölting, W. Salzburger, M. Muschik, and A. Meyer, 2006 Sympatric speciation in Nicaraguan crater lake cichlid fish. Nature **439:** 719–723.

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics **140:** 821–841.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. Camb. **72:** 123–133.

Barton, N. H., A. M. Etheridge, and A. Sturm, 2004 Coalescence in a random background. Ann. Appl. Probab. **14:** 754–785.

Bolnick, D. I., 2006 Multi-species outcomes in a common model of sympatric speciation. J. Theor. Biol. **241:** 734–744.

Bürger, R. and K. Schneider, 2006a Intraspecific competitive divergence and convergence under assortative mating. Am. Nat. **167:** 190–205.

Bürger, R. and K. Schneider, 2006b On the conditions for speciation through intraspecific competition. Evolution**:** in press.

Catania, F., M. O. Kauer, P. J. Daborn, J. L. Yen, R. H. Ffrench-Constant, and C. Schlötterer, 2004 World-wide survey of an *Accord* insertion and its association with DDT resistance in *Drosophila melanogaster*. Mol. Ecol. **13:** 2491–2504.

Depaulis, F., S. Mousset, and M. Veuille, 2001 Haplotype tests using coalescent simulations conditional on the number of segregating sites. Mol. Biol. Evol. **18:** 1136–1138.

Depaulis, F., S. Mousset, and M. Veuille, 2003 Power of neutrality tests to detect bottlenecks and hitchhiking. J. Mol. Evol. **57:** S190–S200.

Depaulis, F., S. Mousset, and M. Veuille, 2005 Detecting Selective Sweeps with Haplotype Tests: Hitchhiking and Haplotype Tests. In D. Nurminsky (Ed.), *Selective Sweep*. Landes Bioscience.

Depaulis, F. and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-sites model. Mol. Biol. Evol. **15:** 1788–1790.

Dieckmann, U. and M. Doebeli, 1999 On the origin of species by sympatric speciation. Nature **400:** 354–357.

Doebeli, M., 1996 A quantitative genetic competition model for sympatric speciation. J. Evol. Biol. **9:** 893–909.

Doebeli, M. and U. Dieckmann, 2005 Adaptive dynamics as a mathematical tool for studying the ecology of speciation processes. J. Evol. Biol. **18:** 1194–1200.

Doebeli, M., U. Dieckmann, A. J. Metz, and D. Tautz, 2005 What we have also learned: adaptive speciation is theoretically plausible. Evolution **59:** 691–695.

Durett, R., 2002 *Probability Models for DNA Sequence Evolution*. New York: Springer.

Durett, R. and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Pop. Biol. **66:** 129–138.

Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger, 2005 An approximate sampling formula under genetic hitchhiking. Preprint.

Ewens, W. J., 2004 *Mathematical Population Genetics (second edition)*. Berlin: Springer.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Harlow, Essex, UK: Addison Wesley Longman.

Fay, J. C. and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Fisher, R. A., 1930 *The genetical theory of natural selection*. Oxford, U.K.: Oxford University Press.

Flint, J., R. M. Harding, J. B. Clegg, and A. J. Boyce, 1993 Why are some genetic diseases common? Distinguishing selection from other processes by molecular analysis of globin gene variants. Hum. Genet. **91:** 91–117.

Gavrilets, S., 2004 *Fitness landscapes and the origin of species*. Princeton, NJ: Princeton University Press.

Gavrilets, S., 2005 "Adaptive speciation"—it is not that easy: a reply to Doebeli et al. Evolution **59:** 696–699.

Geritz, S. A., E. Kisdi, G. Meszéna, and J. Metz, 1998 Evolutionary singular strategies and the adaptive growth and branching of the evolutionary tree. Evol. Ecol. **12:** 35–57.

Geritz, S. A. H. and E. Kisdi, 2000 Adaptive dynamics in diploid, sexual populations and the evolution of reproductive isolation. P. Roy. Soc. Lond. B **267:** 1671–1678.

Gillespie, J., 1991 *The Causes of Molecular Evolution*. New York: Oxford University Press.

Gíslason, D., M. M. Ferguson, S. Skúlason, and S. S. Snorasson, 1999 Rapid and coupled phenotypic differentiation in Icelandic Arctic char (*Salvelinus alpinus*). Can. J. Fish. Aquat. Sci. **56:** 2229–2234.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. D. Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165:** 1269–1278.

Gourbiere, S., 2004 How do natural and sexual selection contribute to sympatric speciation? J. Evol. Biol. **17:** 1297–1309.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and selection history of *Drosophila melanogaster* populations. Genome Research **15:** 790–799.

Haldane, J. B. S., 1927 A mathematical theory of natural and artificial selection. Part V: Selection and mutation. Proc. Camb. Phil. Soc. **23:** 838–844.

Hamblin, M. T. and A. Di Rienzo, 2000 Detection of the Signature of Natural Selection in Humans: Evidence from the Duffy Blood Group Locus. Am. J. Hum. Genet. **66:** 1669–1679.

Hamblin, M. T., E. E. Thompson, and A. Di Rienzo, 2002 Complex signatures of natural selection at the Duffy blood group locus. Am. J. Hum. Genet. **70:** 369–383.

Hansen, T. F., C. Pelabon, W. S. Armbruster, and M. L. Carlson, 2003 Evolvability and genetic constraint in *Dalechampia* blossoms: Components of variance and measures of evolvability. J. Evol. Biol. **16:** 754–765.

Harr, B., M. Kauer, and C. Schlötterer, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99:** 12949–12954.

Hermisson, J. and P. S. Pennings, 2005 Soft Sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics **169:** 2335–2352.

Houle, D., 1992 Comparing evolvability and variability of quantitative traits. Genetics **130:** 195–204.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **28:** 337–338.

Innan, H. and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. USA **101:** 10667–10672.

Kacser, H. and J. A. Burns, 1981 The molecular basis of dominance. Genetics **97:** 6639–6666.

Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The "Hitchhiking Effect" Revisited. Genetics **123:** 887–899.

Keightley, P. D., 1996 A metabolic basis for dominance and recessivity. Genetics **143:** 621–625.

Kelly, J. K., 1997 A test on neutrality based on interlocus associations. Genetics **146:** 1179–1206.

Kern, A. D., C. D. Jones, and D. J. Begun, 2002 Genomic Effects of Nucleotide Substitutions in *Drosophila simulans*. Genetics **162:** 1753–1761.

Kim, Y. and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167:** 1513–1524.

Kim, Y. and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415–1427.

Kim, Y. and W. Stephan, 2002 Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. Genetics **160:** 765–777.

Kim, Y. and W. Stephan, 2003 Selective Sweeps in the Presence of Interference Among Partially Linked Loci. Genetics **164:** 389–398.

Kimura, M., 1955 Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA **41:** 144–150.

Kimura, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat. **28:** 882–901.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge, U.K.: Cambridge University Press.

Kimura, M. and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763–771.

Kirkpatrick, M. and S. L. Nuismer, 2004 Sexual selection can constrain sympatric speciation. P. Roy. Soc. Lond. B **271:** 687–693.

KONDRASHOV, A. S. and F. A. KONDRASHOV, 1999 Interactions among quantitative traits in the course of sympatric speciation. Nature **400:** 351–354.

KOPP, M. and J. HERMISSON, 2006 The evolution of genetic architecture under frequency-dependent disruptive selection. Evolution**:** in press.

LANDE, R. and S. J. ARNOLD, 1983 The measurement of selection on correlated characters. Evolution **37:** 1210–1226.

LI, H. and W. STEPHAN, 2006 The rate and strength of fitness effects among recent adaptive substitutions in Drosophila. in preparation.

LYNCH, M. and J. B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sunderland: Sinauer.

MATESSI, C., A. GIMELFARB, and S. GAVRILETS, 2001 Long-term buildup of reproductive isolation promoted by disruptive selection: how far does it go? Selection **2:** 41–64.

MAYNARD SMITH, J. and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet. Res., Camb. **23:** 23–35.

MAYR, E., 1942 *Systematics and the origin of species*. New York: Columbia Universtiy Press.

MCVEAN, G. A. T., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162:** 987–991.

NIELSEN, R., S. WILLIAMSON, Y. KIM, M. HUBISZ, A. CLARK, and C. BUSTAMANTE, 2005 Genomic scans for selective sweeps using SNP data. Genome Research **15:** 1566–1575.

OLSEN, K. and M. PURUGGANAN, 2002 Molecular evidence on the origin and evolution of glutinous rice. Genetics **162:** 941–950.

OMETTO, L., S. GLINKA, D. D. LORENZO, and W. STEPHAN, 2005 Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol. Biol. Evol. **22:** 2119–2130.

ORR, H. A., 1991 A test of Fisher's theory of dominance. Proc. Natl. Acad. Sci. USA **88:** 11413–11415.

ORR, H. A. and A. J. BETANCOURT, 2001 Haldane's Sieve and Adaptation From the Standing Genetic Variation. Genetics **157:** 875–884.

OTTO, S. and M. C. WHITLOCK, 1997 The Probability of Fixation in Populations of Changing Size. Genetics **146:** 723–733.

PENNINGS, P. and J. HERMISSON, 2006 Soft Sweeps II: Molecular Population Genetics of Adaptation from Recurrent Mutation or Migration. MBE **23:** 1076–1084.

POLECHOVÁ, J. and N. H. BARTON, 2005 Speciation through competition: a critical review. Evolution **59:** 1194–1210.

PRZEWORSKI, M., 2002 The Signature of Positive Selection at Randomly Chosen Loci. Genetics **160:** 1179–1189.

PRZEWORSKI, M., G. COOP, and J. D. WALL, 2005 The signature of positive selection on standing genetic variation. Evolution **59:** 2312–2323.

ROPER, C., R. PEARCE, B. BREDENKAMP, J. GUMEDE, C. DRAKELEY, F. MOSHA, D. CHANDRAMOHAN, and B. SHARP, 2003 Antifolate antimalarial resistance in southeast Africa: A population-based analysis. The Lancet **361:** 1174–1181.

ROPER, C., R. PEARCE, S. NAIR, B. SHARP, F. NOSTEN, and T. ANDERSON, 2004 Intercontinental spread of pyrimethamine-resistant malaria. Science **305:** 1124.

ROSENZWEIG, M. L., 1978 Competitive speciation. Biol. J. Linn. Soc. **10:** 275–289.

ROUGHGARDEN, J., 1972 Evolution of niche width. Am. Nat. **106:** 683–718.

SABETI, P. C., D. REICH, J. M. HIGGINS, H. Z. P. LEVINE, D. J. RICHTER, S. F. SCHAFFNER, S. B. GABRIEL, J. V. PLATKO, N. J. PATTERSON, G. J. MCDONALD, H. C. ACKERMAN, S. J. CAMPBELL, D. ALTSHULER, R. COOPER, D. KWIATKOWSKI, R. WARD, and E. S. LANDER, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

SANTIAGO, E. and A. CABALLERO, 2005 Variation after a selective sweep in a subdivided population. Genetics **169:** 475–483.

Savolainen, V., M. C. Anstett, C. Lexer, I. Hutton, J. J. Clarkson, M. V. Norup, M. P. Powell, D. Springate, N. Salamin, and W. J. Baker, 2006 Sympatric speciation in palms on an oceanic island. Nature **441:** 210–213.

Schlenke, T. A. and D. J. Begun, 2005 Linkage Disequilibrium and recent selection at three immunity receptor loci in *Drosophila simulans*. Genetics **169:** 2013–2022.

Schlenke, T. B. and D. J. Begun, 2004 Strong selective sweep associated with transposon insertion in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **101:** 1626–1631.

Schliewen, U. K., D. Tautz, and S. Pä"abo, 1994 Sympatric speciation suggested by monophyly of crater lake cichlids. Nature **368:** 629–632.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, and T. Mitchell-Olds, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601–1615.

Schneider, K., 2005 Competitive divergence in non-random mating populations. Theor. Pop. Biol. **68:** 105–118.

Shimizu, K., J. Cork, A. Caicedo, C. Mays, R. Moore, K. Olsen, S. Ruzsa, G. Coop, C. Bustamante, P. Awadalla, and M. Purugganan, 2004 Darwinian selection on a selfing locus. Science **306:** 2081–2084.

Stephan, W., Y. S. Song, and C. H. Langley, 2006 The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci. Genetics **172:** 2647–2663.

Stephan, W., T. Wiehe, and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor. Pop. Biol. **41:** 237–254.

Steppan, S. J., P. C. Phillips, and D. Houle, 2002 Comparative quantitative genetics: evolution of the *G* matrix. TREE **17:** 320–327.

Storz, J. F., B. A. Payseur, and M. W. Nachman, 2004 Genome scans of DNA variability in humans reveal evidence for selective sweeps outside Africa. Mol. Biol. Evol. **21:** 1800–1811.

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

Takahashi, A., S. C. Tsaur, J. A. Coyne, and C.-I. Wu, 2001 The nucleotide changes governing cuticular hydrocarbon variation and their evolution in Drosophila melanogaster. PNAS **98:** 3920–3925.

Talisuna, A. O., P. Bloland, and U. DÁlessandro, 2004 History, Dynamics, and Public Health Importance of Malaria Parasite Resistance. Clinical Microbiology Reviews **17:** 235–254.

Teshima, K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res. **16:** 702–712.

van Rheede, T., 2003 Some histories of molecular evolution: amniote phylogeny, vertebrate eye lens evolution, and the prion gene. PhD dissertation, Nijmegen University.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biology **4:** 0446–0458.

Wall, J. D. and R. R. Hudson, 2001 Coalescent simulations and statistical tests of neutrality. MBE **18:** 1134–1135.

Wang, X., W. E. Grus, and J. Zhang, 2006 Gene losses during human origins. PLOS Biology **4:** 0366–0377.

Waxman, D. and S. Gavrilets, 2005 Issues of terminology, gradient dynamics and the ease of sympatric speciation in adaptive dynamics. J. Evol. Biol. **18:** 1214–1219.

Xue, Y., A. Daly, B. Yngvadottir, M. Liu, G. Coop, Y. Kim, P. Sabeti, Y. Chen, J. Stalker, E. Huckle, J. Burton, S. Leonard, J. Rogers, and C. Tyler-Smith, 2006 Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am. J. Hum. Genet. **78:** 659–670.

CV

Pleuni was born and raised in Castricum, the Netherlands. From a very early age it was clear that she had a mind of her own, as evidenced by the big curl that kept poking up through her little woolly hat. Her neverending amount of energy, ideas and creativity was also present from early on. As a child, she was always drawing, doing handicraft or performing stories she and her sisters and friends made up. She was definitely the silliest performer of the bunch. She also learned to play the violin, not without merit, and thoroughly enjoyed making music with others. Later in life, she played in (and even set up) different orchestras and bands. Her sports addiction started with hockey; she has played several times a week for most of her life and even moving to Scotland, France and Germany has not kept her from hitting the ball regularly. 1975

She attended secondary school in Alkmaar, cycling a total of 20 kilometers to and from school every day. She had clear ideas about teaching and especially did not agree with the way in which mathematics was taught. After secondary school, however, she ended up studying mathematics (and theoretical physics) anyway for a year in Aberdeen, Scotland. One of her friends there studied zoology and this triggered her interest in biology. 1987

Thus it happened that one happy day in 1994 she started to study Biology at the University of Amsterdam. She found a home away from home at Anna's Hoeve, where the Biology Department was situated. A real home was more difficult to find in Amsterdam, so the next years she moved around a lot and lived in as many as ten appartments (not counting the one in Paris). At Annas Hoeve she met a lot of interesting people and made friends for life. She became an active member of Congo, the biology student society at the University of Amsterdam, where she organised different activities and of which she was president for one year. She developed a love for evolutionary ecology, research and hitch-hiking. After winning the Congo hitch-hiking competition for three consecutive years, she graciously bowed out to allow others the taste of victory 1994 as well. Her fondness of teaching started when she was given the opportunity to teach the first-year mathematics course as a "student assisten".

1999 After five years in Amsterdam, in 1999 the outside world called again, this time *en français*. She did two research projects in France. First, she went to Poitiers for five months to study *Wolbachia* in woodlice (which drew some disgusted looks from her family and friends). In 2000 she lived in Paris for ten months and participated in different projects on host-parasite interactions at the University of Paris-Sud. For this work the Amsterdam Biology Department rewarded her the prize for the best student research project.

1999 The year 1999 marked the beginning of another great adventure. Pleuni and three friends started a company to promote and support science education in secondary schools. Their first product was a box with supplies for a small science project that could be used directly in classrooms. Today, De Praktijk has a broad spectrum of activities and products. It is a small but steady company. In 2003, Pleuni decided to leave the company to pursue a scientific career.

2000 At the end of 2000 she received her Master of Science degree in Amsterdam.

2003 In 2003 Pleuni had a sudden craving for Lederhosen and left the Netherlands again, this time for Munich, Germany, to do a PhD in theoretical evolutionary biology. Under supervision of Dr. Joachim Hermisson she studied the way in which populations adapt to new environments, which resulted in the book you are reading now. It is a nice coincidence that a large part of her work focusses on the effect of "genetic hitch-hiking". In 2005 she received a grant from the Dutch Science Foundation to spend three months in Vienna. Whilst there, she proved herself able to adapt to new environments extremely well by falling in love with a German, Andi (who by the way has not yet been spotted in Lederhosen). In Munich, she developed and taught a new course on population genetics. Her sports addiction got completely out of hand, as evidenced by the fact that she plays hockey, swims and cycles and goes skiing, hiking or canoeing on weekends.

Near future In the near future she will continue to work in Munich, partly in research, partly as a coordinator of the Munich Graduate School for Evolution, Ecology and Systematics.

More distant future In the more distant future, the pull of Dutch cheese and especially peanut butter will prove too strong. Pleuni will convince Andi to move to Amsterdam, will buy a luxury house on one of the canals and will successfully combine careers in science, education, politics, business and art whilst raising three accomplished sons, maintaining a happy relationship, leading a healthy social life and not neglecting her hockey and violin.

Noor Pennings and Evelien Pennings

List of publications

T. Rigaud, **P.S. Pennings**, P. Juchault
Wolbachia bacteria effects after experimental interspecific transfers in terrestrial isopods.
Journal of Invertebrate Pathology (2001) 77: (4) 251-257.

C.L. Collin, **P.S. Pennings**, C. Rueffler, A. Widmer and J.A. Shykoff
Natural enemies and sex: How seed predators and pathogens contribute to sex-differential reproductive success in a gynodioecious plant.
Oecologia (2002) 131:94-102.

J. Hermisson and **P.S. Pennings**
Soft Sweeps – Molecular population genetics of adaptation from standing genetic variation.
Genetics (2005) 169:2335-2352.

P.S. Pennings and J.Hermisson
Soft Sweeps II – Molecular population genetics of adaptation from recurrent mutation or migration.
Molecular Biolology and Evolultion (2006) 23:1076-1084.

P.S. Pennings and J.Hermisson
Soft Sweeps III – The signature of positive selection from recurrent mutation.
PLoS Genetics (2006) 12:e186

D. Sicard, **P.S. Pennings**, C. Grandclément, J. Acosta, O. Kaltz, J. Shykoff
Specialization and local adaptation for two fitness traits of a fungal parasite on two host plant species.
Accepted for publication in Evolution (2007)

P.S. Pennings, M. Kopp, G. Meszéna, U. Dieckmann and J.Hermisson
A one-locus model of sympatric speciation.
Manuscript in preparation.

Acknowledgments

First of all, I owe a big thanks to Joachim Hermisson. Joachim is the best supervisor I could imagine. He is intelligent, open, patient, and a perfectionist. The first three things made it fun to work with him. The last made me improve my work when I thought it was already good enough. I am convinced that in less than ten years' time Joachim will be one of the big guys in evolutionary biology, and I am proud to have been his first PhD student. I look forward to continuing our collaboration not only to produce Soft Sweeps 4, 5, 6, 7 and 8, but also to revolutionize graduate education in Munich.

My good friend Pieter van Beek and my boyfriend Andreas Gros have helped many many times with all my computer-related problems, and questions about linux, C++, shell scripts, clusters, LaTeX, Mathematica etc... They are the perfect teachers. Often, as a response to one of my many questions, they would give me not only the immediate solution that I was looking for, but also some (understandable!) background information and a hint about where to find more information. And even if I was too lazy to learn and came back with the same question a week later, they would answer me patiently again. Other people also helped with similar things. I want to mention Thijs Brouwer, Stephan Hutter, Andreas Lehnert and Haipeng Li. Yuseob Kim allowed me to use his program for chapter 3 and helped me to change it for my purposes.

I enjoyed working in Joachim's group where I had great colleagues. With José Alvarez-Castro, I enjoyed the trip to the Statistical Genetics summerschool in North Carolina. With Nina Stoletzki I made nice trips to the BBSRC summerschool in Nottingham and the MBE meeting in Arizona. I especially thank Michael Kopp for "jumping" into the speciation project and working really hard to get it finished in time. I want to thank the (former) members of the Wolfgang Stephan group for many discussions and for introducing me to applied population genetics. Especially Steffen Beisswanger, Sascha Glinka, David de Lorenzo, Lino Ometto, Daven Presgraves and Laura Rose. I am happy to have Peter Pfaffelhuber as a colleague. I hope we will continue organizing seminars and courses together (and some time write a paper!). Outside the Munich group, I benefited from discussions with Patrick Meirmans, Géza Meszéna, Tim van Opijnen and Allen Orr. Back in Amsterdam, the courses by Rob Lingeman, Maurice Sabelis, Andre de Roos and Martijn Egas showed me how interesting and fun theoretical biology can be.

When biologists hear the word *self-organization* they think of flocking behaviour in birds orso. It is unfortunate that scientists don't practice self-organization much. I was happy to find some people that were interested in setting up a PhD network and I am convinced that this network (called Volvox) will improve the life and work of PhD students and their supervisors in the evolution, ecology and systematics groups of our department. It would not have worked without Lino Ometto, Rita Verma, Sarah Peter and Steffen Beisswanger.

Clear writing is greatly helped by critical readers. Before this thesis was printed, parts of it were read and criticized by many people. Among them are great scientists, great friends and great parents: Pieter van Beek, Marian Dekker, Sascha Glinka, Andreas Gros, Haipeng Li, Sally Otto, John Parsch, Cees Pennings, Evelien Pennings, Peter Pfaffelhuber, Laura Rose, Saskia Stehouwer, Wolfgang Stephan, Marcy Uyenoyama, Alex Verkade, and various anonymous reviewers. Thanks!

My sisters Noor and Evelien wrote my CV and Andi helped with the cover design. Dank jullie wel!

Pleuni
August 2006.