Detecting selective sweeps in natural populations of *Drosophila melanogaster*

Methods, Applications, and Extensions

Pavlos Pavlidis

München 2010

Detecting selective sweeps in natural populations of *Drosophila melanogaster*

Methods, Applications, and Extensions

Pavlos Pavlidis

Dissertation
an der Fakultät für Biologie
der Ludwig–Maximilians–Universität
München

vorgelegt von
Pavlos Pavlidis
aus Athen, Griechenland

München, den 07.12.2010

Erklärung

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Herrn Prof. Dr. Wolfgang Stephan betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

Ehrenwörtliche Versicherung

Ich versichere hiermit ehrenwörtlich, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt wurde.

München, 07.12.2010

Erstgutachter: Prof. Dr. Wolfgang Stephan

Zweitgutachter: Prof. Dr. Dirk Metzler

Tag der Abgabe: 07.12.2010

Tag der mündlichen Prüfung: 14.02.2011

Contents

Summary **xii**

Zusammenfassung **xv**

Declaration of Contributions as a co-author **xvii**

Acknowledgements **xix**

General Introduction **1**

Methods for detecting selective sweeps 2

Subgenomic data 3

Genome-wide data 8

Methods for detecting selection based on genetic differentiation between populations 10

Methods based on the machine learning paradigm 12

Aims of the thesis 12

1 Searching for footprints of positive selection in whole-genome SNP data from non-equilibrium populations **15**

1.1 Abstract 15

1.2 Introduction 16

1.3 Methods 17

1.4 Theoretical analyses 23

1.5 Statistical performance of the tests in the detection of single hitchhiking events . . 25

1.6 Discussion 36

2 Recent strong positive selection on *Drosophila melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by population genomic analysis **43**

2.1 Abstract 43

2.2 Introduction 44

2.3 Materials and Methods 45

2.4 Results 49

2.5 Discussion 54

3 Selective sweeps in multi-locus models **59**
3.1 Abstract . . . 59
3.2 Introduction . . . 60
3.3 Methods . . . 64
3.3.1 The model . . . 64
3.3.2 Summary statistics of the coalescent and SNP polymorphisms . . . 66
3.4 Implementation . . . 67
3.5 Results . . . 67
3.5.1 Trajectories of new variants . . . 67
3.5.2 Coalescent simulations conditioning on the trajectory of the A_{11} allele . . . 84
3.6 Discussion . . . 86

General Discussion **93**

Bibliography **xxi**

Appendix **xxxiii**

Curriculum vitae **xliii**

List of Figures

1 Coalescent trees on a recombining genomic region 6

1.1 Histogram of the ratio $f = L_n/H_n$ for various demographic scenarios 24
1.2 The relation between Λ_{MAX}, the percentage of star-like genealogies, and the number of segregating sites 26
1.3 A spatial modification of the SFS caused by a selective sweep 27
1.4 The joint distributions of Λ_{MAX} and ω_{MAX} in scenarios with and without selection . 30
1.5 The distributions of Λ_{MAX} for various levels of the decrease of heterozygosity . . . 40

2.1 The demographic model of the European and African population of *D. melanogaster* 48
2.2 Nucleotide diversity of *D. melanogaster* and divergence to *D. simulans* 50
2.3 The likelihood-ratio values calculated by the *CLR* and the *SweepFinder* tests 52
2.4 Sliding window analysis of the fully sequenced 22-kb region 53

3.1 The tetrahedron that represents the state space for the two-locus two-allele model . 62
3.2 Trajectories of A_{11} obtained under the deterministic two-locus two-allele model with symmetrical fitness matrix 69
3.3 Empirical cumulative distribution of the allelic frequencies for the deterministic two-locus two-allele model 70
3.4 The role of parameters c_1 and c_2 on determining the class of the trajectory for the deterministic two-locus two-alleles model 71
3.5 The role of parameters $p_0(A_{11})$ and $p_0(A_{21})$ on determining the class of the trajectory for the comparison of the fixation class versus the polymorphic class 72
3.6 The role of parameter c_1 and c_2 on determining the class of the trajectory when class 1 and class 0 represent fixation and extinction of A_{11} allele, respectively . . . 73
3.7 The role of parameters $p_0(A_{11})$, $p_0(A_{21})$ and allele contributions w_{11} and w_{21} on determining the class of the trajectory for the comparison of the fixation class versus the extinction class 74
3.8 The empirical cumulative distribution for the equilibrium frequency of the stochastic two-locus two-allele model 75
3.9 Examples of trajectories in the stochastic two-locus two-allele model 75
3.10 The relation of six parameters to the class of the trajectory 76
3.11 The relation of six parameters to the class of the trajectory for the comparison of the fixation class versus the extinction class 78

3.12 Empirical cumulative distribution of the frequencies of the trajectories for the deterministic two-locus two-allele model assuming a generalized fitness matrix 79
3.13 The roles of parameters c_1 and c_2 in the deterministic two-locus two-allele model with generalized fitness matrix for the comparison of the fixation class versus the polymorphic class 80
3.14 The effect of the initial trait value on the fate of the trajectory 81
3.15 The empirical cumulative distribution of the five-locus two-allele model with symmetrical fitness matrix 82
3.16 The role of various model parameters in determining the class of the trajectory for the comparison of the fixation class versus the polymorphic class 83
3.17 The growth rates for the $A_{11}A_{21}$ and $A_{11}A_{22}$ 84
3.18 Summary statistics for the coalescent trees as a function of the distance from locus A_1 87
3.19 Sliding window analysis of a 500-kb genomic region flanking the locus A_1 88
1 The distance between a peak of the landscape of the statistic and the selective sweep locations. xxxv
2 The 95^{th} percentile for the ω-statistic and *SweepFinder*. xxxvi
3 The 95^{th} percentile for the ω_{MAX} and Λ_{MAX}. xxxvii
4 Comparisons between recurrent selective sweeps for various levels of reduction of heterozygosity and selection coefficient. xxxviii
5 The fraction of predicted targets within 5 kb from the true location of the selective sweep for a recurrent selective sweep. xxxix
6 The expected fraction of peaks whose distance from the randomized ‘target of selection’ is smaller than 5 kb. xl
7 Variable values of recombination rate affect the ω_{MAX} and Λ_{MAX} values. xli
8 The likelihood curves for each polymorphism class. xlii

List of Tables

1.1 Equilibrium neutrality versus selection in equilibrium populations 29
1.2 Non-equilibrium neutrality versus selection in equilibrium populations 31
1.3 Neutrality versus selection in non-equilibrium populations (deep bottlenecks) . . . 32
1.4 Neutrality versus selection in non-equilibrium populations (shallow bottlenecks) . . 34

3.1 Genotypic values and fitnesses for the symmetric fitness model 63
3.2 The parameter values that were used for the simulations of the two-locus two-allele model 68
1 The matrix used for the pre-calculation of the ω-statistic for all possible configurations xxxiv

To Toma, Alexandra, and Constantinos…

Ζηλεὺω του σταυραετοὺ αποὺ πετὰ στα νὲφη,
και παὶζει με τσι αστραπὲς και με τ΄ αστροπελὲκι.
Στο βρὰχο χτὶζει τη φωλιὰ στο χιὸνι ζευγαρὼνει...

Ψαραντὼνης

Summary

The goal of this study was to gain a deeper understanding of the selective sweep models and the statistical and computational methods that disentangle selective sweeps from neutrality. In the Introduction of the thesis I review the literature on the main approaches that have been developed in the last decade to separate selective sweeps from neutral demographic scenarios. Methods on complete and incomplete selective sweeps are reviewed as well as selective sweeps on structured populations. Further, I analyze the effects of past demographic events, especially bottlenecks, on the genealogies of a sample. Finally, I demonstrate that the ineffectiveness of separating selective sweeps from bottlenecks stems from the lack of robust statistics, and most importantly from the similar genealogies that bottlenecks and selective sweeps may generate locally on a recombining chromosome.

In the first chapter I introduce a method that combines statistical tests in a machine learning framework, in order to disentangle selective sweeps from neutral demographic scenarios. The approach uses support vector machines to learn examples from neutral scenarios and scenarios with selection. I demonstrate that the novel approach outperforms previously published approaches for a variety of demographic scenarios. The main reason for the performance difference is the usage of the scenarios with selection, that are not analyzed by classical statistical methods.

In the second chapter of the thesis I present an application of the methods on detecting a selective sweep in the African population of *D. melanogaster*. Demographic history and ascertainment bias schemes have been taken into account. Results pinpoint to the *HDAC6* gene as a target of recent positive selection. This study demonstrates the variable threshold approach, which remedies the tendency of some neutrality tests to detect selective sweeps at the edges of the region of interest.

In the third chapter I present the results of the analysis of selective sweeps in multi-locus models. I assume that a phenotypic trait evolves under stabilizing or directional selection. In contrast to the classical models of selective sweeps, the evolutionary trajectory of an allele that affects the trait might belong to one of the three categories: it either fixes, disappears or remains

polymorphic. Thereafter, I analyze the properties of coalescent trees and neutral polymorphism patterns that are generated from each of the three categories. I show that for the majority of simulated datasets selection cannot be detected unless the trajectory is either fixed or close to fixation.

Zusammenfassung

Das Ziel dieser Studie ist ein besseres Verständnis von ‘Selective Sweep’ Modellen zu erhalten, sowie den statistischen und computerbasierten Methoden die versuchen ‘Selective Sweeps’ von neutraler Evolution abzugrenzen. In der Einleitung gebe ich einen Überblick über die Literatur der letzten zehn Jahre die Versuche beschreibt ‘Selective Sweeps’ von neutralen demographischen Szenarien zu unterscheiden. Methoden für vollständige und unvollständige ‘Sweeps’ werden besprochen, als auch Methoden für ‘Sweeps’ in strukturierten Populationen. Ich analysiere die Effekte von vergangenen demographischen Ereignissen, insbesondere genetischer Flaschenhälse, auf die Genealogie von Stichproben aus einer Population. Ich zeige auf, dass die Ineffektivität in der Unterscheidung von ‘Selective Sweeps’ und Flaschenhälsen auf einen Mangel an robusten Statistiken zurckzufhren ist, sowie der Tatsache, dass Flaschenhälse und ‘Selective Sweeps’ lokal auf einem Chromosom ähnliche Genealogien erzeugen können.

Im ersten Kapitel stelle ich eine neue Methode zur Unterscheidung von ‘Selective Sweeps’ und neutralen demographischen Szenarien vor, die maschinelles Lernen benutzt um statistische Tests zu kombinieren. Dieser Ansatz benutzt ‘Support Vector Machines’ um Beispiele von neutralen Szenarien sowie Szenarien mit Selektion zu erlernen. Ich zeige, dass dieser neue Ansatz den bisher veröffentlichten Methoden unter einer Vielzahl demographischer Szenarien überlegen ist. Der Hauptgrund für diesen Leistungsunterschied liegt im Gebrauch von Szenarien mit Selektion, welche in klassischen statistischen Methoden nicht berücksichtigt wurden.

Das zweite Kapitel beschreibt die Anwendung von Methoden zum Nachweis von ‘Selective Sweeps’ auf eine Afrikanische Population von *Drosophila melanogaster*. Die demographische Vergangenheit der Population und mögliche statistische Verzerrungen wurden dabei berücksichtigt. Die Ergebnisse deuten darauf hin, dass das Gen *HDAC6* vor kurzem Ziel von positiver Selektion war. Diese Studie benutzt eine Herangehensweise mit variablem statistischem Schwellenwert, welche die Tendenz einiger Neutralitätstests umgeht ‘Selective Sweeps’ an den Rändern von untersuchten Regionen zu entdecken.

Im dritten Kapitel präsentiere ich die Ergebnisse aus einer Analyse von ‘Selective Sweeps’

unter Multi-Lokusmodellen. Ich nehme hierbei an, dass ein Phänotyp unter dem Einfluss von stabilisierender oder gerichteter Selektion evolviert. Im Gegensatz zu klassischen Modellen von ‘Selective Sweeps’, fällt der evolutionäre Verlauf eines Allels das den Phänotyp beeinflusst in eine von drei Kategorien: es fixiert, es geht verloren oder es bleibt polymorph. Des Weiteren untersuche ich die Eigenschaften der Koaleszenzbäume und der neutralen Polymorphismen welche unter den jeweiligen drei Szenarien entstehen. Ich zeige, dass für die Mehrzahl der simulierten Datensätze Selektion nicht nachweisbar ist, außer das Allel ist fixiert oder steht kurz davor.

Declaration of Contributions as a co-author

This thesis was written using the software package LaTeX and is based on the LaTeX template for dissertations developed at the Ludwig-Maximilians-Universität (LMU) by Robert Dahlke and Sigmund Stintzing, 2002 (`http://edoc.ub.uni-muenchen.de/hinweise/LaTeXVorlage.zip`). In this dissertation I present my doctoral research from June 2007 until December 2010. It is organized in three chapters. A large part of the research has been conducted in collaboration with other scientists.

In Chapter 1, I developed the method, implemented the software and performed the analysis. The study was supervised by Prof. Dr. Wolfgang Stephan. The manuscript was written by me, Dr. Jeffrey Jensen (UMass), and Prof. Dr. Wolfgang Stephan. An article that describes the results of the research has been published in *Genetics* (PAVLIDIS *et al.* 2010).

In Chapter 2, Dr. Nicolas Svetec (University of Toronto), generated the data and performed part of the analysis. I applied advanced statistical approaches, and wrote part of the manuscript. The research was supervised by Prof. Dr. Wolfgang Stephan. An article that presents the results of the research has been published in *Molecular Biology and Evolution* (SVETEC *et al.* 2009).

Chapter 3 is an ongoing project. I developed the methods, implemented the software and wrote the manuscript. Prof. Dr. Dirk Metzler contributed ideas and software for the initial steps of the research. The work was supervised by Prof. Dr. Wolfgang Stephan.

A large part of the Introduction has been published in *Molecular Ecology* (PAVLIDIS *et al.* 2008). Dr. Stephan Hutter and Prof. Dr. Wolfgang Stephan contributed to this publication.

Pavlos Pavlidis Prof. Dr. Wolfgang Stephan

Acknowledgements

This work was carried out at the Department of Biology II, Ludwig-Maximilians-Universität, Munich, Germany. I would like to express my gratitude to many excellent people I was lucky to work with. First of all, I would like to thank my supervisor, Prof. Dr. Wolfgang Stephan for his enthusiasm and believe in my work. His thoroughness and patience from the initial to the final level enabled me to develop an understanding of the subject and made this thesis a reality. His support in difficult moments helped me not to give up.

I am grateful to Prof. Dr. Dirk Metzler for the help on statistical and programming problems. His comments improved the quality of the manuscripts that were published during my Ph.D. studies. It was a pleasure to collaborate with Prof. Dr. John Parsch and his student Sarah Saminadin-Peter on one of their projects. I am thankful to Prof. Dr. John Baines for collaborating on detecting selective sweeps in mice. Prof. Dr. Peter Pfaffelhuber and Dr. Pleuni Pennings gave insightful answers to my many questions in the beginning of my doctoral studies.

Many thanks to co-authors and collaborators from other Universities. Prof. Dr. Kentaro Shimizu, Dr. Takashi Tsuchimatsu (University of Zurich, Switzerland) and Dr. Thomas Städler (ETH, Zurich) gave me the opportunity to contribute important results regarding the evolution of self-compatibility in *Arabidopsis*. With Prof. Tõnis Timmusk (Tallinn University of Technology, Estonia) and his group we contributed to the findings on the regulation of the *BDNF* gene. It is my pleasure to thank Dr. Jeffrey Jensen (University of Massachusetts) for collaborations and friendship as well as Sven Puppel (MPI for Evolutionary Anthropology, Leipzig) and Johannes Engelken (Universitat Pompeu Fabra, Barcelona).

I would like to express my gratitude to all colleagues and lab-mates. Many thanks to Aurélien Tellier who was sharing office with me; to Annegret Werzner, Stefan Laurent, Stephan Hutter, Nicolas Svetec, Robert Piskol, Daniel Zivkovic, Pablo Duchen and the other members of the *Drosophila* group for invaluable discussions and support. Many thanks to Lisha Naduvilezhath and Martin Hutzenthaler from the group of Prof. Metzler and to Claus Kemkemer, Sonja Grath and Miri Linnenbrink from the group of Prof. Parsch. It is a pleasure to recall wining and joyful

discussions with Elena Berg. Their friendship and support helped me to overcome many difficulties during the three years of doctoral studies. I would like especially to thank Justyna Wolinska for an honest friendship since the first moment we met at the Volkswagen symposium at Potsdam. It is an honor to acknowledge Alexis Stamatakis and George Themelis because they were always there to help when it was needed.

I would like to show my gratitude to the Volkswagen Foundation, and especially to Dr. Henrike Hartmann, for financial support during my Ph.D. studies, and for organizing the annual symposia on Evolutionary Biology. This dissertation was supported by grants from the Volkswagen Foundation (grant I/82770).

I want to thank my wife, Tamara Aid-Pavlidis, who motivated me to pursue the doctoral degree, not to give up, but also to respect and love myself. Working with her on the microarray data analysis was a lot of fun, and I hope we will continue to generate discoveries (and children). Also, my parents, my brother and my sister were of a great support during all the years of studies. Finally, I am grateful to my children, Alexandra and Constantinos, who were simultaneously a challenge and a motivation for me.

General Introduction

Based on nearly complete genome sequences from a variety of organisms data on naturally occurring genetic variation on the scale from hundreds of loci to entire genomes have been collected in recent years. In parallel, new statistical tests have been developed to infer evidence of recent positive selection from these data and to localize the target regions of selection in the genome. These methods have now been successfully applied to *Drosophila melanogaster*, human, mouse and a few plant species. In genomic regions of normal recombination rates, the targets of positive selection have been mapped down to the level of individual genes.

Searching for strong positive selection in the genomes of individuals of a natural population has been the focus of a multitude of studies over the past ten years HARR *et al.* (2002); KIM and STEPHAN (2002); GLINKA *et al.* (2003); AKEY *et al.* (2004); ORENGO and AGUADÉ (2004). The goals of these studies have been (i) to provide evidence of positive selection, (ii) estimate the strength of selection, and (iii) localize the targets of selection. A long-term goal is that the genes that experienced recent, strongly positive selection could be identified and the associated functions and phenotypes characterized.

In general, these studies followed a two-tier approach: at first, levels of DNA polymorphism are measured for a very large number of loci on a genome-wide scale within populations. For humans, the best-studied species, continuous single nucleotide polymorphism (SNP) data are also available along the entire genome, though with some varying density. The goal of this initial step is to identify loci that display patterns of variability suggesting recent positive selection. Some studies employed microsatellite markers to measure polymorphism and looked for regions of depleted variability as an indicator of a selective sweep due to genetic hitchhiking in the region (see Box on page 2). Other studies analyzed SNP by directly sequencing small fragments of DNA at multiple loci, which allows for the estimation of properties of the site frequency spectrum (SFS) of SNPs and linkage disequilibrium (LD). While this approach might seem straight forward, the actual definition of a candidate locus can be challenging, especially in populations that have undergone demographic perturbations. Most studies up to now have employed rather simple methods such

as outlier analysis, in order to select candidate loci (*e.g.* KAUER *et al.* 2003; OMETTO *et al.* 2005). Only recently more sophisticated methods have been developed for analyzing genome-wide polymorphism data, including tests based on the background SFS (NIELSEN *et al.* 2005), F_{ST} (BEAUMONT and BALDING 2004; RIEBLER *et al.* 2008) and explicit modeling of the population history (LI and STEPHAN 2006).

As a second step following the identification of a candidate locus, polymorphism patterns of the surrounding region are obtained by fine-scale sequencing. The resulting high-density SNP data is then used for tests of deviation from neutral expectations [including the standard tests of HUDSON *et al.* (1987); TAJIMA (1989) and FAY and WU (2000)]. In addition, however, specific tests for positive selection in these subgenomic regions such as the *CLR-GOF* KIM and STEPHAN (2002); JENSEN *et al.* (2005) tests are used; they can also estimate the strength of selection and the approximate location of the beneficial mutation within the region. Below, we describe these new tests and show that they have been successfully used to identify the targets of recent, strongly positive selection. If the rate of local recombination is not too low, individual genes or even regions within a gene can be mapped using this approach.

The hitchhiking effect

When a strongly beneficial mutation occurs and spreads in a population, it is inevitable that the frequency of linked neutral (or weakly selected) variants increases. In a seminal paper, MAYNARD SMITH and HAIGH (1974) described this process, which they termed genetic hitchhiking. They show that in very large populations hitchhiking can drastically reduce genetic variation near the site of selection (thus causing a selective sweep).
According to Maynard Smith and Haigh's deterministic model, in recombining chromosomal regions diversity vanishes at the site of selection immediately after the fixation of the beneficial allele and is predicted to increase as a function of the distance to the selected site (scaled by the selection coefficient). This result is also roughly correct in finite populations (KAPLAN *et al.* 1989; STEPHAN *et al.* 1992). Further signatures of the hitchhiking effect include (i) shifts in the site frequency spectrum of polymorphisms such as an excess of low- and high-frequency derived alleles (BRAVERMAN *et al.* 1995; FAY and WU 2000), and (ii) distinct patterns of linkage disequilibrium such as an elevated level of LD in the early phase of the fixation process of a beneficial mutation (KIM and NIELSEN 2004; STEPHAN *et al.* 2006). In a suite of statistical tests, these properties of the hitchhiking effect have been used to map recent, strongly positive directional selection along recombining chromosomes of several species.

Methods for detecting selective sweeps

Subgenomic data

CLR test: Using predictions of the hitchhiking model (MAYNARD SMITH and HAIGH 1974, see Box on page 2), KIM and STEPHAN (2002) developed a composite-likelihood ratio (*CLR*) test to detect local reductions of nucleotide variation along a recombining chromosome and to predict the strength and the location of a selective sweep. The *CLR* test compares the probability of the observed polymorphism data under the standard neutral model with the probability of the data under a model of selective sweep. Under the standard neutral model the expected number of sites where the derived variant is in the frequency interval $[p+dp]$ in the population (the SFS) is given by

$$\phi_0(p)dp = \frac{\theta}{p}dp \tag{1}$$

(FU 1995; EWENS 2004). FAY and WU (2000) have shown that immediately after a hitchhiking event this distribution is transformed approximately to

$$\phi_A(p) = \begin{cases} \frac{\theta}{p} - \frac{\theta}{C} & \text{for } 0 < p < C \\ 0 & \text{for } C < p < 1-C \\ \frac{\theta}{C} & \text{for } 1-C < p < 1 \end{cases} \tag{2}$$

where the parameter C depends on the strength of selection $\alpha = 2Ns$ and the recombination rate r between the neutral site and the site where the beneficial mutation has occurred (KIM and STEPHAN 2002). N is the effective population size and s the selection coefficient.

The probability of observing a site where k derived alleles are found in a sample of n sequences is obtained by binomial sampling as

$$P_{n,k} = \binom{n}{k} p^k (1-p)^{n-k} \phi(p)dp, \tag{3}$$

where $\phi(p) = \phi_0(p)$ applies under the standard neutral model and $\phi(p) = \phi_A(p)$ under the hitchhiking model. KIM and STEPHAN (2002) compare two hypotheses:

- (H_0) The observed allelic class at each position of the subgenomic region under consideration is derived from a standard neutral model, and

- (H_A) The observed allelic class at each position of the subgenomic region is due to a selective

sweep that occurred at some position X of the fragment and is characterized by the selection parameter α.

The probabilities of the data under these hypotheses are calculated as the product of the probabilities of all sites of the fragment under consideration (Equation 3) using the densities $\phi_0(p)$ and $\phi_A(p)$, respectively. The maximum log-likelihood-ratio statistic Λ_{CLR} is then given by

$$\Lambda_{CLR} = \log \frac{\max P(\text{Data}|H_A)}{P(\text{Data}|H_0)}, \tag{4}$$

where max refers to the maximization of $P(\text{Data}|H_A)$ with respect to the parameters X and α.

Since the null and alternative hypotheses that are compared in the *CLR* test are explicitly modeled, the interpretation of the test results is rather simple. That means that the expectation of the SFS is well formulated under both evolutionary scenarios. On the other hand, it is important to realize that the null hypothesis of the test is based on the standard neutral model. That means that any violation of the assumptions of the null hypothesis may influence the results and favor the alternative hypothesis (JENSEN *et al.* 2005; THORNTON and JENSEN 2007). Therefore, the application of the *CLR* test is not appropriate for detecting selective events when severe demographic events (especially bottlenecks) have occurred in the recent history of the population.

The effect of bottlenecks on the genealogies: Bottleneckes may generate similar polymorphism patterns as the selective sweeps. Therefore, disentangling selective events from neutral demographic events can be challenging. There are two reasons for this problem. First, the statistics that are used in population genetics to summarize a full dataset (alignment of sequences from one or more populations) are not robust to bottlenecks. This has been demonstrated extensively for summary statistics such as Tajima's (1989) D, θ_W, the number of halplotypes, and other classical summaries of the data. The lack of robust summary statistics motivates researchers to develop a test that captures some aspects of selective sweeps that are absent from the polymorphism patterns generated by bottlenecks. As we will demonstrate below, several modern tests have been developed that use extensive modeling or modern techniques from the machine learning field (KIM and NIELSEN 2004; NIELSEN *et al.* 2005; PAVLIDIS *et al.* 2010; LIN *et al.* 2010). Second, to some extent, there are intrinsic reasons for the resemblance of a bottleneck to a selective sweep. Therefore, the resemblance does not depend on the way that the data are summarized. Apparently, the second reason implies that any summary statistic or any test will fail to separate bottlenecks from selective sweeps simply because the data appear to be very similar.

In order to illustrate this problem we plot coalescent trees (genealogies) of a sample of individuals from a selective sweep and from a bottleneck scenario (Figure 1). The coalescent (KINGMAN

1982; HUDSON 1990) employs a sample of individuals from a population to trace all alleles of a gene to a single ancestral copy, known as the most recent common ancestor (MRCA). The inheritance relations between alleles are represented as a gene genealogy, which can be drawn as a binary tree for non-recombining loci. The relations between the gene copies of the sample affect the generated polymorphism patterns on a dataset. In Figure 1 the genealogies of a recombining genomic segment are shown. Due to the presence of recombination more than one coalescent trees are needed to describe the genealogy of the sample. The upper panel illustrates bottleneck genealogies, whereas the bottom panel illustrates selective sweep genealogies. Apparently, as Figure 1 shows, the bottleneck genealogies can be very similar to selective sweep genealogies. Consequently, the polymorphism data generated by these coalescent trees will be similar as well, and it will be nearly impossible to separate the bottleneck-derived dataset from the selective sweep-derived dataset.

In the following sections we describe modern approaches to disentangle selective sweeps from bottlenecks.

Distinguishing between selective sweeps and demography: JENSEN *et al.* (2005) showed that the *CLR* test is not robust in the cases of structured populations or recent bottlenecks. Under these scenarios, the false positive rate may be as high as 80% (JENSEN *et al.* 2005). They proposed a goodness-of-fit (*GOF*) test to distinguish between the true positives that come from the rejection of the standard neutral scenario because of a selective sweep event, and the false positives that come from the rejection of the standard neutral hypothesis due to demographic factors.

The *GOF* test is based on the hypothesis that non-selective evolutionary processes influence the frequency spectrum globally (*e.g.* the whole region under investigation) and not locally as a selective sweep does. This assumption is adopted widely, although it is possible that a recent strong bottleneck combined with recombination may create local patterns that resemble those of a selective sweep (BARTON 1998; THORNTON and JENSEN 2007).

The *GOF* approach tests whether the observed data is drawn from a selective sweep model. Thus, the latter represents the null hypothesis. The alternative hypothesis claims that the data is not drawn from a selective sweep scenario. Thus, for H_A an alternative model is not specified explicitly, except that it is assumed that the evolutionary forces in action affect the whole region under investigation. The likelihood of the alternative model is calculated as

$$\begin{aligned} P(\text{Data}|H_A) &= \prod_{i=1}^{l} P(Y = y_i|H_A) \\ &= \prod_{i=1}^{l} \binom{n}{y_i} P_i^{y_i}(1-p_i)^{n-y_i} \end{aligned} \tag{5}$$

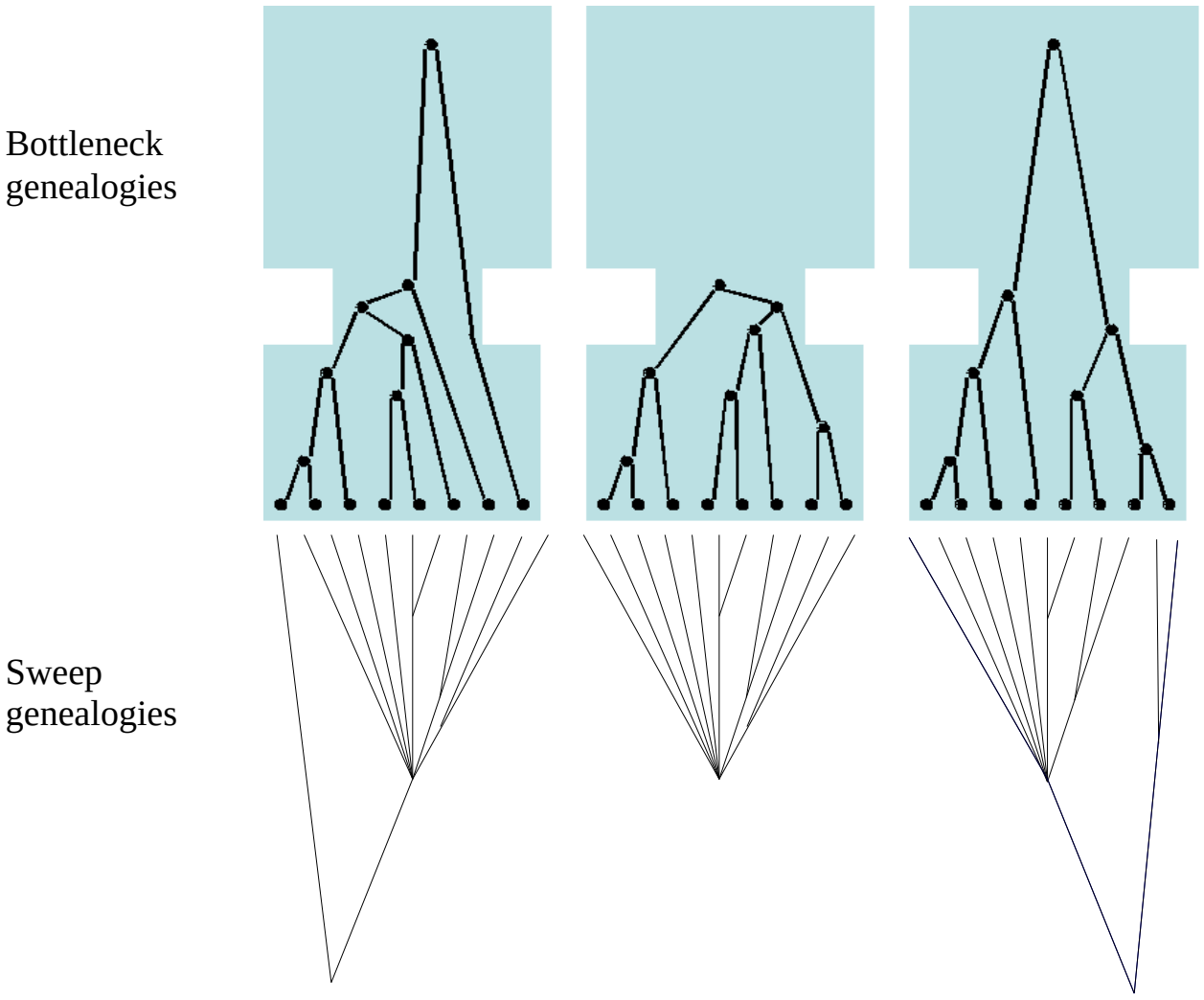

FIGURE 1: Coalescent trees on a recombining genomic region for a bottleneck (upper panel), and a selective sweep (bottom panel). The genealogies can be very similar, and this creates an intrinsic problem in disentangling selective sweeps from bottlenecks

The composite-maximum-likelihood estimates of p_i are given by the empirical frequencies $p_i = \frac{k}{n}$, where y_i is the number of sequences that carry the derived allele at site i, and l is the length of the region under study. The proposed goodness-of-fit statistic is then formulated

$$\Lambda_{GOF} = \log \frac{\max P(\text{Data}|H_A)}{\max P(\text{Data}|H_0)} \tag{6}$$

For the null hypothesis H_0, the maximization refers to the X and α parameters. For the alternative hypothesis H_A the maximization is related to the estimates of p_i as mentioned above. Λ_{GOF} values cannot directly reveal the favorable model even if, intuitively, small Λ_{GOF} values support the selective sweep hypothesis. More importantly, it is difficult to predict the effect of various evolutionary forces on the value of Λ_{GOF}. This is because the alternative hypothesis lacks the function $\phi(\cdot)$, which is specific to the evolutionary model. Simulations under a selective sweep scenario are employed in order to obtain the null distribution of the Λ_{GOF} statistic. The parameters X and α are estimated using the *CLR* approach of KIM and STEPHAN (2002). If the p-value for Λ_{GOF} is smaller than a cutoff value P_c, then H_0 is rejected, otherwise it is accepted. JENSEN *et al.* (2005) suggest a cutoff value of 0.15.

Simulating neutral data under various bottleneck scenarios allows for the estimation of the false positive rate of the *GOF* approach under various values of P_c. We may use $P_{0.05}$ such that the false positive rate of the *GOF* test would be 0.05. Simulations, however, show that both the false positive rate (using a certain cutoff P_c) and the value $P_{0.05}$ depend on the demographic scenario. Thus, results obtained using the *GOF* test should be interpreted carefully, when there is evidence that the population has experienced recent demographic changes, especially bottlenecks. It should be noted that there is not a single $P_{0.05}$ value appropriate for all the demographic changes.

The combined *CLR* and *GOF* tests are used extensively in subgenomic scans for the detection of selective sweeps. Subgenomic datasets are usually obtained by re-sequencing short fragments of DNA segments. Subsequently, a particular 'interesting' region that shows evidence for a selective sweep may be selected for fine-scale sequencing and parameters like the position of the sweep or the strength of selection are estimated from the data. However, the pre-selection of interesting regions creates an ascertainment scheme that has been shown to result in high false positive rates (THORNTON and JENSEN 2007). Both the *CLR* and *GOF* tests are not robust to this combination of ascertainment and demography. THORNTON and JENSEN (2007) propose to control the false positive rate by using the null distribution of the Λ_{CLR} statistic that is both generated from the correct demographic model and conditional on the ascertainment scheme. This strategy can be applied when the demographic model is known or can be estimated from the data. The source codes of the *CLR* and the *GOF* tests as well as their documentation are freely available online or

may be requested from Yuseob Kim and Jeffrey Jensen, respectively.

Genome-wide data

***SweepFinder*:** The availability of whole-genome or chromosome-wide SNP data, mainly from the HapMap Project (INTERNATIONAL HAPMAP CONSORTIUM 2003), motivated NIELSEN *et al.* (2005) to develop a method for the detection of selective sweeps, which would allow for an analysis of genome-wide data. Full genomic scans, however, face several challenges. First, the confounding effects of demography obscure the detection of selective events in similar ways as in subgenomic scans. Second, data usually consists of SNPs that were initially identified in an ascertainment process, which may be quite complicated in some cases and can generate biases that should be taken into account.

The test *SweepFinder* proposed by NIELSEN *et al.* (2005) is a composite-likelihood ratio test that is based on the ideas of the *CLR* approach of KIM and STEPHAN (2002). However, it differs from the latter one in that the null hypothesis is not derived from a specific evolutionary model, but estimated from the empirical background distribution of the data. The idea behind the use of the background distribution is similar to the arguments presented in JENSEN *et al.* (2005) for formulating the alternative hypothesis. That means that the non-selective evolutionary processes that shape SFS affect the whole genome. Additionally, the method relies on the assumption that a class of neutral DNA exists in the genome.

SweepFinder is also based on the principles of the hitchhiking theory. That is, when a beneficial mutation occurs on a chromosome and goes to fixation, variation at linked neutral loci is reduced as the beneficial mutation spreads through the population. A selective sweep is modeled by assuming that each ancestral lineage escapes the sweep with a probability p_e, which is given as a function of the recombination distance from the selected site and $\alpha = \frac{r}{s}\ln(2N)$. Given that some lineages have escaped the selective sweep by recombination, the method calculates the probability to observe a mutant allele of frequency B. In order to calculate this quantity, the method estimates the number of ancestral lineages that carry the neutral mutation after the end of the selective phase and assumes that the SFS after the selective sweep is the same as at present (*i.e.* at the time of sampling).

Similarly to KIM and STEPHAN (2002), *SweepFinder* uses a composite-likelihood ratio approach to choose between a neutral and a selective model. The alternative hypothesis H_A states that a beneficial mutation has occurred at some position X. The likelihood of H_A is calculated as the product of the site probabilities (p^*_B) for all the sites and maximized with regard to the parameters X and α. When only polymorphic sites are included in the dataset the method is properly standardized. The null hypothesis is formulated as the probability to observe the data given the

empirical frequency spectrum. That means that if the probability of a specific allelic class is f_i, $i = 1, \ldots, n-1$, in the case of an unfolded spectrum and the allelic class at position j is given by ξ_j, then the likelihood of H_0 is equal to

$$L_{H_0} = \prod_{j=1}^{l} f_{\xi_j}. \tag{7}$$

Obviously, L_{H_0} depends only on the empirical frequency spectrum. Similarly to KIM and STEPHAN (2002), the composite-likelihood ratio statistic is given by

$$\Lambda_{SF} = \log \frac{\max P(\text{Data}|H_A)}{P(\text{Data}|H_0)}, \tag{8}$$

The null distribution of the statistic Λ_{SF} is obtained by using the specific demographic scenario that might have shaped the observed data. Even if the method is robust against several demographic scenarios that have been investigated in NIELSEN *et al.* (2005), our simulations have shown that this does not hold in general, especially in cases of severe and recent bottlenecks. Additionally, it is unknown how the method behaves in cases that SNP data is retrieved from the ascertainment schemes described in THORNTON and JENSEN (2007). Thus, these factors should be included when the null distribution of the statistic Λ_{SF} is constructed and from this a threshold value is calculated. The method is robust against multiple testing.

SweepFinder may also be applied to subgenomic data. In this case, the program offers the flexibility to employ a user-specified frequency spectrum instead of calculating it from the data. This may be useful when the genomic region under study is not representative of the whole genome.

The source code and the documentation of *SweepFinder* are available from Rasmus Nielsen's webpage `http://www.binf.ku.dk/~rasmus/webpage/sf.html`. The program is written in C and tested successfully on 32-bit and 64-bit machines. The simulations for the calculation of the threshold of Λ_{SF} may also be done on computer clusters.

Joint inference of demography and positive selection: While the *CLR* and *GOF* approaches do not use explicit demographic models, LI and STEPHAN (2006) describe a statistical method to detect footprints of selection in chromosome- or genome-wide data (multiple loci), while taking fluctuations of the population size into account (LI and STEPHAN 2006). They analyze X chromosomal SNP data from a Zimbabwe and European *Drosophila* population. Initially, they infer the demographic scenario of the African population (from the ancestral range). This is characterized by a stepwise expansion such that population size changed instantaneously some generations ago. The European population is derived from the African population thereby undergoing a recent severe bottleneck. The parameters of this model are estimated by applying maximum-likelihood

techniques based on the SFS for the African population and the joint SFS for the European population. In the analysis it is assumed that there is no recombination within loci (which are only about 500 bp long in this dataset), but the loci are partially linked.

Performing simulations for the whole X chromosome, and considering that the SFSs between the loci are independent given their genealogy, they inferred the parameters of the demographic scenario by maximum likelihood.

LI and STEPHAN (2006) avoid the problem of inefficient sampling of genealogies by calculating the likelihood as a function of the expected branch lengths that may produce the observed pattern

$$L_k = P(\mathrm{SFS}|G_k) = \prod_{i=1}^{n_k-1} P(\zeta_{ik}|E(l_{ik})) \tag{9}$$

However, this is just an approximation and its accuracy has still to be demonstrated.

After estimating the demography, LI and STEPHAN (2006) perform a sliding window analysis to find genomic regions which are affected by the action of strong positive selection. They conduct a likelihood-ratio test which employs two hypotheses. The null hypothesis considers that the population has experienced the inferred demographic scenario, and the alternative one assumes that the forces that shape the data consist of the inferred demographic scenario together with a selective sweep. In order to overcome the problem of inefficient sampling of genealogies for the loci that belong to the sliding window, they consider a compact frequency spectrum. In this approach, all high frequency variants are pooled together and hence the number of inconsistent coalescent trees is diminished (LI and STEPHAN 2005).

It is encouraging and promising that methods that incorporate demographic events explicitly in the inference of selection are being developed. Even if the *CLR-GOF* and the *SweepFinder* approaches do that only indirectly, demographic models can be incorporated in the estimation of the null distributions of the relevant statistics. Simulations have shown that this strategy can control the false positive rate (THORNTON and JENSEN 2007).

LI and STEPHAN (2006) implemented a software package called Mosy `http://www.zi.biologie.uni-muenchen.de/~li/mosy/` to detect recent selective sweeps and estimate parameters in populations of varying size.

Methods for detecting selection based on genetic differentiation between populations

F_{ST} based methods: Bayesian approaches have been shown to be powerful for quantifying dif-

ferentiation between populations (BALDING and NICHOLS 1995) and for the estimation of demographic events (BEAUMONT 2003). More recently, Bayesian methods have also been applied to whole-genome data (multiple loci) in order to reveal genetic regions that have experienced selective sweeps (BEAUMONT and BALDING 2004; RIEBLER *et al.* 2008). These methods combine information from multiple populations. Thus, they are able to use data produced from recent genotyping (*e.g.* INTERNATIONAL HAPMAP CONSORTIUM 2005) and sequencing projects (*e.g.* GLINKA *et al.* 2003). Additionally, they can infer both positive and balancing selection. Since RIEBLER *et al.* (2008) extended the method introduced by BEAUMONT and BALDING (2004) we discuss here the RIEBLER *et al.* (2008) approach.

This approach infers selective events using the F_{ST} value of a population for a given locus in a hierarchical two-level Bayesian framework. F_{ST} represents the probability that two randomly chosen alleles from the locus in the same subpopulation are identical because of common ancestry within the subpopulation (CROW and KIMURA 1970). In a coalescent framework, F_{ST} may be seen as the probability that a coalescent event precedes a migration event (going backwards in time) (Hudson, 1998). That means that F_{ST} values may be used for inferring balancing or positive selection since positive selection may increase the F_{ST} value and balancing selection decreases it.

In the two-level model of RIEBLER *et al.* (2008), the lower level expresses the likelihood for the allele-frequency counts as a function of F_{ST} using a multinomial Dirichlet model (BEAUMONT 2003; BEAUMONT and BALDING 2004). The higher level describes the logistic regression of locus-specific, population-specific and locus-by-population-specific effects on F_{ST}. The advancement of the RIEBLER *et al.* (2008) approach consists in a reparameterization of the original framework of BEAUMONT and BALDING (2004) and the subsequent use of an auxiliary Bayesian variable that indicates if a locus is neutral or not.

It should be noted, however, that both the BEAUMONT and BALDING (2004) and the RIEBLER *et al.* (2008) approaches are based on haplotype statistics since distinct haplotypes (*e.g.* sequences) are treated as alleles. As a consequence, the calculation of genetic differentiation based on loses information when many haplotypes in the sample are unique. This may be the case when the sample size is small, the mutation rate high and/or the sequence of a locus long. The source code, C executable files and R programs, is available from Andrea Riebler (andrea.riebler@ifspm.uzh.ch).

Haplotype-based methods: All methods discussed thus far are designed to detect complete sweeps within a panmictic population or, in the case of population structure, within a subpopulation. To discover incomplete sweeps (*i.e.* sweeps that are ongoing within a subpopulation or sweeps that are complete within one subpopulation, but not with regard to the total population), haplotype-based methods have been developed. These methods analyze the length of haplotypes

around a given locus of interest, which is thought to be the target of selection.

If a selective sweep is ongoing in a subpopulation, the hitchhiking haplotype is expected to be rather long (see Box on page 2). This feature of the hitchhiking effect has been exploited by SABETI *et al.* (2002) who developed the so-called long-range haplotype (LRH). A slight modification of this is the iHS statistic (VOIGHT *et al.* 2006). A disadvantage of these approaches is that they lose power when the beneficial allele is close to fixation. To overcome this problem, TANG *et al.* (2007) and SABETI *et al.* (2007) apply the ideas of the haplotype-based tests not to a single (local) subpopulation but contrast the haplotype profiles between subpopulations.

Until now, little is known about the power and robustness of haplotype-based methods. Additional research is needed to investigate the false positive rate of these methods under various demographic scenarios or migration models when more than one subpopulation is involved.

Methods based on the machine learning paradigm

Given the parameter values of a selective sweep and the parameter values of a bottleneck, disentangling a selective sweep from a bottleneck can be treated as a binary classification problem, where a dataset must be assigned to either the neutral class or to the selection class. In the computer science and mathematics disciplines theoretical and algorithmic advancements have been developed the last decades that perform classification of datasets. These advancements can be grouped as machine learning methods, because first they teach computers to understand patterns from the data, and then to use this knowledge in order to classify an unknown sample. However, their application in population genetics still remains limited. The first application of the machine learning in population genetics to our knowledge was developed by PAVLIDIS *et al.* (2010), who used a support vector machine approach to perform the classification. PAVLIDIS *et al.* (2010) used as features the results from the *SweepFinder* (NIELSEN *et al.* 2005), the ω-statistic (KIM and NIELSEN 2004), and the distance between the peaks of the statistics. More recently, LIN *et al.* (2010) also developed a machine learning approach based on the 'boosting' algorithm, a statistical method that combines simple classification rules using summary statistics to maximize their joint predictive performance. Details about the machine learning approach are provided in Chapter 1 of the thesis.

Aims of the thesis

This study deals with the detection of selective sweeps in natural populations. The model organism is *D. melanogaster*; however, the methods developed as a result of this research can be adapted to most of the organisms of relatively large effective population size and outcrossing reproduction.

The aims of the thesis are: (i) implementation of a method that is able to detect selective sweeps in natural populations that have experienced past demographic changes; (ii) application of the methods to real data; (iii) extension of selective sweeps in multi-locus models.

This thesis is organized in three chapters. In Chapter 1 I implement computer simulations of a single population that has experienced past demographic changes with or without selection in order to scrutinize the polymorphism patterns that selection may generate in the genome. Two algorithms have been used. First, the *SweepFinder* (NIELSEN *et al.* 2005) identifies genomic regions where a selective sweep fits better than a demographic model based on SFS information. Second the ω-statistic (KIM and NIELSEN 2004) identifies genomic regions using the LD information instead of the SFS. In Chapter 1 I combine the results of *SweepFinder* and the ω-statistic. Since the LD and the SFS are partially independent, combining the *SweepFinder* with the ω-statistic may be advantageous for disentangling selective sweeps from neutrality. Their combination is implemented in a machine learning framework using support vector machines (VAPNIK 1995). The power of support vector machines has been demonstrated mainly in microarray analysis studies (*e.g.* FUREY *et al.* 2000), where combinations of gene expression values are used to separate classes of individuals (*e.g.* high-risk versus low-risk patients). In Chapter 1 the two classes are characterized by the presence or the absence of a selective sweep. For many demographic scenarios, combining the *SweepFinder* with the ω-statistic outperforms both algorithms when they are applied separately. However, in general, as shown in Chapter 1, the problem of disentangling demography (especially bottlenecks) from selection is considerably challenging, because both demographic bottlenecks and selective sweeps can generate similar SFS and LD patterns.

Real data pose challenges on the application of the methods because they deviate partially from the assumptions of the methods. Often, errors in the data, violations of the model assumptions, and ascertainment biases must be taken into account. In Chapter 2 I apply the *SweepFinder* algorithm and the *CLR* test on the subgenomic region that includes the gene *HDAC6* (SVETEC *et al.* 2009). An African population of *D. melanogaster* is examined. HDAC6 is an unusual histone deacetylase being localized in the cytoplasm. Recent discoveries have shown that HDAC6 is a key regulator of cytotoxic stress resistance. The first evidence for a selective sweep in the *HDAC6* region was supported by a previous genome scan (LI and STEPHAN 2006). LI and STEPHAN (2006) discovered a 100-kb fragment that overlaps with the *HDAC6* region and showed evidence of recent positive selection in the European population of *D. melanogaster*. This prior information generated ascertainment bias in the analysis underestimating the p-values. Performing a joint analysis of the African and the European populations of *D. melanogaster* remedies the ascertainment bias.

While the first and the second aims of the thesis are based on the classical one-locus two-allele

model of selective sweeps, the third goal represents the extension of selective sweeps to multi-locus models. Even if the evolution of multi-locus models has been studied elsewhere (BÜRGER 2000), the study of allelic frequency trajectories is still limited. In Chapter 3 I use computer simulations in order to obtain the trajectory of an allele that initially is in low frequency. Deterministic and stochastic simulations have been implemented for the two-locus two-allele model as well as the stochastic five-locus two-allele model. In contrast to the classical one-locus two allele model, the trajectories may reach polymorphic equilibria (*i.e.* equilibria where both of the alleles of the focal locus are maintained in the population). The establishment of a polymorphic equilibrium generates profoundly different polymorphism patterns than classical selective sweeps. Therefore, many of the selective events that might occur in a multi-locus regime will be unidentified by the neutrality tests that have been developed for one-locus two-allele models.

Chapter 1

Searching for footprints of positive selection in whole-genome SNP data from non-equilibrium populations

Pavlos Pavlidis[†,*], Jeffrey D. Jensen[‡], Wolfgang Stephan[†]

[†] Department of Biology II, Ludwig-Maximilians-University Munich, 82152 Planegg, Germany

[‡] Program in Bioinformatics and Integrative Biology, University of Massachusetts, Medical School, Worcester, MA

Genetics 2010, 185:907–922

1.1 Abstract

A major goal of population genomics is to reconstruct the history of natural populations and to infer the neutral and selective scenarios that can explain the present-day polymorphism patterns. However, the separation between neutral and selective hypotheses has proven hard, mainly because both may predict similar patterns in the genome. The present study focuses on the development of methods that can be used to distinguish neutral from selective hypotheses in equilibrium and non-equilibrium populations. These methods utilize a combination of statistics based on the site frequency spectrum (SFS) and linkage disequilibrium (LD). We investigate the patterns of genetic variation along recombining chromosomes using a multitude of comparisons between neutral and selective hypotheses, such as selection or neutrality in equilibrium and non-equilibrium popula-

tions, and recurrent selection models. We perform hypothesis testing using the classical p-value approach, but we also introduce methods from the machine learning field. We demonstrate that the combination of SFS- and LD-based statistics increases the power to detect recent positive selection in populations that have experienced past demographic changes.

1.2 Introduction

Genomes contain information related to the history of natural populations. Past neutral and selective processes may have left footprints in the genome. Recent advances in population genetics aim to understand the patterns of genetic diversity and identify events that have led to genetic adaptations. Among them, positive selection has been a focus of many recent studies (HARR *et al.* 2002; KIM and STEPHAN 2002; GLINKA *et al.* 2003; AKEY *et al.* 2004; ORENGO and AGUADÉ 2004). Their goal is to (i) provide evidence of positive selection, (ii) estimate the strength and the rate of selection, and (iii) localize the targets of selection. These objectives form the basis of a long-term pursuit, which is the understanding of the molecular basis of adaptation of populations in a changing environment.

Positive selection can cause genetic hitchhiking when a beneficial mutation spreads in the population (MAYNARD SMITH and HAIGH 1974). When a strongly beneficial mutation occurs and spreads in a population, linked neutral or slightly deleterious variants hitchhike with it, and their frequency increases. According to Maynard Smith and Haighs model, three patterns are generated locally around the position of the beneficial mutation. First, the level of variability will be reduced since standing variation of the population that is not linked to the beneficial allele vanishes, and tightly linked polymorphisms may fix (KAPLAN *et al.* 1989; STEPHAN *et al.* 1992). Second, the site frequency spectrum (SFS), which describes the frequency of allelic variants, shifts from its neutral expectation towards rare and high-frequency derived variants (BRAVERMAN *et al.* 1995; FAY and WU 2000). The third signature describes the emergence of specific linkage disequilibrium (LD) patterns around the target of positive selection, such as an elevated level of LD in the early phase of the fixation process of the beneficial mutation and a decay of LD across the selected site at the end of the selective phase (KIM and NIELSEN 2004; STEPHAN *et al.* 2006).

The availability of genome-wide SNP data has made possible the scanning of genomes and the identification of loci that may have been targets of recent selective events. Several approaches have been developed within the last years that can detect the molecular signatures of positive selection (KIM and STEPHAN 2002; JENSEN *et al.* 2005; NIELSEN *et al.* 2005). While the methods of KIM and STEPHAN (2002) and JENSEN *et al.* (2005) are designed to analyze subgenomic SNP data,

the approach of NIELSEN *et al.* (2005) can be applied to both subgenomic and whole-genome data (reviewed in PAVLIDIS *et al.* (2008)). For this reason we concentrate here on the latter procedure. This method, called *SweepFinder*, calculates the probability $P(x)$ that a polymorphism of multiplicity x is linked to a beneficial mutation using a simple selective model and the SFS prior to the selective event. Then, for each location in the genome it compares a selective with a neutral model assuming independence between the SNPs, therefore calculating the composite likelihood ratio Λ. Thus, it identifies regions where the likelihood of the selective sweep is greater than that of the neutral model using the maximum value Λ_{MAX} of Λ.

The ω-statistic, developed by KIM and NIELSEN (2004), detects specific LD patterns caused by genetic hitchhiking (described above). In the study by KIM and NIELSEN (2004) the maximum value of the ω-statistic was used to identify the targets of selective sweeps. Later, JENSEN *et al.* (2007b) studied its performance in separating demographic from selective scenarios. An important result by JENSEN *et al.* (2007b) is the demonstration that for demographic parameters relevant to non-equilibrium populations (such as the cosmopolitan populations of *D. melanogaster*) the ω-statistic can distinguish between neutral and selective scenarios. This paper will further develop *SweepFinder* and the ω-statistic such that they can eventually be applied to whole-genome SNP datasets that have been collected from non-equilibrium populations. In particular, populations undergoing population size bottlenecks are of interest as these size changes may confound the patterns of selective sweeps (BARTON 1998). For this reason we use the following approach: first, we theoretically analyze the genealogies of bottlenecked populations under neutrality and show to what extent they resemble the genealogies of single hitchhiking (SHH) events. We also point out the importance of high-frequency derived variants in the identification of selective sweeps. Second, we study the statistical properties of *SweepFinder* and the ω-statistic separately and in combination. As the main result, we demonstrate that the combination of these two methods (that include both SFS and LD information) increases the power for detecting recent SHH events in non-equilibrium populations, in particular when machine-learning techniques are employed. Third we analyze the performance of *SweepFinder* and the ω-statistic in the detection of recurrent hitchhiking (RHH) events.

1.3 Methods

Modifications of the ω-statistic and *SweepFinder*: The proposed modifications aim at (i) adapting the ω-statistic for the analysis of whole-genome data, and (ii) increasing the accuracy of *SweepFinder* to predict the target of selection. Instead of fixed windows, variable-size windows

are used in the ω-statistic, and in the *SweepFinder* algorithm a fraction of monomorphic sites is incorporated.

The hitchhiking model by MAYNARD SMITH and HAIGH (1974) predicts that an excess of LD arises after the completion of the selective sweep within each of the two regions flanking the selected site, but does not extend across the two regions (STEPHAN *et al.* 2006; MCVEAN 2007; PFAFFELHUBER *et al.* 2008). This is due to the assumption that any observed polymorphism around the sweep has been introduced in the population prior the selective sweep, and entered the beneficial genetic background through recombination. Since independent recombination events are necessary to explain polymorphisms on both sides of the selective sweep, the LD vanishes across the site of the beneficial mutation, but not within each side. This genomic footprint may be captured using the ω-statistic (KIM and NIELSEN 2004). Assume a genomic window with S segregating sites that is split into a left and right sub-region with l and $S-l$ segregating sites, respectively. The ω-statistic (equation 1.1) quantifies to what extent average LD is elevated on each side of the selective sweep (see the numerator of equation 1.1) but not across the selected site (see the denominator of equation 1.1).

$$\omega = \frac{\left(\binom{l}{2} + \binom{S-l}{2}\right)^{-1}\left(\sum_{i,j\in L} r_{ij}^2 + \sum_{i,j\in R} r_{ij}^2\right)}{(l(S-l))^{-1}\sum_{i\in L, j\in R} r_{ij}^2}. \tag{1.1}$$

The ω-statistic considers the space between the left and right sub-regions as the center of the selective sweep. Thus, a genomic region may be scanned and scores are reported for each position. Then, using simulations, a significance threshold is determined. The maximum value ω_{MAX} predicts the target of recent positive selection. In the original version of the ω-statistic, the borders of the left and right sub-regions are assumed constant (KIM and NIELSEN 2004; JENSEN *et al.* 2007b). This may be valid for a subgenomic analysis, when the recombination rate ρ and mutation rate θ do not fluctuate much or a single selective event may have occurred. However, in a whole-genome study these parameters that affect the extent of LD may vary dramatically. Additionally, the polymorphism patterns may have been shaped by recurrent selective sweeps. Thus, the constant-border approach implemented by KIM and NIELSEN (2004) may be limited. If the sub-regions are large, then ω_{MAX} tends to decrease and the signal disappears. On the other hand, short sub-regions might contain no SNPs and the ω-statistic cannot be calculated.

We have implemented a variable-window size ω-statistic. The borders of the left and right sub-regions vary and the configuration that maximizes ω is reported. This approach overcomes the afore mentioned problems inherent in the constant-border approach of KIM and NIELSEN (2004). Thus, it may be suitable for scanning large genomic regions or whole chromosomes characterized

by variable ρ or θ parameters and shaped by recurrent adaptive substitutions.

A naive implementation of the ω-statistic scanning algorithm would re-calculate the LD of the positions as the center of the sweep moves along the chromosome. This is particularly critical for the variable-window size approach since the number of calculations increases. Our implementation, as illustrated in Appendix, in Table *S*1, guarantees a single calculation between any two sites that may participate in the ω calculation. Thus, it results in an algorithm that is efficient when the number of polymorphisms is large. Calculations are performed using a matrix *Z* (Table *S*1 in Appendix), which stores the unweighted Z_{nS} (KELLY 1997) values (not divided by the number of comparisons) for all possible windows. For a pair $(i, i+1)$, $Z_{i,i+1}$ equals the correlation coefficient between these two positions. This value is then added to all cells $Z_{j,i+1}$, with $j < i$ to form the Z_{nS} for the region $[j, i+1]$. With this method all possible numerators of the ω-statistic are formed. When the left and right sub-regions are defined by $[i,k]$ and $[k+1, j]$, respectively, then the denominator is simply a weighted version of $Z_{i,j} - Z_{i,k} - Z_{k+1,j}$.

SweepFinder detects the shift of the SFS as a signature of hitchhiking. Demographic effects are incorporated through the neutral SFS, which is either provided by the user or calculated from the data itself. Monomorphic sites are generally excluded from the analysis (NIELSEN *et al.* 2005; SVETEC *et al.* 2009) since tests that include them may be more sensitive to assumptions regarding the mutation rate (NIELSEN *et al.* 2005). Additionally, for realistic mutation rates, the majority of the sites remain monomorphic. Thus, by including invariant sites the dataset and the computational time required for the analysis increase dramatically. On the other hand, the decrease of diversity represented by the monomorphic sites constitutes a well-known signature of the hitchhiking effect. Omitting them may decrease the power of the tests (NIELSEN *et al.* 2005) and lead to inaccurate predictions about the target of selection. Inaccuracies mainly emerge due to changes in the input site density when only polymorphic sites are included. We incorporate a fraction of the monomorphic sites into the analysis in a way that (i) generates a uniform input site density and (ii) preserves the signature of low diversity in regions of depleted variation. Additionally, since only a small fraction of monomorphic sites are used, the computational time is only increased slightly. Given a genomic region with *S* polymorphic sites we include *Sq* monomorphic sites, where $0 < q < 1$. In the present study $q = 0.1$, so that the number of monomorphic are in the same order as the polymorphic sites. We proceed as follows. In the first step, there are $S-1$ intervals between the *S* polymorphic sites. A monomorphic site is included at a random location within the largest interval. In the second step there are $S+1$ sites and *S* intervals and the process is repeated. The cutoff value is defined by treating the neutral simulations in the same way. With this process the SNP density differences are reduced and monomorphic sites are embedded in regions of depleted variation.

Quantifying the effects of population bottlenecks on neutral genealogies: The ω-statistic and *SweepFinder* can scan genomes from natural populations that have experienced demographic changes and detect targets of selection. We investigated whether the neutral demographic scenarios inferred by LI and STEPHAN (2006) and THORNTON and ANDOLFATTO (2006) to describe the demography of a European population of *D. melanogaster* can result in patterns along a recombining chromosome that resemble selective sweeps. In particular, we examined which effects of population bottlenecks are responsible for the polymorphism patterns that mimic the effects of selective sweeps. We focused on the properties of genealogies that are generated by those two demographic models because genealogies reflect demographic properties more comprehensively than summary statistics.

A way to measure the effect of a bottleneck on the genealogies of a recombining genome is through the ratio $f = \frac{L_n}{H_n}$ of the total length to the height of the coalescent. Short, star-like genealogies have large ratios and $\max(\frac{L_n}{H_n}) = n$ is obtained for a n-furcated star-like tree. On the other hand, for genealogies with long internal branches the ratio takes small values and $\min(\frac{L_n}{H_n}) = 2$ is obtained when the genealogy is dominated by two very long internal branches. Using simulations we first calculate the percentage of n-furcated star-like genealogies (with large f values) in a region of 50 kb. Then, for each simulated instance we relate the percentage of n-furcated star-like genealogies with the resemblance to a selective sweep as this is measured using *SweepFinder* (see **Theoretical analyses**).

The joint effects of population bottlenecks and selective sweeps on high-frequency derived alleles: A hallmark of selective sweeps in constant populations is the excess of high-frequency derived variants around the target of positive selection. High-frequency derived variants consist of mutations that were present in the population prior the selective sweep, hitchhike with the beneficial allele and, due to recombination, appear as polymorphisms. This signature forms the basis of a multitude of neutrality tests that are based on the SFS (FAY and WU 2000; KIM and STEPHAN 2002; NIELSEN *et al.* 2005) and contributes to the precise detection of the target of selection. However, in natural populations positive selection may occur simultaneously with demographic changes. Using simulations from the demographic models that were inferred by LI and STEPHAN (2006) and THORNTON and ANDOLFATTO (2006), we examine whether high-frequency derived alleles occur when demographic changes occur simultaneously with positive selection.

Measuring the precision of the inferred selective sweep position: An objective of the genome-scanning studies is the precise prediction of the selective sweep locations. Usually, every position or a subset of them is scored for a given statistic (for example the ω-statistic or the *SweepFinder*). Thus, peaks and valleys are formed along the genomic region. Then, some of the

peaks may survive a cutoff value delimiting the potential targets of selection. As illustrated in Figure *S*1 in Appendix, we determine the distance between a peak on the landscape of the statistic and the closest location where a selective sweep has occurred given a user-defined threshold. In Figure *S*1, two selective sweeps have occurred recently in the history of the population. The positions of the sweeps are illustrated as vertical green lines. A peak is defined as the highest point in an isolated region by the cutoff value. Thus, five peaks (*a* to *e*) have been formed in the example of Figure *S*1. *D* measures the distance between a peak and the closest selective sweep location. Based on this approach we can measure the accuracy of the different methods. Furthermore, we implemented a simple randomization of the peaks to evaluate the quality of the predictions. This is necessary because finite genomic regions are simulated, and therefore the distance between any location and the target of selection is bounded.

Supervised learning techniques: We introduce supervised learning approaches from the field of machine learning that can be useful for the classification of a genomic region as either neutral or selected. In a classification problem, the goal is to separate these classes using a function, which is inferred from the available data. Such a process is called 'learning from the data' or 'supervised' learning and is related to finding the optimal hyperplane that distinguishes the two classes. Typically, in a supervised learning problem, data consist of pairs of input and output objects. Input consists of a vector of multiple entries that summarize the data and are called features. Inputs can be set arbitrarily depending on the specific problem. However, the efficiency of the algorithm increases when they are independent and capture the whole information of the data. Output can be binary, denoting the class that the object belongs to. In supervised learning the goal is to use the input to predict the value of the output, and the problem can be formulated as teaching the computer the combinations of feature-values that are associated with either of the classes. In the specific problem we examine here, the output is coded as 'neutrality/selection'. Then, using simulations of the neutral demographic model and the model with selection we train the algorithm to separate these two classes. As input for the machine learning approach we use ω_{MAX}, Λ_{MAX} (from the original algorithms) and combinations of ω and Λ, such as the distance between the genomic positions of ω_{MAX} and Λ_{MAX} and the correlation coefficient between ω and Λ. The reasoning for this choice of inputs is as follows. First, Λ_{MAX} and ω_{MAX} capture different aspects of the data. Λ_{MAX} is affected mostly by the SFS, whereas ω_{MAX} is affected by LD. Even if SFS and LD can be correlated (KIM and NIELSEN 2004), it is expected that this correlation is lower than using statistics that are based exclusively on the SFS or LD. Second previous studies have shown that Λ_{MAX} and ω_{MAX} are relatively robust to demographic changes (but see ORENGO and AGUADÉ (2010)). Third, it seems intuitively obvious that the peaks of ω and Λ profiles should

point to the same genomic location if a selective sweep has occurred. Thus, using the distance between the peaks or the correlation of the profiles should increase the classification performance of the algorithm. In this study, both the distance between the peaks and the correlation between the profiles are used.

For each demographic scenario that was simulated in this study, we used a subset of simulations for training, and the remaining for testing the performance. The supervised learning approach can be employed to classify a certain genomic region as either neutral or selected. However, within a region the specific target of selection cannot be specified by the method itself. In order to achieve this, the features of the method (*i.e.* the ω and Λ profiles) should be inspected. Tables 1 to 4 provide information about the accuracy of the features under various demographic scenarios.

Traditionally, when neutrality tests are employed to detect targets of positive selection neutral simulations are performed and the 5% percentile is used as a threshold. This methodology assumes that neutrality tests produce significantly larger values in data with selection. This may be the case when the population size remains constant. However, in non-equilibrium models the values of the neutrality tests may overlap significantly between neutral models and models with selection, and therefore their performance decreases. Combining different statistics that capture different aspects of the data may contribute to increasing the classification performance.

Several methods have been developed for data classification. For example, Bayesian classifiers, rule-based classifiers, *k*-nearest-neighbors, and linear discriminant analysis are some of the approaches that have been applied to supervised learning problems (DUDA *et al.* 2000; HAN and KAMBER 2000; HASTIE *et al.* 2001). Here, we demonstrate the use of Support Vector Machines (SVMs) with a radial basis kernel, which is the most widespread kernel. In general, SVM uses a nonlinear mapping to transform the original training data into a higher-dimensional space and to search for an optimal linear hyperplane in this space. A great advantage of the SVMs is that they are highly accurate and less prone to overfitting; *i.e.* they have desirable generalization properties (HAN and KAMBER 2000).

Implementation and code availability: The C++ source code is available from `http://www.bio.lmu.de/~pavlidis`. For the ω-statistic, the user is able to choose between constant- or variable-window size scanning modes. Additionally, besides r^2 various other measurements of LD, such as abs(D) and abs(D_ω) (LANGLEY *et al.* 1974), may be used in equation 1. There are no specific library dependencies and the software can be installed on any Linux machine that runs the g++ compiler. Also, the modified version of *SweepFinder* that has been used here to analyze data with monomorphic sites is provided. In this version the likelihood curve of monomorphic sites has been modified so that the probability to observe a monomorphic site is high in the prox-

imity of the sweep position but becomes negligible as distance increases (the rate of decrease is larger than in the original version). The original version of *SweepFinder* is provided by the web site of Rasmus Nielsen `http://people.binf.ku.dk/rasmus/webpage/sf.html`. Furthermore, perl scripts that have been used in the analysis are available from `http://www.bio.lmu.de/~pavlidis` or upon request from the authors.

1.4 Theoretical analyses

The genealogies of bottlenecked populations may resemble those of SHH in constant-size populations: Past demographic changes such as bottlenecks may confound the patterns of a selective sweep (BARTON 1998). Similarly to a selective sweep, a bottleneck scenario may result in coalescent trees dominated by either external or internal branches. Short coalescent trees with long external branches are obtained when, due to a rapid, recent, and severe decrease of population size, the time of the most recent common ancestor of the sample is found within the bottleneck period. On the other hand, if some of the lineages escape the bottleneck, then long internal branches will be created. In recombining genomic regions short and long trees may alternate, creating sweep-like patterns in the SFS (BARTON 1998).

We illustrate the effect of bottlenecks on genealogies using the demographic scenarios that have been inferred by LI and STEPHAN (2006) and THORNTON and ANDOLFATTO (2006) to describe the history of the European population of *D. melanogaster*. Scaling the time in units of $4N$ generations (where N is the present effective population size) the LI and STEPHAN (2006) model describes a 4-epoch scenario. Backward in time, the population experiences a bottleneck from 0.0367 time units until 0.0375 time units. Within this bottleneck period $N_b = 0.002N$, where N_b denotes the effective population size in the bottleneck. Then, instantly, the size of the population size changes to $7.5N$, and eventually at the time 0.1395 it becomes $1.5N$. The bottleneck phase models the founding of the European population from the ancestral population, whereas the transition from $7.5N$ to $1.5N$ models a (forward-in-time) expansion of the ancestral population. The demographic scenario inferred by THORNTON and ANDOLFATTO (2006) implements a 3-epoch model. The values of the parameters depend on the ratio $\frac{\rho}{\theta}$ and here we use the results obtained when $\frac{\rho}{\theta} = 10$. The present population size N is estimated to be 2.4×10^6, and backward in time at 0.0042 it contracts to $0.029N$. Finally, the population reaches instantly the present-day level at time 0.022.

The demographic model of LI and STEPHAN (2006) produces both star-like and long genealogies in the same genomic region of a recombining chromosome (Figure 8). The length of these

trees is on average shorter than that of the standard neutral trees, thus reducing variation. The effect of the THORNTON and ANDOLFATTO (2006) demographic model is similar, however milder. On average, it creates shorter genealogies and effectively reduces the nucleotide polymorphism. However, it does not result in extreme star-like coalescent trees as often as the LI and STEPHAN (2006) model (Figure 8). This is because the population size changes are milder, the bottleneck period is longer, and starts (backward in time) very recently in the usual coalescent time scale, allowing for a series of coalescent events.

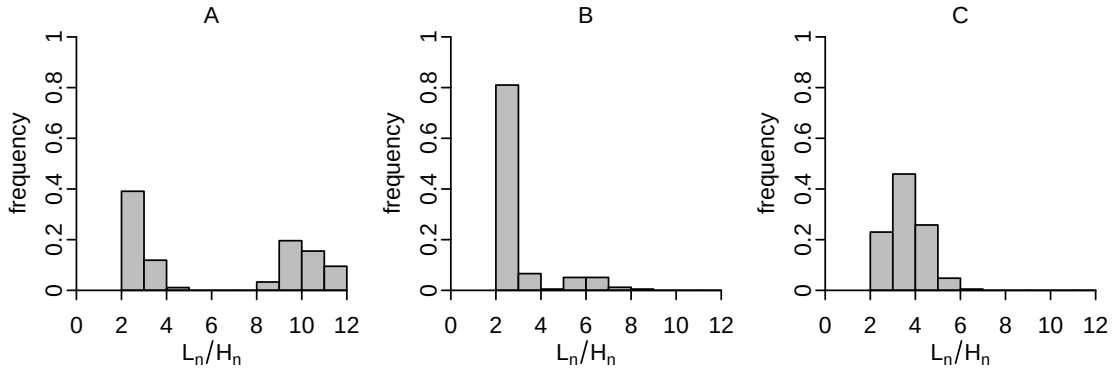

FIGURE 1.1: Histogram of the ratio $f = \frac{L_n}{H_n}$ for the following demographic scenarios: A) a single realization of the bottleneck scenario inferred by LI and STEPHAN (2006). Long coalescent trees that escape the bottleneck tend to produce small ratios (< 4). On the other hand, genealogies that coalesce within the bottleneck period produce star-like trees because of the recent, rapid and severe contraction of the population. B)a realization of the bottleneck scenario inferred by THORNTON and ANDOLFATTO (2006). In contrast to LI and STEPHAN (2006), coalescent events occur continuously. C) for the standard neutral model. For the LI and STEPHAN (2006), THORNTON and ANDOLFATTO (2006) and the neutral scenario, 12 chromosomes of 50 kb have been simulated. The recombination rate is $\rho = 0.05$/bp and the mutation rate $\theta = 0.004$/bp. The parameter values for the LI and STEPHAN (2006) and THORNTON and ANDOLFATTO (2006) scenarios are described in the main text.

Next we used simulations to examine the relationship between the percentage of star-like genealogies, the number of segregating sites, and Λ_{MAX} of *SweepFinder*, which can be considered a proxy for the resemblance of polymorphism patterns (based on the SFS) to a signature of a selective sweep. A 50-kb genomic region was simulated using *ms* (HUDSON 2002) for a sample of 12 chromosomes. The recombination rate $\rho = 0.05$/bp and the mutation rate $\theta = 0.004$/bp. The demographic model describes a recent population bottleneck (as inferred by LI and STEPHAN (2006)). As illustrated in Figure 1.2, a small number of star-like trees create a large number of segregating sites and small Λ_{MAX} values. Similarly, when a genomic region is dominated by short, star-like genealogies, the number of segregating sites and Λ_{MAX} decrease. Even if this constitutes a poly-

morphism valley, the pattern does not look like a sweep because of a lack of the high-frequency derived variants (KIM and STEPHAN 2002). On the other hand, the simultaneous presence of star-like and long genealogies creates sweep-like patterns. For intermediate frequencies of star-like genealogies, Λ_{MAX} assumes large values. Since neighboring genealogies are not independent, star-like genealogies form clusters and effectively create valleys of reduced polymorphism resembling a selective sweep. These results help to interpret some of our findings below.

Selective sweeps in non-equilibrium populations may result in a loss of high-frequency derived variants and violate the assumptions of *SweepFinder* and the ω-statistic: We examined the effects of selective sweeps on polymorphisms, when they occur within demographic bottlenecks. A 50-kb genomic fragment was simulated under the bottleneck model inferred by THORNTON and ANDOLFATTO (2006), and a selective sweep ($\alpha = 2500$) was assumed to take place within the bottleneck period (THORNTON and JENSEN 2007). First, we show that the combined action of selective sweeps and bottlenecks results in SFS that differ considerably from those generated by selective sweeps in equilibrium populations. Figure 1.3 compares the modifications of the average SFS around the target of selection in a constant-size demographic scenario with the model inferred by THORNTON and ANDOLFATTO (2006). It is apparent that in equilibrium demographic models there is a dramatic increase of the class '$n-1$' in the proximity of the selective sweeps (Figure 1.3a). Neutrality tests based on the SFS can detect the increase of the high-frequency derived variants and therefore the accurate prediction of the target of selection is possible. In non-equilibrium scenarios, when population contraction and selective sweeps co-occur, the '$n-1$' class vanishes in a large genomic region around the target of selection (Figure 1.3b). The joint effect of selection and population contraction increases the probability of coalescences, resulting in short genealogies where the most recent common ancestor is located within the bottleneck phase. Consequently, the frequency of the '$n-1$' class vanishes in the present-day sample. Furthermore, the part of the genealogy that is older than the selective sweep/bottleneck phase is eliminated. Therefore the vast majority of the present-day polymorphisms are younger than the selective sweep. This violates the assumptions of *SweepFinder* and the ω-statistic and may result in imprecise prediction of the target of selection.

1.5 Statistical performance of the tests in the detection of single hitchhiking events

In this section, the discrimination capacity of *SweepFinder* and the ω-statistic is scrutinized, and the distance between the predicted and the true target of selection is evaluated for single sweeps

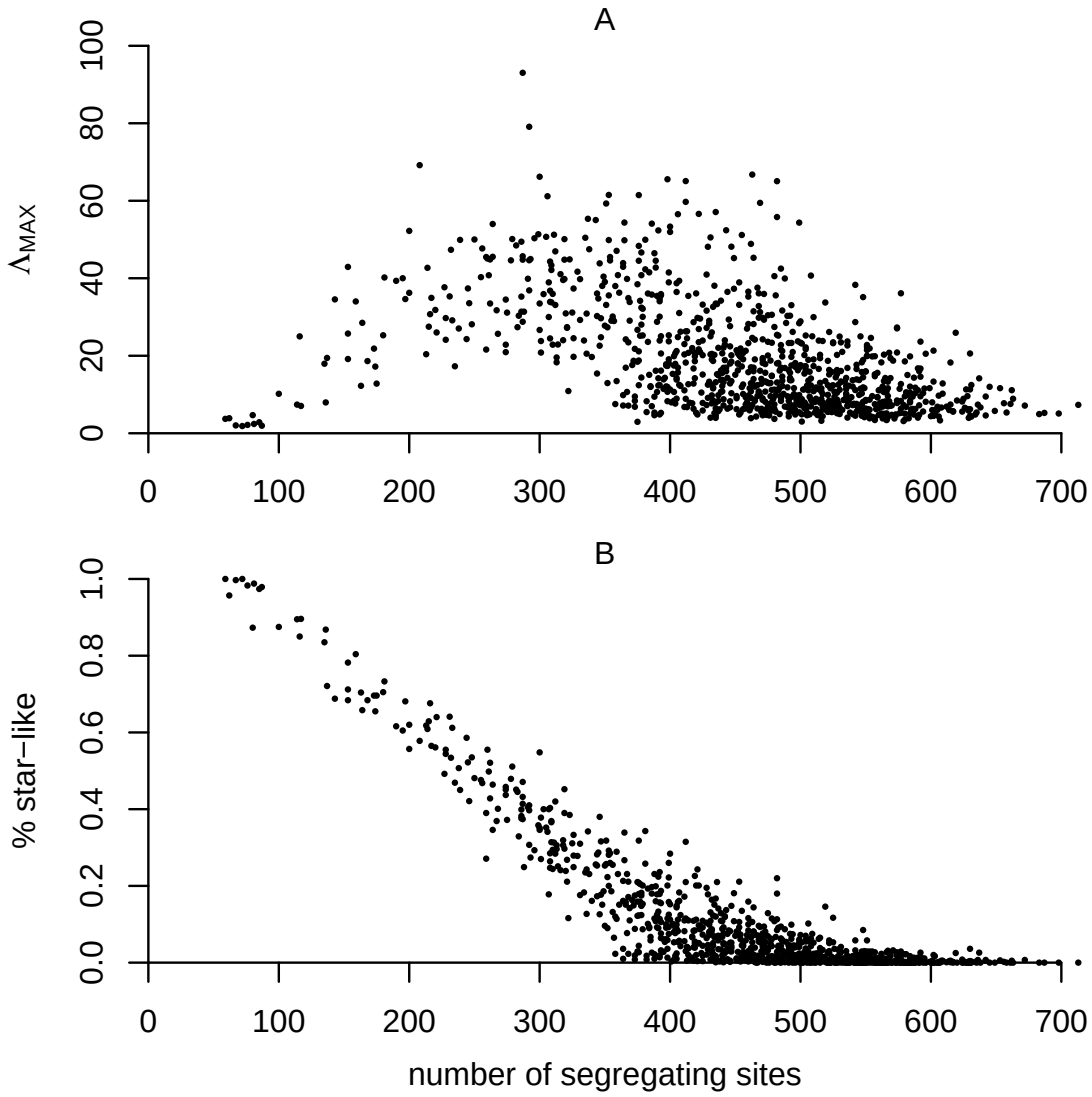

FIGURE 1.2: The relation between A) Λ_{MAX} and B) percentage of star-like genealogies, and the number of segregating sites in the LI and STEPHAN (2006) demographic scenario. We have performed neutral simulations for 12 recombining chromosomes, assuming a length of 50 kb. The recombination rate $\rho = 0.05$/bp and the mutation rate $\theta = 0.005$/bp. The parameter values for the demographic model inferred by LI and STEPHAN (2006) are described in the main text. The number of short genealogies in the LI and STEPHAN (2006) scenario determines both the number of segregating sites and the sweep-resemblance (measured by the *SweepFinder* statistic). When a genomic region is dominated by short star-like genealogies only a few segregating sites are present. Even if this constitutes a polymorphism valley, the pattern does not look like a single sweep because of a lack of the high-frequency derived variants (KIM and STEPHAN 2002). Similarly, when the star-like trees are absent Λ_{MAX} is small. On the other hand, the simultaneous presence of star-like and long genealogies creates sweep-like patterns. This is because star-like trees tend to cluster together along the recombining chromosome, creating valleys within polymorphism islands.

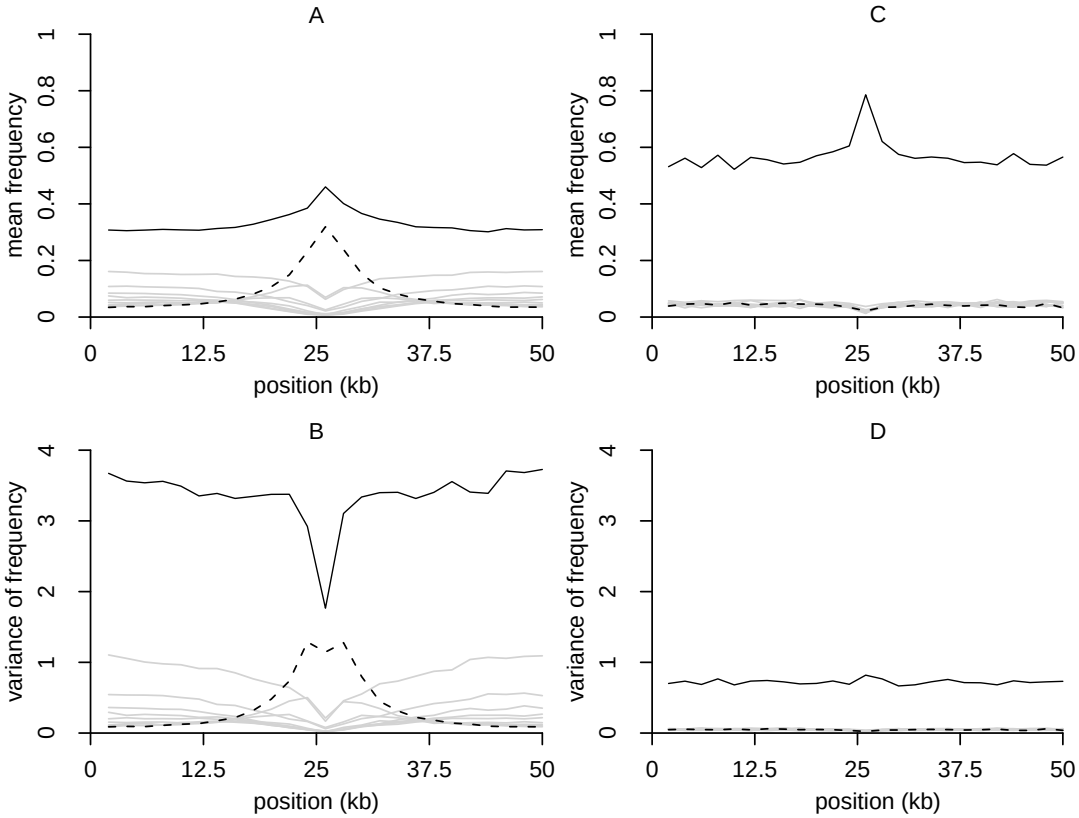

FIGURE 1.3: A selective sweep causes a spatial modification of the SFS. The mean and the variance of the frequency are modified when a selective sweep has occurred in the middle of a 50-kb genomic fragment. The 50-kb region is split in 2-kb non-overlapping windows and in each one the average mean(f_i) [A) and C)] and the variance var(f_i) [B) and D)] of the frequency f_i of the polymorphism class i is calculated. In A) the plots refer to a selective event in equilibrium populations ($\alpha = 2500$) that has been completed recently, whereas in C) to the non-equilibrium model of THORNTON and ANDOLFATTO (2006) ($\alpha = 2500$). The solid black lines refer to the singletons, the dashed black lines to the class '11', and the gray lines to the classes $2-10$. The dramatic change of the high-frequency derived alleles in A) contributes to the precise localization of the selective event. On the contrary, in C) the high-frequency derived SNPs are absent even in the proximity of the selective sweep. This is because the length of the branches of the coalescent tree that may generate high-frequency derived variants are very small due to the simultaneous action of the sweep and the bottleneck. Therefore, the observed polymorphisms (mostly singletons) are younger than the selective event and spread over the whole genomic region, obscuring the location of the selective sweep.

under the scenarios (i) selection versus neutrality in equilibrium populations (*i.e.* standard neutral populations), (ii) selection in equilibrium populations versus neutrality in non-equilibrium populations (*i.e.* populations that have experienced past demographic changes), and (iii) selection versus neutrality in non-equilibrium populations. The performance is assessed as follows. First, the false positive (FP) rate of the SVM is estimated. Using this false positive rate we compare the true positive (TP) rates of each test. Thus, all comparisons refer to the same false positive rate. Second, for the evaluation of the distance between the true and predicted targets we use only simulated results that survive the threshold defined by the false positive rate. Finally, for the non-equilibrium models with selection we implement a simple randomization process to assess the quality of results (see **Methods**).

SHH versus neutrality in equilibrium populations: We simulate a single selective sweep in the middle of a 50-kb genomic region using the *ssw* software (KIM and STEPHAN 2002). The parameter values have been chosen for their relevance to natural populations of *D. melanogaster*. Specifically, the parameter $\alpha = 2Ns$, where s is the selection coefficient of the beneficial mutation, assumes the values 500, 2500, and 5000 that are realistic for *D. melanogaster* (BEISSWANGER and STEPHAN 2008). For all datasets the mutation rate $\theta = 0.005$/bp, similar to the estimation of θ for the European population of *D. melanogaster* by LI and STEPHAN (2006). The scaled recombination rate ρ equals 0.05/bp, so that the ratio $\frac{\rho}{\theta} = 10$ (THORNTON and ANDOLFATTO 2006). The standard neutral simulations were performed using the same value of ρ. We used a sample size of 12 for all simulations.

Each realization of the selective sweep was compared with those of the standard neutral model that are obtained using $\theta_{NEU} = \theta_W = \frac{S_n}{h_n}$. θ_{NEU}. θ_{NEU} denotes the θ value used in standard neutral simulations, θ_W is Watterson's (1975) estimator of θ obtained using the number of segregating sites S_n of the selective sweep realization, and $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$. Thus, a selective sweep is compared with the standard neutral realizations that on average create the observed number of polymorphic sites [$F\theta$ procedure (RAMOS-ONSINS *et al.* 2007)]. Alternative approaches to calculate the threshold value may use the observed number of segregating sites S_n or to take into account the uncertainty on θ by considering a prior distribution of θ. In neutral equilibrium populations these approaches result in the same threshold values for the models tested in this study (Figure *S*2 in Appendix). Here, for the calculation of thresholds we use the $F\theta$ approach. Since, the null model is represented by an equilibrium standard neutral model, θ can be estimated using the estimator θ_W. Figure *S*2 shows that the cutoff value of the ω-statistic decreases as S_n increases and the opposite tendency is seen for the *SweepFinder* statistic.

Consistent with previous studies (JENSEN *et al.* 2007b) a selective sweep is discriminated

Table 1.1: Equilibrium neutrality versus selection in equilibrium populations

Parameter	Performance	SF	SF*	ω	ω*	SVM
$\alpha = 500$	TP (FP = 0.03)	0.85	0.97	0.13	0.14	0.9
	Median distance in bp from target (SD)	1728 (5597)	754 (1333)	528 (480)	540 (525)	-
$\alpha = 2500$	TP (FP = 0)	0.97	0.99	0.82	0.85	0.98
	Median distance in bp from target (SD)	5383 (4509)	4582 (3905)	789 (657)	794 (680)	-

Using the SVM approach a false positive rate (FP) is estimated for various parameter values. For this FP rate, the true positive rates (TP) of the various neutrality tests are compared. The median distance and the standard deviation (SD) are also shown. SF: original *SweepFinder*, SF*: modified *SweepFinder*, ω: ω algorithm with constant-size windows, ω*: ω algorithm with variable-size windows.

easily from the standard neutral model. Indeed as illustrated in Figure 1.4a, the ω_{MAX} and Λ_{MAX} are distributed to a large extent distinctly even for relatively small values of α (*e.g.* 500). Results are summarized in Table 1.1. Next, the distance between the true target of selection and the predicted target of selection is estimated (Table 1.1). The ω-statistic is more accurate than the *SweepFinder* and the median distance from the target of selection is about 0.5 kb. However, the performance of *SweepFinder* in discriminating the two scenarios is higher. Combining *SweepFinder* with the ω-statistic increases the classification performance (last column in Table 1.1).

SHH in equilibrium populations versus neutrality in non-equilibrium populations: Using simulations, selective sweeps have been generated as described above. For realizing past bottleneck events we used the LI and STEPHAN (2006) demographic history for the European population of *D. melanogaster*. We follow a similar approach as described in the previous section in order to assess the cutoff value. However, since the null hypothesis is not represented by the standard neutral model, θ_W is not an appropriate estimator of θ. Instead, we use the generalized unbiased estimator $\widehat{\theta} = \frac{2S_n}{E(T_c)}$, where $E(T_c)$ is the expected total length of the coalescent of *n* sequences (ZIVKOVIC and WIEHE 2008). $E(T_c)$ depends only on the demographic history of the population.

For large values of α ($\alpha = 2500$) the true positive rate of the statistics ω_{MAX} and Λ_{MAX} is greater than 70% when the false positive rate is 18% (Table 1.2). For the same false positive rate, the true positive rate of the modified version of *SweepFinder* is above 90%. However, when smaller selection coefficients (*e.g.* $\alpha = 500$) define the hitchhiking effect, the selective sweep may be inseparable from bottleneck scenarios similar to that inferred by LI and STEPHAN (2006), using the original version of *SweepFinder* or the ω-statistic (TP rates $< 10\%$, Table 1.2 and Figure 1.4b). The modified version of *SweepFinder* has a larger discrimination performance (true positive rate $\sim$ 40%). The low discrimination performance is indicated by the resemblance of genealogies between

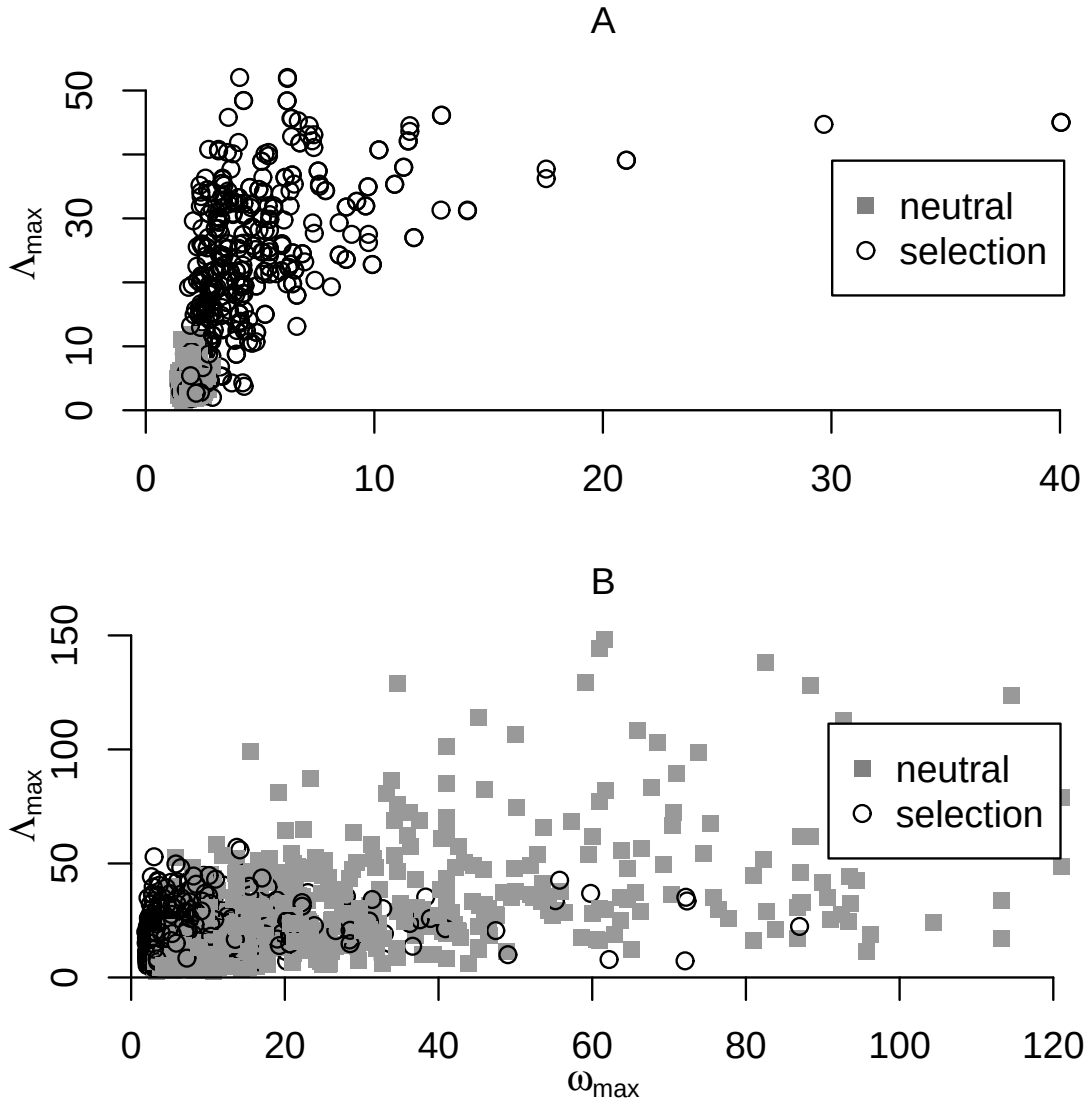

FIGURE 1.4: The joint distributions of Λ_{MAX} and ω_{MAX} in scenarios with and without selection. In A) we compare the joint distribution of Λ_{MAX} and ω_{MAX} between a model with selection ($\alpha = 500$) in a constant population and a standard neutral model. The overlap between the distributions is limited and the scenarios can be discriminated by the *SweepFinder* (y-axis) and to a lesser extent by the ω-statistic (x-axis). In B) we compare a model with selection ($\alpha = 500$) with a neutral model that has experienced a bottleneck as it has been inferred by LI and STEPHAN (2006). Neither of the statistics can discriminate accurately the two scenarios (see also Table 1.2). Notice that the scales of the statistics are different in A) and B).

bottleneck models and selective sweeps in constant populations (see also **Theoretical analyses**). The distributions of ω_{MAX} and Λ_{MAX} are largely overlapping as illustrated in Figure 1.4b. The SVM approach performs considerably better than any of the tests alone. The true positive rate is 75% when the false positive is 26% (Table 1.2). The main reason for the superior performance of the SVM approach is that it uses information about the distance of the peaks. In the scenarios with selection the target can be predicted accurately (Table 1.2), therefore the distance between the peaks is considerably smaller than in the neutral scenarios.

Table 1.2: Non-equilibrium neutrality versus selection in equilibrium populations

Parameter	Performance	SF	SF*	ω	$\omega^\star$	SVM
$\alpha = 500$	TP (FP = 0.26)	0.1	0.41	0.04	0.03	0.75
	Median distance in bp from target (SD)	899 (878)	522.982 (824)	423 (428)	603 (513)	-
$\alpha = 2500$	TP (FP = 0.18)	0.73	0.93	0.72	0.74	0.84
	Median distance in bp from target (SD)	3065 (3209)	2074 (3361)	917 (1653)	956 (1629)	-

SHH versus neutrality in non-equilibrium populations: In this section we examine the statistical performance of the neutrality tests to detect selection in a genomic region and assess the distance between the true and the predicted targets of selection. We focus on two bottleneck scenarios. The first one describes a deep and short-lasting bottleneck (model A), whereas the second scenario describes a shallow and long-lasting bottleneck (model B). In both cases the severity (*i.e.* the product depth $\times$ length) is the same ($= 0.375$ in units of $4N$), and the bottleneck begins (backward in time) at 0.01. The present effective population size is assumed 10^6, and the simulated region 50 kb. The recombination rate ρ for the whole region is set to 500. In the deep bottleneck scenario, the depth ($= \frac{\text{present population size}}{\text{bottlenecked population size}}$) equals 500 and the length 0.00075. In the shallow bottleneck scenario, the depth equals 20 and the length 0.01875.

Neutral simulations have been performed using Hudson's *ms* (HUDSON 2002) and simulations with selection using the *mbs* algorithm (TESHIMA and INNAN 2009). The design of simulations is as follows. In both cases we fix the number of polymorphic sites (=50) by employing broad uniform priors on θ and accepting only those instances that result in 50 segregating sites. This is justified by the dependence of the ω-statistic and *SweepFinder* on the number of segregating sites (Figures *S*2 and *S*3 in Appendix) and the large variance on segregating sites that neutral bottleneck scenarios generate. Furthermore, the rejection process guarantees that the total length of the tree, the posterior θ values and the number of segregating sites are coupled. The 25^{th} and 75^{th} quantiles of the posterior distribution of θ are 32 and 52, respectively, for the deep-bottleneck scenario and

32 and 48 for the shallow scenario; therefore the ratio $\frac{\rho}{\theta}$ is close to 10. In the simulations with selection, we examine scenarios of selective sweeps occurring recently (between the present and the bottleneck; sweep in phase 1), within the bottleneck (sweep in phase 2), and after the bottleneck (backward in time; sweep in phase 3). The parameters of the models with selection are described in Table 1.3 and Table 1.4 for the deep and shallow models, respectively. Similar to the neutral cases, a broad uniform prior on θ has been used, and we condition on observing 50 segregating sites. The posterior range of θ depends on the timing of the selective sweep; therefore, the ratio $\frac{\rho}{\theta}$ is close to 10 when the sweep is either recent or old, but it decreases when the selective sweep occurs within the bottleneck phase.

Table 1.3: Neutrality versus selection in non-equilibrium populations (deep bottlenecks)

Parameter	Performance	SF	SF*	ω	$\omega^\star$	SVM
	TP (FP=0.51)	0.64	0.66	0.39	0.49	0.71
phase 1	Median distance in bp from target (SD)	10813 (6768)	10497 (6832)	11986 (6595)	10239 (6186)	-
	Random target distance (SD)	11053 (6827)	11308 (6803)	11575 (6645)	11944 (6945)	-
	TP (FP=0.20)	0.62	0.64	0.36	0.44	0.73
phase 2	Median distance in bp from target (SD)	9666 (6531)	10828 (6896)	11854 (6500)	10469 (6123)	-
	Random target distance (SD)	11508 (6885)	11397 (6808)	11877 (6750)	11555 (6804)	-
	TP (FP=0.08)	0.72	0.78	0.63	0.12	0.97
phase 2*	Median distance in bp from target (SD)	9512 (6659)	10986 (6977)	10905 (6482)	11328 (6487)	-
	Random target distance (SD)	12067 (6983)	12265 (6920)	11647 (6950)	13236 (7213)	-
	TP (FP = 0.56)	0.53	0.55	0.48	0.46	0.63
phase 3	Median distance in bp from target (SD)	10377 (6831)	10845 (6833)	11342 (6662)	10624 (6541)	-
	Random target distance (SD)	12202 (6908)	11641 (6860)	12151 (6920)	12220 (6824)	-

A deep bottleneck, named model A, is examined. The ratio $\frac{\text{present population size}}{\text{bottlenecked population size}} = 500$ and the length of the bottleneck is 0.00075. A beneficial mutation may appear within each phase of this 3-epoch model (where time is measured backwards in units of $4N$ generations): a recent sweep at time 0.01 (sweep in phase 1), a sweep within the bottleneck at time 0.0107 (sweep in phase 2), and an old sweep at 0.115 (sweep in phase 3). The selection coefficient is 0.002. Additionally, in the 'sweep in phase 2*' model we describe a sweep which completes within the bottleneck ($s = 0.8$). The true positive rates of the neutrality tests are shown for each sweep model. The other rows depict the distance between the predicted and true targets and the random expectations for the distance.

First, we examined the performance of the ω-statistic and *SweepFinder* to detect whether a genomic region has been shaped by positive selection. Results are presented in Table 1.3 and Table 1.4. For all comparisons, we used the false positive rate that is reported by the SVM. Then, we compare the TP rates between the various tests; the performance of a test is better when the TP rate is higher. The combination of *SweepFinder* and ω-statistic performs better than each test (SVM column in Table 1.3 and Table 1.4). Also, *SweepFinder* outperforms the ω-statistic. In model A (deep bottleneck), when the sweep is either recent or old, the discrimination between neutral and selective models becomes problematic; when the false positive rate is about 50%, the true positive is as low as 70% and 63%, respectively, for the SVM approach. For the separate tests, the performance is even lower. This result suggests that recent or old selection in populations that have experienced deep bottlenecks cannot be discriminated from neutrality. However, when selection has occurred within the bottleneck phase, the false positive rate decreases to 20% and the true positive rate is 73% for the SVM and about 10% lower for the *SweepFinder* (Table 1.3, sweep phase 2). Higher discrimination performance is achieved when the sweep completes within the bottleneck (Table 1.3, sweep phase 2*), but this requires unrealistically high values of *s*.

In model B (shallow bottleneck), the discrimination performance is slightly better than that of model A. However, again the most challenging scenarios are either recent or old sweeps and the performance increases when the sweep occurs within the bottleneck phase (Table 1.4). Finally, the distances between the true target and the predicted target of selection are estimated. For both models A and B the distance is large and close to random expectations (Table 1.4).

Distinguishing RHH from neutrality in equilibrium populations: In contrast to single selective sweep (SHH) models, recurrent selected substitutions occur randomly along a chromosome according to a time-homogeneous Poisson process at a rate *v* per generation (KAPLAN *et al.* 1989; WIEHE and STEPHAN 1993; STEPHAN 1995). Well-known patterns of SHH models are modified under RHH. As an example, the SFS is skewed toward the rare variants; however, the excess of high-frequency derived alleles decreases (KIM 2006; JENSEN *et al.* 2008). Previously, JENSEN *et al.* (2007b) have shown that it is difficult to separate RHH models from neutrality based on ω_{MAX}-values or site frequency spectrum statistics. We explore the same problem with our new versions of the ω-statistic and the *SweepFinder* algorithm. Using the software developed by JENSEN *et al.* (2008) we simulated 100-kb genomic regions for a given reduction of heterozygosity (WIEHE and STEPHAN 1993), namely $\frac{H_{RHH}}{H_{NEU}} = 0.05, 0.25, 0.5, 0.75$ or 0.95. $\frac{H_{RHH}}{H_{NEU}}$ denotes the ratio of heterozygosity in the RHH model to the heterozygosity in the absence of selective sweeps. The selection coefficient $s = 0.0001$ or 0.01. The null hypothesis is represented by the standard neutral model.

Table 1.4: Neutrality versus selection in non-equilibrium populations (shallow bottlenecks)

Parameter	Performance	SF	SF*	ω	$\omega^\star$	SVM
	TP (FP=0.27)	0.46	0.49	0.22	0.25	0.5
phase 1	Median distance (SD)	10116 (6872)	10691 (7001)	10268 (6658)	10868 (6670)	-
	Random target distance (SD)	11604 (6862)	11452 (6835)	10744 (6895)	11192 (7115)	-
	TP (FP=0.22)	0.58	0.56	0.27	0.32	0.6
phase 2	Median distance (SD)	10233 (6866)	11059 (6807)	11659 (6721)	11531 (6643)	-
	Random target distance (SD)	11725 (6889)	11375 (6855)	10846 (6829)	11245 (6882)	-
	TP (FP=0.35)	0.67	0.74	0.65	0.4	0.67
phase 2*	Median distance (SD)	9610 (6814)	10148 (6962)	11356 (6683)	10680 (6539)	-
	Random target distance (SD)	11906 (6889)	12102 (6846)	12432 (6894)	11583 (7079)	-
	TP (FP = 0.25)	0.4	0.38	0.23	0.27	0.46
phase 3	Median distance (SD)	10232 (6710)	10447 (6744)	11693 (6965)	10829 (6625)	-
	Random target distance (SD)	11372 (6906)	11574 (6857)	11666 (6817)	13068 (6914)	-

A shallow bottleneck, named model B, is examined. The ratio $\frac{\text{present population size}}{\text{bottlenecked population size}} = 20$ and the length of the bottleneck is 0.01875. A recent sweep at time 0.01 (sweep in phase 1), a sweep within the bottleneck at time 0.0107 (sweep in phase 2), and an old sweep at 0.115 (sweep in phase 3) are described. The selection coefficient in the model 'sweep in phase 2*' is 0.1.

The null model used for the *SweepFinder* calculations and represented by the SFS of the population prior to the selective sweep in the SHH cases (n-SFS) cannot be described precisely by the standard neutral model. The population size is assumed to be constant. However, since adaptive mutations occur according to a time-homogeneous Poisson process it remains obscure what the 'prior to the sweep' SFS should be. Here, we follow two approaches. First, we assume that the n-SFS is derived from the standard neutral model and second, the n-SFS is obtained from the genomic region itself. Clearly, both approaches are approximations. On one hand, using the standard neutral model we increase the sensitivity of the *SweepFinder*. On the other hand, the nucleotide polymorphism patterns of the genomic region under investigation have been shaped by selective sweeps, so the n-SFS forms a conservative null model with small sensitivity. However, if real data are consistent with the RHH model, the standard neutral model cannot be supported as a null model since the whole genome will be affected by recurrent sweeps.

When the n-SFS is derived from the data itself then the power of the *SweepFinder* is greater for small values (*e.g.* 0.0001) than large values (*e.g.* 0.01) of the selection coefficient *s* (Figure *S*4 in Appendix). Even if this appears to be counterintuitive, it is reasonable because when *s* is small the footprints of the selective sweep are local, and a large fraction of the genome remains neutral. On the other hand, for large values of *s* almost the entire genomic region may be affected by RHH contradicting the assumption of the *SweepFinder* test that only a small fraction of the genome has been shaped by positive selection (Figure *S*4).

Under RHH models selective sweeps occur in different genomic locations during the evolution of the population following a time-homogeneous Poisson process (WIEHE and STEPHAN 1993). When subgenomic data are analyzed it is possible that the target of selection is either inside or outside of the sequenced genomic region. Furthermore, since selective events occur with a certain probability per generation (WIEHE and STEPHAN 1993), patterns of polymorphism are shaped by both old and new selective events. However, the ω-statistic and *SweepFinder* are based on the assumption that a single selective sweep has just been completed. Thus, it is important to test whether the algorithms are able to predict the correct position of the adaptive events.

Incorporating a fraction of monomorphic sites into *SweepFinder* analysis increases the precision of the algorithm (Figure *S*5 in Appendix). Similarly, the variable-size sliding window approach appears more accurate than the constant-size sliding window method for high cutoff values. When $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.25$, *SweepFinder* and the ω-statistic predict that a target of selection is within a 5-kb distance from a true selective sweep position in about 40% of the cases. However, this fraction becomes smaller for higher values of $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}}$ (Figure *S*5).

1.6 Discussion

The demography of natural populations: A major challenge of population genomics studies is to identify the loci that driven by positive selection contribute to the adaptation of natural populations, and to localize the beneficial mutation accurately (KIM and STEPHAN 2002; SABETI *et al.* 2002; JENSEN *et al.* 2005; NIELSEN *et al.* 2005; AKEY 2009; NIELSEN *et al.* 2009; PICKRELL *et al.* 2009). In order to address these questions, it is important to consider the demographic history of the population, as this neutral non-equilibrium model represents the null (LI and STEPHAN 2006; THORNTON and ANDOLFATTO 2006). Since the standard neutral model does not reflect accurately the demography of most natural populations, neutrality tests should not be performed using the standard neutral scenario as the null model. In this study, we examined two bottleneck scenarios that are relevant to the demographic history of the European population of *D. melanogaster* (LI and STEPHAN 2006; THORNTON and ANDOLFATTO 2006). The properties of the coalescent trees that underlie these demographic models differ considerably. In a recombining genomic region, the model inferred by LI and STEPHAN (2006) produces both star-like short coalescent trees, and genealogies with long internal branches. Star-like genealogies are generated less frequently by the THORNTON and ANDOLFATTO (2006) model (Figures 8 and 1.2). As a consequence, the null distributions of the neutrality statistics may differ. Thus, inferring the demographic history of a population is a prerequisite for performing genomic scans for selective sweeps, which has been shown to be a challenging task (MYERS *et al.* 2008).

Separating single selective sweeps from neutral models: When the value of the selection intensity α is large, the joint distribution of Λ and ω overlaps only partially between a model of selection in an equilibrium population and the bottleneck model inferred by LI and STEPHAN (2006). However, for smaller values of α the two distributions overlap greatly. A useful approach for classifying an observation as either a neutral or selective model is by combining the Λ and ω profiles. Here, we use the distance between the peaks and the correlation of ω and Λ. These features can be used in a classifier (*e.g.* SVM). Training requires that there are known instances of both neutral and selective models. For simple selective and neutral models this is currently possible, using coalescent-based programs. However, it remains challenging for more complicated scenarios. Forward simulations provide greater flexibility when selective events occur in non-equilibrium populations and they can be used efficiently when the population size is relatively small (*i.e.* on the order of thousands) or diffusion scaling applies (HOGGART *et al.* 2007; CHADEAU-HYAM *et al.* 2008; HERNANDEZ 2008).

The rationale for employing combinations of Λ and ω is that under a selective model the two

statistics assume high values close to the target of selection. This implies that the target of selection can be localized accurately. Under selection models in equilibrium populations this assumption is met even for small α values. Modifying *SweepFinder* to include a fraction of non-polymorphic sites in the analysis increased the accuracy of the algorithm and the performance in separating neutral scenarios from scenarios with selection. Furthermore, both versions of the ω-statistic, the constant- and the variable-size sliding window approach, are very accurate for selection models in equilibrium populations.

However, in severe non-equilibrium scenarios (*e.g.* the estimated bottlenecks of LI and STEPHAN (2006) and THORNTON and ANDOLFATTO (2006)), when selection and past demographic changes occur within the same model, the target of selection cannot be predicted, neither by *SweepFinder* nor by the ω-statistic. The accuracy of the target prediction when a selective sweep has occurred within the bottleneck period is comparable to that of randomized experiments. The reason is that polymorphism valleys and short coalescent trees may extend over large genomic regions, and the often used sweep signature of an excess of high-frequency derived alleles vanishes. This result should be taken into account when regions of strong and recent positive selection are identified in genome scans. Since natural populations can be described by equilibrium demographic models only rarely, the true target of selection may be tens of kilobases away from the predicted target.

In the case of a severe bottleneck, such as the model A, recombinants (carrying the selected mutation and the derived neutral allele) are most likely formed in the early period of the selective phase (forward in time), but they will be lost with high probability due to drift after the population size crashes. Therefore, high-frequency derived variants may not be observed. In contrast, the frequency of rare variants (singletons) will dramatically increase. Therefore, based on site frequency spectrum it is possible to discriminate, to some extent, neutral from non-neutral scenarios (Table 1.3).

The analysis of the likelihood curves of *SweepFinder* can provide further insights into the technical reasons that, in the cases of selection in non-equilibrium populations, make the prediction of the target of selection challenging. *SweepFinder* implements a model of selective sweep which assumes that each observed SNP was existing prior to the sweep. It uses the compound parameter $\gamma = \frac{r}{s} log(2N)$ (named α in NIELSEN *et al.* (2005)) and the position x where the selective event occurred. (Here r denotes the recombination rate per bp). As Figure *S*6 (Appendix) illustrates, low- and high-frequency SNPs affect the likelihood in a similar way by contributing high values in the proximity of the sweep. Examining how the SFS changes over a genomic region under an equilibrium demographic model with selection and the THORNTON and ANDOLFATTO (2006) model with selection ($\alpha = 2500$), it is apparent that there is a dramatic increase of the class '$n-1$'

in the proximity of the selective sweep in the equilibrium model (Figure 1.3), but a very slight change of singletons in the non-equilibrium model. In the equilibrium-model case the precise localization of the sweep is possible, due to the spatial patterns of the rare and high-frequency derived variants. However, in the THORNTON and ANDOLFATTO (2006) model with selection this pattern vanishes, the high-frequency derived variants disappear and the singletons spread over the whole genomic region. Thus, the target of selection cannot be estimated accurately.

It should be noted, however, that the poor performance of *SweepFinder* and the ω-statistic under the non-equilibrium models (bottlenecked populations with selection) does not imply that the performance of the tests is poor under any non-equilibrium model with selection. These models represent extreme cases that violate major assumptions of the algorithms. The slightly improved performance of the machine learning approach is due to the fact that it uses information from the sweep scenarios and, furthermore, it combines information from both the ω-statistic and *SweepFinder*.

Studying a scenario where a selective event took place in a bottleneck period is of great biological importance. Often, population bottlenecks are associated with a major migration event. For example, the bottleneck inferred by LI and STEPHAN (2006) for the European population of *D. melanogaster* describes the colonization of Europe from the African ancestral population. Therefore, positive selection may have occurred in the new habitat that contributed to the adaptation of flies to the environmental conditions of Europe. As Tables 1.3 and 1.4 show, the performance of the tests (especially the SVM, and to a lesser extent, the *SweepFinder*) is high when the sweep occurs within the bottleneck. This suggests that the approaches tested in this study can be used for the detection of selective sweeps in populations that have recently migrated to new environments. Furthermore, Tables 1.3 and 1.4 suggest that the power of SFS-based tests is higher than LD-based tests.

A difficulty which arises from using simulations with selection in order to train the algorithms is that the parameters of the scenarios with selection are unknown, *i.e.* the selection intensity α, the position of the sweep x, and the time at which the sweep occurred. In the models that we presented it was assumed that these parameters are known. However, when real data are analyzed these parameters are generally unknown, and moreover there are no methods available that can estimate them in scenarios with past demographic changes. Thus, heuristic approaches have to be used. First, the position x can be assumed to be in the center of the fragment. Then, in real-data analysis overlapping windows should be used so that there will exist windows where x is located near their center. The time of the sweep should be recent ($< 0.1N$). In the classical approach this parameter is also implicitly specified by assuming that the sweep has just been complete. Finally,

the selection intensity can be drawn from a prior uniform distribution. In this case the training set is composed of a mixture of models with various selection intensities.

Recurrent selective sweep analysis: Recurrent selective sweeps invalidate the assumption that a single hitchhiking event has just been completed. In agreement with JENSEN *et al.* (2007b), we find that for greater rates v of selective events per generation the power of the tests increases for a given $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}}$. One possible explanation is that for smaller v a few strong selective sweeps have occurred which affect a large portion of the genome and shift the SFS of large genomic regions. Thus, the local characteristic of the signature of a selective event is lost. Another possible explanation is that for smaller v the selective events are old on average and the signature of selective sweep has faded away (JENSEN *et al.* 2007b).

The variable-size sliding window approach increases the accuracy of the ω-statistic to predict the target of selection. However, the performance is still poor. In ∼20% of the peaks above a certain threshold found in a scan of a given genomic region, the real position of the sweep is located within a 5-kb distance. The performance of the constant-size sliding window is about half that of the variable-size approach and comparable to the randomization experiments. A similar improvement has been achieved with the modified *SweepFinder* algorithm. RHH models imply that adaptive substitutions occur at a time-homogeneous rate, *i.e.* uniformly in the history of the population. This assumption may be violated in domesticated populations or in populations that experienced environmental changes. Thus, an increase of the performance of the tests (lower false positive rate, greater accuracy in target prediction) may result when RHH models are incorporated within the *SweepFinder* or the ω-statistic algorithms.

Recurrent selective sweep parameters such as the rate v of adaptive substitutions and the decrease of heterozygosity have been estimated recently. JENSEN *et al.* (2008) and LI and STEPHAN (2006) have estimated that heterozygosity has decreased in genomic regions of normal recombination by 50% whereas the estimate of MACPHERSON *et al.* (2007) and ANDOLFATTO (2007) is about 20% (*i.e.* $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.8$). We examined the performance of the *SweepFinder* and the ω-statistic for various levels of heterozygosity reduction, $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.25, 0.5, 0.75$, and 0.95, and selection coefficients $s = 10^{-2}$ and 10^{-4} (Figure 1.5). The power of *SweepFinder* is greater for the LI and STEPHAN (2006) and JENSEN *et al.* (2008) estimations than that of MACPHERSON *et al.* (2007) and ANDOLFATTO (2007), given that selection is strong ($s = 10^{-2}$). For $s = 10^{-4}$ the differences in the performance of *SweepFinder* for various levels of $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}}$ are small. The reason is that for $s = 10^{-4}$ the diversity is similar for values of $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}}$ between 0.05 and 0.95. This may be due to inaccuracies of the RHH theory when s is small or due to the stochastic trajectory of the beneficial mutation (COOP and GRIFFITHS 2004; SPENCER and COOP 2004).

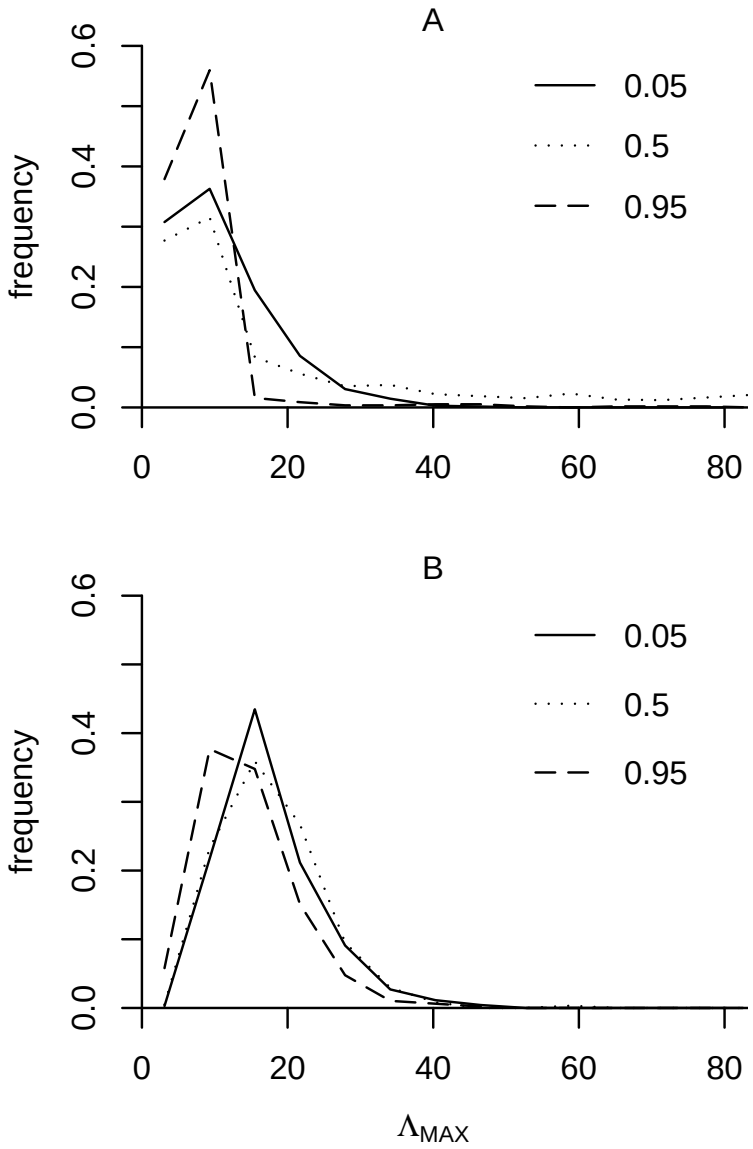

FIGURE 1.5: The distributions of Λ_{MAX} for various levels of the decrease of heterozygosity and $s = 10^{-2}$. Each distribution is discrete and the size of each bin has been set to 6. A) For $\frac{H_{\text{RHH}}}{H_{\text{NEU}}} = 0.05$, 0.5, and 0.95 the cutoff values (95$^{\text{th}}$ percentile) are 5.7, 9.7, and 11.9, respectively, and the sensitivities of the test (percentage of true positives) given the cutoff values are 0.74, 0.48, and 0.07. The power of *SweepFinder* is greater for the LI and STEPHAN (2006) and JENSEN *et al.* (2008) estimations than those of MACPHERSON *et al.* (2007) and ANDOLFATTO (2007) because selection is strong ($s = 10^{-2}$). B) When $s = 10^{-4}$ the amount of diversity is similar for $\frac{H_{\text{RHH}}}{H_{\text{NEU}}} = 0.05$, 0.5, and 0.95. Therefore, the performance of *SweepFinder* is relatively independent of the $\frac{H_{\text{RHH}}}{H_{\text{NEU}}}$.

Time of the selective sweep: For SHH models (in demographic equilibrium) we assume that the selected mutation has reached fixation very recently. The selective model that underlies the *SweepFinder* algorithm assumes a recent and strong selective sweep. Therefore, the power of *SweepFinder* is expected to be higher for recently completed hitchhiking effects. Indeed, simulations have shown that the power decreases exponentially after the selective sweep (P. Pavlidis, unpublished results). It should be mentioned that the demographic scenario that follows the selective sweep (*i.e.* between the time of completion of the selective sweep and the time of sampling) affects the performance of *SweepFinder*. Simulations have shown that if the completion of a selective sweep is followed by population expansion, the performance of the likelihood ratio test implemented in *SweepFinder* remains high even after the completion of the selective sweep (P. Pavlidis, unpublished results). The rationale behind this is that a population expansion decreases the coalescent rate; therefore the return to the equilibrium SFS is slower and the signature of the selective sweep is preserved for a longer period. In contrast to *SweepFinder*, which is based on the low- and high-derived variants, the ω-statistic is more sensitive to the time since the completion of the selective sweep. Indeed, the LD pattern captured by the ω-statistic, vanishes rapidly (JENSEN *et al.* 2007b), comparable to the fixation rate of the high-frequency derived alleles (KIM and STEPHAN 2000; PRZEWORSKI 2002; JENSEN *et al.* 2007b).

Overlapping selective sweeps: In the present study we focused on non-overlapping selective sweeps. The RHH model we have used describes successive and non-overlapping selective events. CHEVIN *et al.* (2008) have shown that two interfering selective sweeps may modify the pattern of linked neutral variation. A related process, when the targets of selection are located closely to each other in the genome, causes trafficking (KIRBY and STEPHAN 1996; KIM and STEPHAN 2003). A most extreme scenario, which describes the appearance of beneficial mutations at the same site, is described as 'soft' sweep (HERMISSON and PENNINGS 2005). Soft sweeps may emerge during the evolution of organisms (*e.g.* Plasmodium) with high mutation rates (NAIR *et al.* 2007). Conversely, they may be of limited importance in the evolution of *D. melanogaster* or *H. sapiens*, for instance. The patterns of neutral variation under these selective scenarios are different from those of single selective events. For example, the skew of Tajima's *D* toward negative values vanishes in the interference scenarios described by CHEVIN *et al.* (2008), and can be even positive between the selected sites. In general, SFS-based approaches may not work under overlapping selective sweeps because the frequency of the class of polymorphisms in intermediate frequency may be quite large. In such cases, LD-based statistics can be useful because a multitude of extended haplotypes may exist on the left and right sides of the selected region (SABETI *et al.* 2002; VOIGHT *et al.* 2006; TANG *et al.* 2007).

Machine learning approaches in population genetics: Machine learning approaches are widely used in a variety of applications from image processing to classification of microarrays. Here, we are interested in the subfield of machine learning that is related to supervised learning or classification. Typically, in a classification problem a training set teaches the algorithm to predict the class label of an input object (DUDA *et al.* 2000; HASTIE *et al.* 2001). The goal is to decide between a selective and a neutral model. However, classifying a dataset as either neutral or selective is challenging because the parameters of the neutral and selective models are unknown. Therefore, parameter estimation is required prior to the classification. In the cases that an equilibrium model with selection is employed, the selection intensity α can be estimated using the *clsw* software (KIM and STEPHAN 2002) or the *SweepFinder* algorithm (given that ρ is known). To our knowledge, currently the only method able to estimate α given a non-equilibrium (stepwise) model with selection has been developed by LI and STEPHAN (2006). On the other hand, several approaches exist for the estimation of parameters in a neutral demographic model (NIELSEN 2000; EXCOFFIER *et al.* 2005; LI and STEPHAN 2006; HEY and NIELSEN 2007). Usually, these approaches require multiple loci in order to infer the demographic parameters of a population. The next step in a classification problem is feature selection, which aims at using a subset of the features available from the data. Here, Λ_{MAX}, ω_{MAX}, and their combinations (distance between peaks and correlation of ω and Λ) have been used. Combining ω and Λ is powerful in comparisons between equilibrium models with selection and neutral non-equilibrium models when the selection intensity is small (Table 1.2). Alternatively, various summary statistics, such as TAJIMA (1989)'s *D*, FAY and WU (2000)'s *H* or Z_{nS} (KELLY 1997) can be used. Our choice is based on the fact that *SweepFinder* uses SFS information whereas the ω-statistic is based on LD. The choice of the classification technique is important and depends on the problem and the nature of the data. Here, we demonstrate an application using the SVM classifier (with the radial kernel), as it is implemented in the 'e1071' package of the R-project. To our knowledge, there are no studies in separating neutral from selective scenarios that use supervised learning approaches. Future work will provide insight into the feature selection problem and will also evaluate the performance of the supervised learning approaches.

Chapter 2

Recent strong positive selection on *Drosophila melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by population genomic analysis

Nicolas Svetec[†,*], Pavlos Pavlidis[†], and Wolfgang Stephan[†]
[†] Department of Biology II, Ludwig-Maximilians-University Munich, 82152 Planegg, Germany

Molecular Biology and Evolution 2009, 26:1549–1556

2.1 Abstract

Based on nearly complete genome sequences from a variety of organisms, data on naturally occurring genetic variation on the scale from hundreds of loci to entire genomes have been collected in recent years. In parallel, new statistical approaches (such as the *CLR* and *SweepFinder* tests) have been developed to infer evidence of recent positive selection from these data and to localize the target of selection. Here we apply these methods to the X chromosome of *Drosophila melanogaster* in an effort to map genes involved in ecological adaptation. Using slight modifications of these tests that increase their robustness against past demographic changes, we detected evidence that recent strongly positive selection has been acting on a 2.7-kb region in an ancestral African population. This region overlaps with the 3' end of *HDAC6*, a gene that encodes a newly characterized stress

surveillance factor. HDAC6 is an unusual histone deacetylase being localized in the cytoplasm. Its ubiquitin-binding and tubulin-deacetylase activities suggest that HDAC6 is very different from other histone deacetylases. Indeed, recent discoveries have shown that HDAC6 is a key regulator of cytotoxic stress resistance.

2.2 Introduction

Recent advances in population genomics allow us to detect footprints of strong positive selection in the genome and to identify the targets of selection on the scale of individual genes (reviewed in PAVLIDIS *et al.* 2008). Based on nearly complete genome sequences from a variety of organisms data on naturally occurring DNA sequence variation from hundreds of loci to entire genomes have been collected in the past 5 years. Most of these studies concentrated on *D. melanogaster* (*e.g.* GLINKA *et al.* 2003; ORENGO and AGUADÉ 2004) and humans (*e.g.* AKEY *et al.* 2004; INTERNATIONAL HAPMAP CONSORTIUM 2007). In parallel, new statistical tests have been developed to infer evidence of recent positive selection from these data (KIM and STEPHAN 2002; JENSEN *et al.* 2005; NIELSEN *et al.* 2005).

These new tests are based on the hitchhiking model by MAYNARD SMITH and HAIGH (1974). When a beneficial mutation arises in a population and goes to fixation driven by positive selection ('selective sweep'), theory predicts the emergence of a specific polymorphism pattern: (i) diversity vanishes around the site of selection, (ii) the site frequency spectrum (SFS) of polymorphisms is shifted toward low- and high-frequency derived variants (BRAVERMAN *et al.* 1995; FAY and WU 2000), and (iii) linkage disequilibrium (LD) is elevated in the early phase of the fixation process (KIM and NIELSEN 2004; STEPHAN *et al.* 2006). Importantly, the width of the valley of reduced variation is mainly determined by the ratio of the rate of recombination around the site of selection and the strength of selection.

A multitude of studies has used the predictions of the hitchhiking model to detect footprints of positive selection in the genome of various organisms, estimate the strength of selection, and map the target of selection (PAVLIDIS *et al.* 2008). However, demographic factors such as population size bottlenecks may stochastically produce patterns of nucleotide diversity across the genome that resemble those of selective sweeps. Therefore, a major challenge of these analyses has been (and still is) to distinguish the effects of selection from those of demography. Recent progress in this area of research could be made based on the insight that demography affects the entire genome whereas selection acts on individual loci. This improved the robustness of the tests for selection (JENSEN *et al.* 2005; NIELSEN *et al.* 2005; THORNTON and JENSEN 2007).

The candidate regions of selection identified by these tests, however, were generally very large (often $\sim$ 100 kb) and contained many genes ($\sim$ 10). This is particularly the case for humans (*e.g.* WILLIAMSON *et al.* 2007). Although some progress has been made in *Drosophila* (POOL *et al.* 2006; JENSEN *et al.* 2007a; ORENGO and AGUADÉ 2007), a major challenge ahead is to develop strategies that help to narrow down the target regions of selection such that it is possible to map the site of selection to individual genes or gene regions. This is essential for ultimately understanding adaptation at the functional level.

Here we use selection mapping to identify genes in *D. melanogaster* that may have been involved in ecological adaptation. We were able to identify a 2.7-kb region as the putative target of selection that contains the last exon of *HDAC6* harboring a ubiquitin-binding domain. HDAC6 is an unusual histone deacetylase with two catalytic domains and is localized in the cytoplasm. Its activities (ubiquitin binding and tubulin deacetylase) mark a distinct departure of HDAC6 from the known action of other HDACs. Recent discoveries have shown that HDAC6 is a key regulator of cytotoxic stress resistance (reviewed in MATTHIAS *et al.* 2008). It appears to be both a sensor of stressful environmental stimuli and an effector, which mediates and coordinates appropriate cell responses.

2.3 Materials and Methods

***Drosophila* lines and DNA sequencing:** DNA sequence data were collected from 12 highly inbred lines sampled in Africa (Lake Kariba, Zimbabwe). Furthermore, sequence data were obtained from 12 inbred European lines from The Netherlands. Both samples are described in detail in GLINKA *et al.* (2003). All *Drosophila* strains were kept at 23°C in glass bottles of 250 ml containing 80 ml standard cornmeal and yeast medium under a 6-18 dark-light cycle with 45% humidity.

DNA primers were designed based on the *D. melanogaster* genome sequence (flybase) and obtained from Metabion (Martinsried, Germany). Genomic DNA from each line was extracted from pools of 20 females using the Puregene DNA isolation kit (Gentra System, Minneapolis). Short DNA fragments of about 300 to 700bp long were amplified by standard PCR using the Taq DNA polymerase recombinant kit (Invitrogen, Carlsbad, USA). PCR products were purified using the Exosap-It kit (USB, Cleveland) and sequence reactions were conducted with ABI PRISM Big Dye Terminator v1.1. Sequence data were then obtained by an ABI 3730 DNA analyzer (Applied Biosystems +Hitachi, Foster City, USA).

Sequence editing and alignments were performed with the DNAstar software package, including Editseq, Seqman and Megalign (DNASTAR, Madison, USA). Alignments were performed

using the ClustalV option of Megalign. However, in cases of ambiguous alignments, we manually chose the most parsimonious scenario. Insertion and deletion polymorphisms were excluded from further analysis. Absolute positions of the DNA sequence follow the Flybase release 5.10.

Mapping strategy: To identify and map the target of selection, we proceeded as follows. First, we selected a subgenomic region of about 70 kb on the X chromosome that contained several ecologically interesting genes, including a gene encoding a putative antifreeze protein (*CG6227*). This region partially overlaps with the window 47 in LI and STEPHAN (2006). Re-sequencing an additional (limited) number of short fragments of 500-600 bp in the 70-kb subgenomic region, we found very low levels of variation across most of the region in the European sample (data not shown), while the valley of reduced variation in the African sample appeared much narrower; *i.e.* the situation was similar as in the case of the *roughest* and *wapl* regions (POOL *et al.* 2006; BEISSWANGER *et al.* 2006). To be able to localize the target of selection as precisely as possible, we therefore decided to follow the same strategy as in the *wapl* analysis (BEISSWANGER and STEPHAN 2008) and concentrated on the African sample (see 'Standard analyses of a candidate region of selection' in the Results section). In a second step, we narrowed this 70-kb region down to 22 kb, re-sequenced this segment completely, and applied the specific tests for selective sweeps to this region (see Results).

Outlier analysis: We used DnaSP 4.50.3 (ROZAS *et al.* 2003) to calculate the basic summary statistics π, θ_W, Tajima's *D* (TAJIMA 1989), divergence, Fu and Li's *D* (FU and LI 1993) and Fay and Wu's *H* (FAY and WU 2000). Divergence was calculated between the sample from the African population of *D. melanogaster* and the available online release of the *D. simulans* sequence (Flybase consortium; `http://www.flybase.org`). The ancestral states were defined using either *D. simulans* or (when not available) its close relative *D. sechellia*.

We compared the mean value of each summary statistic of the 70-kb candidate region to its average value obtained for the whole X chromosome (OMETTO *et al.* 2005). For each summary statistic, we used the Mann-Whitney test to infer whether the region represents an outlier compared to the rest of the X chromosome.

Ascertainment bias correction: THORNTON and JENSEN (2007) describe an approach that generates a uniform distribution of p-values when some of the assumptions of the neutrality tests are violated. They study cases when past demographic events have shaped the polymorphism patterns of a subgenomic region, which is a biased sample based on a priori information (for example, from a genome scan). The *HDAC6* subgenomic region was selected based on the genes in this region that may contribute to the ecological adaptation of *D. melanogaster*. Even if such a sampling is not random, it is unclear whether it generates any bias on selective sweep scanning

and how to sample conditional on this biological information.

Performing a genome scan analysis, LI and STEPHAN (2006) discovered a 100-kb fragment that overlaps with the *HDAC6* region and showed evidence of recent positive selection in the European population of *D. melanogaster*. Among the fragments LI and STEPHAN (2006) analyzed was a 560-bp fragment located within the *HDAC6* subgenomic region that contained no polymorphic sites. This information was not considered important for the initial choice of the 70-kb region. However, we decided to include it into the analysis as *a priori* information making this analysis more conservative. Thus, we simulate a sample of 24 lines (12 European and 12 African ones) according to the demographic scenario inferred by LI and STEPHAN (2006). Conditioning on the existence of a monomorphic 560-bp fragment within the European sample, we create the null distribution of the neutrality test statistics used in this paper.

Composite Likelihood Ratio (*CLR*) test: The *CLR* test (KIM and STEPHAN 2002) was used to infer selection. It computes the composite-likelihood ratio (Λ_{CLR}) between a standard neutral model and a selective sweep model. The null distribution of the statistic is derived using the approach described in the 'Ascertainment bias correction' section (see also Figure 2.1). This modification follows a suggestion of THORNTON and JENSEN (2007) who showed that the false positive rate can be controlled if the correct demographic null model is used. For the generation of the simulated datasets we used the estimated value of the parameter θ_W (0.0499) under the demographic scenario of Figure 2.1. Furthermore, the B test of the KIM and STEPHAN (2002) method was performed because it is more conservative. The *CLR* test was also used to estimate the target site of selection. However, its confidence interval could not be determined (in contrast to BEISSWANGER and STEPHAN (2008)), as population recombination rate was too high to run simulations of the sweep model in reasonable times.

***SweepFinder* test:** To infer selection, we also used the *SweepFinder* test. It takes into account the SFS of the whole chromosome (background SFS) in order to calculate the likelihood of the neutral model. Non-polymorphic sites were excluded from the analysis, as NIELSEN *et al.* (2005) suggest. *SweepFinder* uses the same principles as the *CLR* test: by comparing two hypotheses, a model of neutral evolution and a model of a selective sweep that just completed, it calculates the maximum likelihood estimates of the position of the beneficial allele as well as the strength of selection. Additionally, it reports the likelihood ratio Λ_{SF} between the null and the alternative model. Similarly to the *CLR* test, a null distribution is required to decide about the statistical significance of the selective sweep hypothesis. The main advantage of the *SweepFinder* is that a specific population genetic model is not considered in the null hypothesis, but the SFS is derived from the whole-chromosomal pattern of variation; *i.e.*, from the data itself.

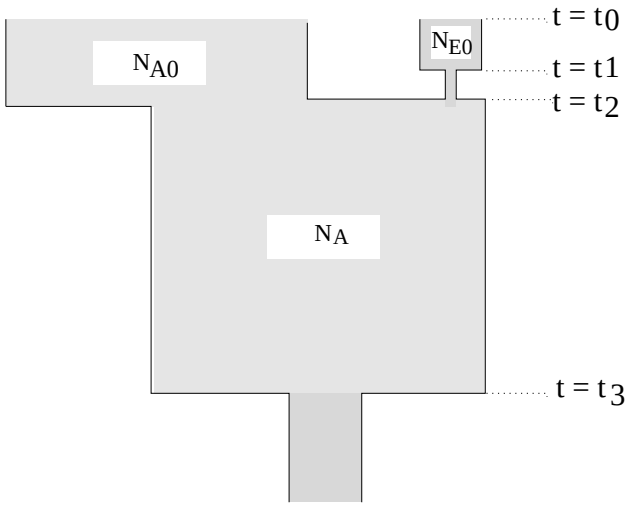

FIGURE 2.1: The demographic model of the European and African population of *D. melanogaster* as it was inferred by LI and STEPHAN (2006) and used in this study. The present European effective population size is approximately $N_{E0} = 10^6$ whereas the African population (N_{A0}) is 8 times larger. Backward in time the model can be described by a severe bottleneck in the European population that took place $t_1 = 15460$ years ago and lasted for ~ 340 years. During the bottleneck the effective population size of the European population was decreased to 2200. Approximately at $t_2 = 15800$ years ago the European population merges with the African population forming the ancestral population ($N_A = N_{A0}$). Finally, the ancestral population decreases to a fifth of the present day African population at $t_3 = 60000$ years ago.

We have extended the original approach for calculating the significance threshold for the *SweepFinder*. According to NIELSEN *et al.* (2005) the 95^{th} percentile of the statistic Λ_{SF} denotes the threshold value. Our approach, however, splits the region of interest into k fragments and for each one the $100 - \frac{5}{k}$ percentile is used as the cut-off value, resulting in a variable region-specific threshold. This approach helps to remedy the tendency of the *SweepFinder* to produce higher Λ_{SF} values at the borders of the region under study (P. P., unpublished results). Here we chose $k = 10$. The demographic model of Figure 2.1 (LI and STEPHAN 2006) with the ascertainment bias described in the 'Ascertainment bias correction' section is used to create the null distribution of the test statistics for all performed neutrality tests.

Estimation of the time since fixation of the beneficial allele: The time since the fixation of the beneficial allele was estimated by the methods described in PRZEWORSKI (2003) and SLATKIN and HUDSON (1991). For the Przeworski test, mutation rate 1.45×10^{-9}/bp/gen (LI and STEPHAN 2006) and recombination rate $r = 4.718 \times 10^{-8}$/bp/gen (COMERON *et al.* 1999) were used. The local parameters were estimated from a 925-bp long region located between the 7^{th} and 9^{th} exon of *HDAC6* [as exon 8 is very short (88bp), it has presumably no special effect on the parameter estimates, and was thus kept in the analysis]. This region contains 10 segregating sites forming 8 haplotypes, and Tajima's D = -1.74221. Two positions of the beneficial mutation were tested: one in the last exon of *HDAC6* and one in the last exon of *CG9123*.

We also used the Slatkin-Hudson method (SLATKIN and HUDSON 1991) assuming a star-like genealogy since the fixation of the beneficial allele. We based this estimation on the DNA region between positions 9.865 and 12.443 kb. In this region 19 segregating sites were detected and divergence to *D. simulans* is 0.056. To convert the obtained estimates into years, we assumed 10 generations per year for both methods.

2.4 Results

Standard analyses of a candidate region of selection: The region analyzed here is about 70 kb long. It is located in a highly recombining portion of the X chromosome ($r = 4.718^{-8}$/bp/gen) and is relatively gene dense. This region contains 12 genes, five of which have unknown molecular functions (*CG15032*, *CG9114*, *CG9123*, *CG12608*, and *CG9164*). The other genes have been functionally characterized (*gce*, *Top1*, *dah*, *HDAC6*, *CG6227*, *acj6*, and *Pp1*). In order to perform a fine-scale analysis of the African sample, we sequenced 15 non-coding (intronic or intergenic) DNA fragments of 511 bp on average, in addition to the four already sequenced by OMETTO *et al.* (2005) (Figure 2.2). For each of these 19 fragments, basic summary statistics were calculated,

averaged over the whole candidate region, and then compared to the chromosomal average. Only 15 of the 19 fragments could be aligned with *D. simulans*.

The region exhibits a strong reduction in nucleotide polymorphism. On average the 259 fragments sequenced by OMETTO *et al.* (2005) for the African population contained twice as many segregating sites as the 70-kb candidate region ($p < 0.0001$). π and θ_W were significantly lower than the chromosomal average ($p < 0.0001$ for both). As can be seen in Figure 2.2, the θ_W curve is roughly U shaped (with a minimum between 10 and 15 kb), except for two positions at -10 and around 40 kb where divergence is very low. In general, divergence is rather high in the region of reduced variation between positions 0-22 kb (~ 0.09).

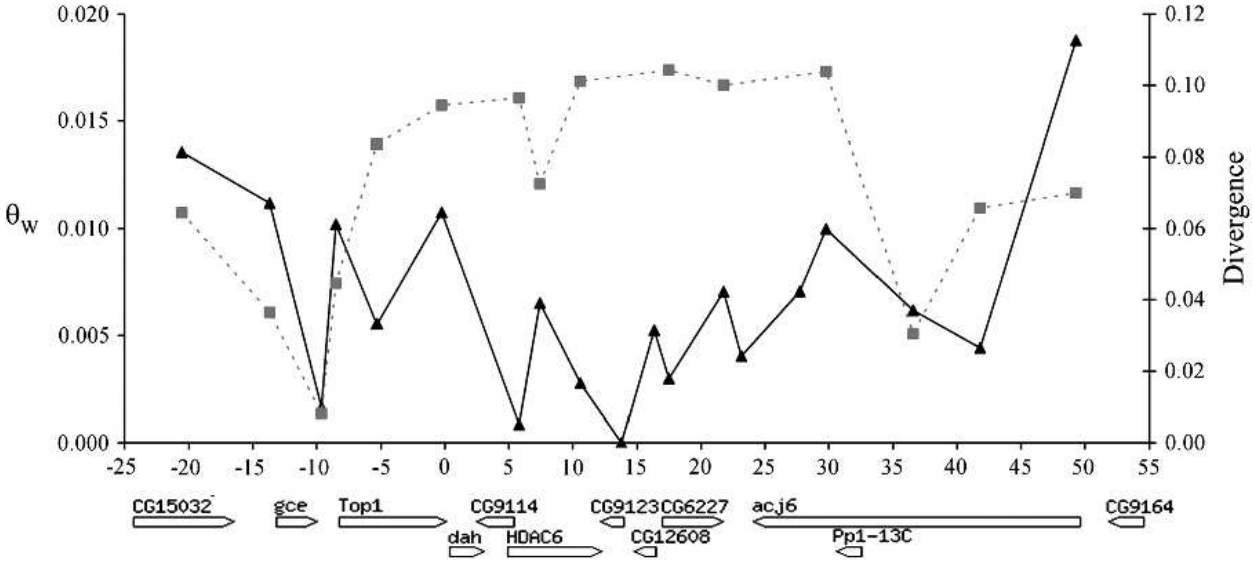

FIGURE 2.2: Nucleotide diversity θ_W (solid line) and divergence to *D. simulans* (dashed line) across the candidate region for selection. The relative positions in kb are on the X-axis. Gene spans (according to Flybase) are at the bottom of the graph.

Furthermore, the region shows deviations from the chromosomal expectation with regard to the SFS. Indeed, Tajima's *D* values are more negative than the X chromosome average (-1.143 vs. -0.667), which is highly significant ($p = 0.001$). Four fragments show significantly negative Tajima's *D* values (data not shown). In contrast, Fay and Wu's *H* statistic does not depart from the chromosomal average. This illustrates that the SFS is lacking intermediate frequency variants and shows an excess of low frequency SNPs.

The number of haplotypes ranges from 1 to 12 in the candidate region, but its mean is significantly lower than the chromosomal average ($p < 0.001$). Similarly, haplotype diversity is significantly lower ($p < 0.001$). LD as measured by the Z_{nS} statistic is relatively constant over the whole region (< 0.3) and does not deviate from the chromosomal average.

The genes *CG9123* and *CG12608* are paralogs. Among the 12 *Drosophila* genomes examined (*Drosophila* 12 GENOMES CONSORTIUM 2007) this duplication is present only in *D. melanogaster*. Both copies are highly diverged from *D. simulans*. Investigating the pattern of polymorphism at both genes, we did not find evidence for extensive gene conversion; for instance, there is only one SNP shared between both copies (out of 48 SNPs in total). *CG9123* contains many non-synonymous SNPs in relatively high frequency, most of which produce drastic amino acid changes. In addition, we observed some deletions in the coding region, one of which causes a frame shift change. This may suggest that *CG9123* is under weak functional constraints or even a pseudogene.

Application of the *CLR* and *SweepFinder* tests: In order to perform more advanced neutrality tests, we defined a region of about 22 kb (corresponding to the segment between absolute positions 15222319 and 15244496 in Flybase release 5.10, and to positions 0 to 22 kb in Figure 2.2). This region was then completely sequenced and subjected to the *CLR* and *SweepFinder* tests. The *CLR* test was marginally significant (p = 0.048) when the null distribution of the statistic Λ_{CLR} was constructed from the demographic scenario of the African population inferred by LI and STEPHAN (2006) (Figure 2.1). Figure 2.3A shows Λ_{CLR} along the region. The beneficial mutation is estimated to have occurred at position 11.378 kb relative to the beginning of the 22-kb region, and $\alpha = 2Ns$ is approximately 13076 (where N is the effective population size and s the selection coefficient). This value is much higher than most other reported estimates, which is consistent with the observed width of the valley of reduced variation and the fact that population recombination rate $4Nr$ is very high in this part of the genome.

The *SweepFinder* test was also significant (p = 0.034) for the 22-kb completely sequenced region. In Figure 2.3B we show the Λ_{SF} values along the region. Consistent with the result of the *CLR* test, three positions (11.315, 12.474 and 13.110 kb) show the highest Λ_{SF} values. The high value around position 1.0 kb is probably not a target of selection as it is not confirmed by Tajima's *D* and the *CLR* test.

Age of the selective sweep: The age of the sweep in the 22-kb region was estimated by the Przeworski and Slatkin-Hudson methods (cf. Material and Methods). We used Przeworski's approach with two positions as input parameter values that are near the estimated selected sites: position 11.787 kb gave a time since fixation of the beneficial allele of 63,334 years (95% C.I.: 23,382-628,432 years), while position 12.787 kb gave 56,770 years (95% C.I.: 21,121-577,307 years). Using the Slatkin-Hudson method the age of the sweep was estimated as 50,047 years.

These estimates suggest that the sweep occurred before the European lineage split off from the African one (about 16,000 years ago; LI and STEPHAN (2006)). In order to confirm this hypothesis, we re-sequenced the region between position 8.0 and 15.0 kb in 12 lines of a European sample from

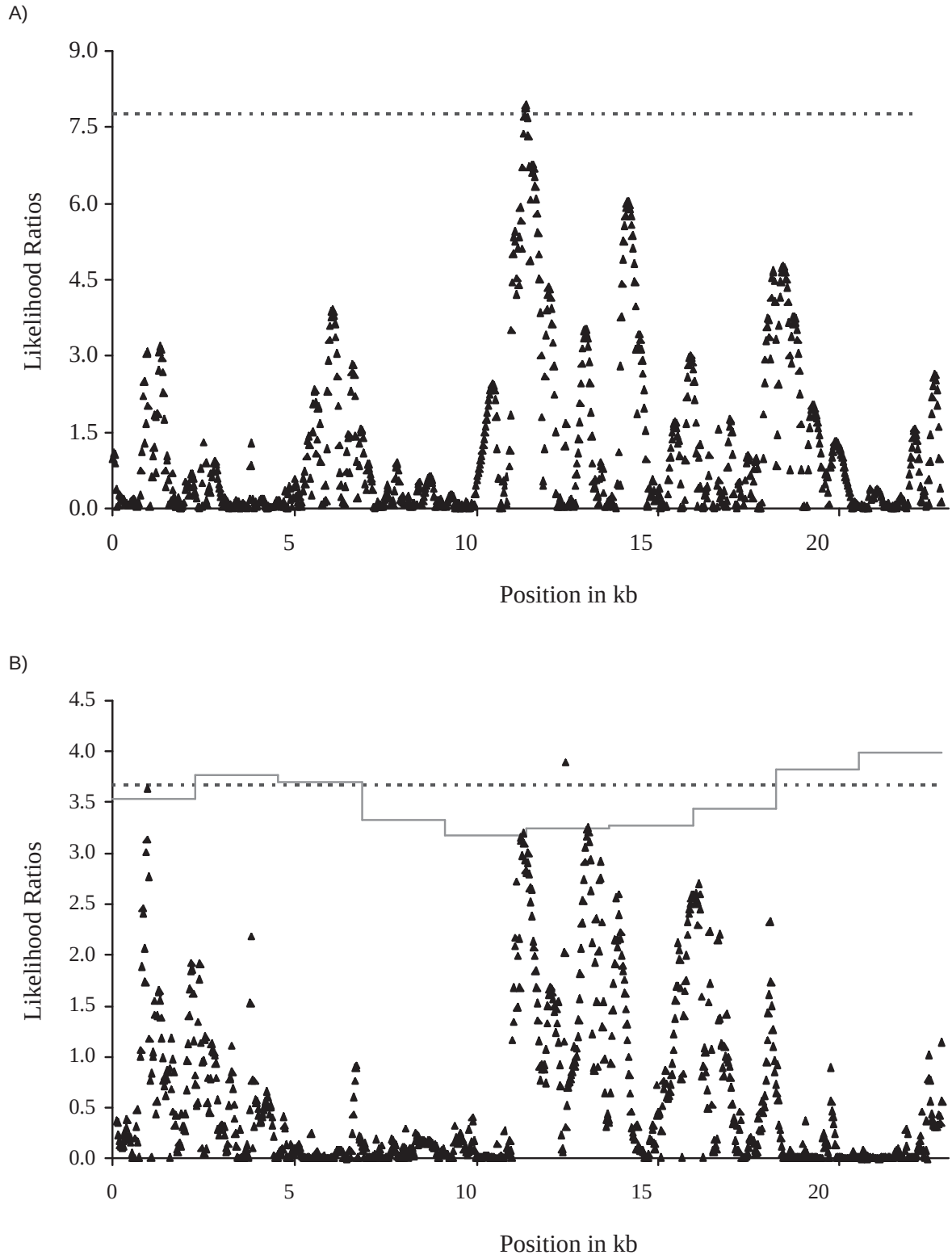

FIGURE 2.3: The likelihood-ratio values calculated by A) the *CLR* and B) the *SweepFinder* tests for a 22-kb subregion of the 70-kb region (for 1000 bins). Each triangle denotes the value of the test statistics for a selective sweep model for which the beneficial mutation occurred at that specific position. In B) the dashed line depicts the constant threshold calculated according to NIELSEN *et al.* (2005), whereas the solid line shows the variable threshold (see Materials and Methods).

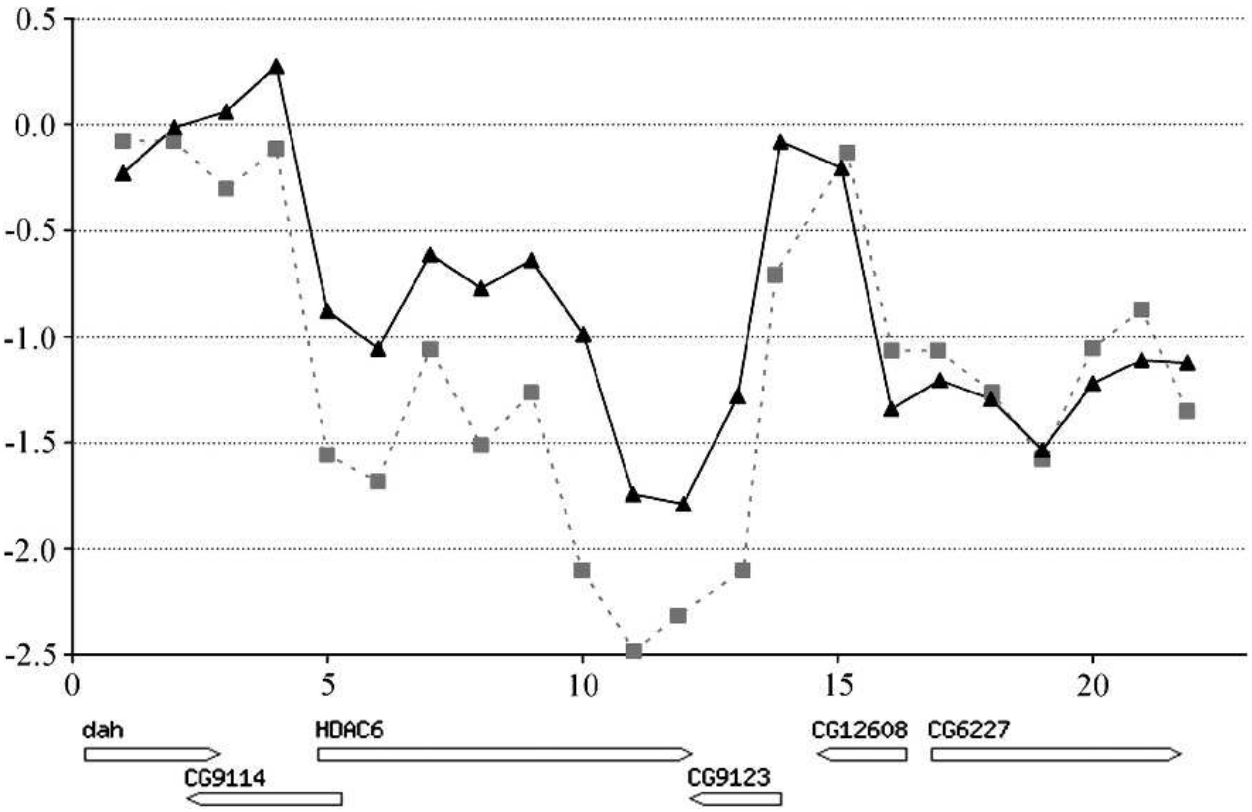

FIGURE 2.4: Sliding window analysis of the fully sequenced 22-kb region. Tajima's *D* and Fu and Li's *D* are represented by solid lines (black triangles) and dashed lines (grey squares), respectively. Each data point represents the midpoint of a 2000-bp long window and the step size is 1000 bp. In order to describe the neutral frequency spectrum we excluded the non-synonymous sites from this analysis.

The Netherlands (Materials and Methods). We found that the European lines were identical with those of the African sample in a limited segment of approximately 2.7 kb from position 9.8 to 12.5 kb (except for three derived singletons and one doubleton. This suggests, in conjunction with the estimated age of the sweep, that the selected allele has been exported to Europe during the colonization process.

Sliding window analysis: To corroborate our mapping results, we also performed a sliding window analysis on the SFS of the sequenced 22-kb region (Figure 2.4). Indeed, significantly negative Tajima's *D* and Fu and Li's *D* values were found near the estimated targets of selection, consistent with the *CLR* and *SweepFinder* results. The analysis revealed a small segment showing a local reduction of nucleotide diversity and a SFS shifted toward low-frequency variants despite normal levels of divergence. This region coincided with the 2.7-kb fragment mentioned above. Four exons lie in this region: the three last exons of *HDAC6* and a portion of the last exon of CG9123. The relatively low value of Tajima's *D* around position 19 kb is probably due to purifying selection (causing the observed low divergence in the helicase functional domain of *CG6227*; data not shown).

In order to identify candidate substitutions under selection, we aligned the 2.7-kb region of *D. melanogaster* to that of *D. sechellia*, *D. simulans*, *D. erecta*, and *D. yakuba*. As the 2.7-kb region centers on *HDAC6*, we focused our investigations on this gene. The *HDAC6* introns were poorly conserved between species but we obtained a good alignment of the 3' UTR and of the three last exons of the gene. In the 3' UTR, we found 6 nucleotide substitutions specific to the *D. melanogaster* lineage. In exon 7, we identified three non-synonymous substitutions specific to *D. melanogaster*. All of them cause non-polar to non-polar amino-acid replacements. We also found a deletion of 9 nucleotides that is specific to *D. melanogaster* at the end of exon 9. This exon also carries two non-synonymous substitutions. One of them generates a drastic amino-acid change: a valine to glutamic acid substitution. In addition, this substitution is in a region predicted by the program MyHits (`http://myhits.isb-sib.ch`) to be the ubiquitin-binding site of HDAC6.

2.5 Discussion

Evidence for a selective sweep in the *HDAC6* region of African *D. melanogaster*: By completely re-sequencing a 22-kb region around *HDAC6* in a sample of 12 African *D. melanogaster* X chromosomes and applying two likelihood tests (*CLR* and *SweepFinder*), we found evidence consistent with the presence of a selective sweep in this region. Furthermore, our mapping showed that the target of selection is most likely located in a 2.7-kb DNA region, centering on the last exon

of *HDAC6*.

The expected age of the sweep was estimated as 50,000 to 63,000 years, depending on the method and input parameter values. This suggests that the sweep occurred before the European lineage split off from the African one (which occurred about 16,000 years ago; LI and STEPHAN (2006). Our age estimates are consistent with the observation that the sequences of the *HDAC6* alleles from our European sample are identical with that of the African haplotype in the swept region of approximately 2.7 kb (except for some derived low-frequency variants). Consistent with the relatively old age of the selective sweep, we did not identify any pattern of LD that is characteristic of a sweep (according to PFAFFELHUBER *et al.* (2008)). Interestingly, a PAML analysis (YANG 2007) of *HDAC6* sequences from five species of the *D. melanogaster* subgroup found no evidence of selection (data not shown). This suggests that, prior to the inferred selective sweep, *HDAC6* has not undergone accelerated evolution in the past few million years.

It is clear that the evidence we provided is subjected to some uncertainty. First, the results inferred by the *CLR* and *SweepFinder* tests may depend to some degree on demography. In particular, complex demographies could be a confounding factor (for instance, population size bottlenecks; PAVLIDIS *et al.* (2008)). However, the demographic history of the African population we inferred previously is probably relatively simple and may be summarized by an expansion model (LI and STEPHAN 2006; HUTTER *et al.* 2007). Furthermore, we have improved the original *CLR* test by KIM and STEPHAN (2002) and have now taken demography into account. Finally, the problem of demography is alleviated by applying *SweepFinder*, because the chromosome-wide background SFS is used rather than a specific model. Second, a more general concern may be that if selection is a frequent and major pervasive force our two-step approach for inferring selection may not work (HAHN 2008). Then a joint inference of selective and demographic parameters would be a more appropriate approach. However, we emphasize that we search for very strong selection. In such a case, our method of separating demography from selection is expected to be a reasonable first approximation. Third, the uncertainty in the estimates of the target site of selection needs to be mentioned. Unlike BEISSWANGER and STEPHAN (2008), we were not able to obtain confidence intervals of our estimates, as the rate of recombination in the *HDAC6* region is too large. However, based on the site frequency spectrum we were able to support our conclusion that the most likely target of selection is located in a 2.7-kb region (between positions 9.8 and 12.5 kb; see Figure 2.4). This result is consistent with the observation that the European alleles are identical in this region with the selected African allele. This latter argument, however, requires that the sweep occurred in Africa before the African and European lineages split, which is indeed supported by the estimated lower bound of the age of the sweep of > 20000 years.

Can the polymorphism patterns in the *HDAC6* region be explained by selective pressures other than positive directional selection? It is possible that at least part of the polymorphism pattern is associated with the action of purifying selection. The entire 70-kb region contains several functional elements that give rise to low divergence levels (Figure 2.4). In the identified 2.7-kb region between positions 9.8 and 12.5 kb, however, divergence is everywhere in the range of 5-8% and thus comparable to the average of the whole 70-kb region of 6.8%. This suggests that purifying selection is not likely a major cause of the observed pattern of variation in the 2.7-kb region.

Significance of the selective sweep in relation to the function of *HDAC6*: The 2.7-kb region we mapped by the selection approach overlaps with the last exons of two genes, *HDAC6* and *CG9123*. The latter is a duplicate of *CG12608*. According to the alignment of the 12 fully sequenced *Drosophila* species (*Drosophila* 12 GENOMES CONSORTIUM 2007), this duplication event occurred in the *D. melanogaster* lineage. However, based on the polymorphism pattern mentioned above, *CG9123* is probably a pseudogene (or on its way to becoming one). Furthermore, *CG9123* is located at the boundary of the identified 2.7-kb region. We therefore concentrate the following discussion on *HDAC6*.

HDAC6 is a unique member of the histone deacetylase family harboring a ubiquitin-binding site and two catalytic deacetylase domains (VERDEL *et al.* 2000; KHOCHBIN *et al.* 2001). In addition, its localization in the cytoplasm is very unusual for an histone deacetylase (VERDEL *et al.* 2000). It has been shown that its role is not limited to gene regulation. Rather, it is also important for the general cytotoxic stress response. It is involved in the two major cellular mechanisms degrading misfolded protein aggregates: autophagy and the ubiquitin-proteasome system (PANDEY *et al.* 2007). HDAC6 detects and mediates the cytotoxic stress response at three different levels. First, its strong ubiquitin-binding ability coupled with its ability to move along microtubules allows HDAC6 to transport ubiquitinated protein aggregates, thus favoring the formation of aggresomes. Second, HDAC6 is able to stimulate autophagy when the ubiquitin-proteasome system is impaired (PANDEY *et al.* 2007), and finally it mediates the activation of heat shock proteins (BOYAULT *et al.* 2006). More generally, HDAC6 is believed to be involved in several other cell stress response pathways such as antiviral responses (BOYAULT *et al.* 2006). In *D. melanogaster*, *HDAC6* is mainly expressed in an insect specific organ: the Malpighian tubule (CHINTAPALLI *et al.* 2007). Its tissues might be exposed to a broad range of cellular stress as it carries out most of the osmoregulation and the excretion of organic solutes as well as xenobiotics (DOW and DAVIES 2006).

To identify possible targets of selection, we aligned the *HDAC6* sequence of five *Drosophila* species. It revealed that *HDAC6* carries a limited number of *D. melanogaster*-specific changes. But we could neither confirm nor exclude that any of them is a positively selected substitution.

Indeed, any nucleotide change in the introns or 3' UTR could affect *HDAC6*'s regulation or expression and any of the non-synonymous changes observed in the exons could modify the protein's properties. However, in the last exon of *HDAC6* one non-synonymous substitution may well have significant functional consequences: a valine-to-glutamic acid replacement that occurred in the *D. melanogaster* lineage and is located in the ubiquitin-binding site of *HDAC6*. Could this substitution affect the ubiquitin-binding affinity of HDAC6 and thus the response of cells to stress? Ubiquitin-binding assays (BOYAULT *et al.* 2006) comparing the *D. melanogaster* and *D. simulans* alleles may provide an answer to this question.

Chapter 3

Selective sweeps in multi-locus models

Unpublished work

3.1 Abstract

We study the trajectories of a new selected mutation that affects a quantitative trait which is determined by multiple loci. Then, given the trajectory, we analyze the properties of the coalescent trees around the new mutation and the neutral polymorphism patterns, and compare them with those of classical selective sweeps and those under neutrality. Trajectories are generated with forward-in-time simulations. Coalescent trees and neutral polymorphism patterns have been implemented conditioning on the trajectory. The fitness function of the trait is Gaussian. The model assumes that the population size is finite; the recombination rate between two adjacent loci is arbitrary. A major objective of the article is to scrutinize the similarities and differences between the multi-locus model affecting a quantitative trait and the classical one-locus selective sweep model, and consequently to study whether the statistical tests that have been developed to detect one-locus selective sweeps are useful for the multi-locus scenario. In the case of multi-locus scenarios, the trajectories of a new mutation, even beneficial, might not reach fixation. The alleles of the loci compete against each other and a polymorphic equilibrium may be obtained. In general, the trajectories that reach polymorphic equilibria generate different polymorphism patterns than the trajectories that result in fixation. If the polymorphic equilibrium point has been reached recently, then the coalescent trees and the polymorphism patterns resemble the coalescent trees and the polymorphism patterns of standard neutral model. Consequently, current neutrality tests would not be able to detect a large proportion of selective events in multi-locus models. On the other hand, if fixation is achieved then the polymorphism patterns are similar to the classical one-locus selective sweeps

and neutrality tests can detect the selective event.

3.2 Introduction

According to the classical one-locus selective sweep theory (MAYNARD SMITH and HAIGH 1974), three distinct signatures of selection may emerge after the fixation of a beneficial mutation. First, the level of polymorphism is reduced in the neighborhood of the beneficial mutation; second, the site frequency spectrum shifts towards low- and high-frequency derived variants, and third, linkage disequilibrium is high on each side of the beneficial mutation but low across the selected site. For ongoing sweeps the main signature consists of extended haplotypes in high frequency (VOIGHT *et al.* 2006). In the last decade, a multitude of tests have been developed that aim at detecting those patterns in whole genome scans (KIM and STEPHAN 2002; KIM and NIELSEN 2004; NIELSEN *et al.* 2005; JENSEN *et al.* 2007a; PAVLIDIS *et al.* 2010). The next step after detecting genomic regions that show signatures of selection attempts to associate the genes that are located in the region with an (advantageous) phenotype (SVETEC *et al.* 2009).

On the other hand, a phenotype may be determined by a multitude of genes as well as the environment. Multi-locus population genetics has been developed in the last decades to describe the evolution of multi-locus systems and phenotypes. Selective forces, such as directional, stabilizing, or disruptive selection modify the genetic constitution of the population and drive the population to either extreme or optimal genotypic values. In this study we focus on stabilizing selection, *i.e.* the type of selection toward a phenotypic optimum. However, here this optimum might not coincide with the genotypic value of heterozygotes. Historically, of special interest is the maintenance of genetic variability under stabilizing selection, because stabilizing selection is assumed to control traits in various organisms, for example the color coat in mice (VIGNIERI *et al.* 2010), human facial features (PERRETT *et al.* 1994), plant defense mechanisms (MAURICIO and RAUSHER 1997), enhancer elements in *Drosophila* (LUDWIG *et al.* 2000), and vocalization in frogs and toads (GERHARDT 1994); see also ENDLER (1986, chapter V) for examples and discussion. Furthermore, it has been suggested that this type of selection exhausts genetic variation (ROBERTSON 1956; FISHER 1930). Studies that underpin this view are based on a large number of loci of very small effect on the trait and they are supported by approximations that are focused on one arbitrary locus (*e.g.* ROBERTSON 1956). By contrast, many quantitative traits exhibit high levels of genetic variability. This contradiction motivated researchers to study the role of mutation (LANDE 1975; TURELLI 1984; GAVRILETS and HASTINGS 1994; BÜRGER 1998), overdominance (BULMER 1973; GILLESPIE 1984), migration (TUFTO 2000), frequency-dependent selection through

intraspecific competition for some resource (BÜRGER 2002; BÜRGER and GIMELFARB 2004), genotype-environment interaction (GILLESPIE and TURELLI 1989), or pleiotropy and epistasis. Additionally, a lot of work has been put into exploring the ability of stabilizing selection *per se* to maintain genetic variability of quantitative traits that are controlled by multiple loci. Theoretical focus was mainly on two-locus models, but also models of more than two loci have been analyzed. Surprisingly, predictions about genetic variability depend profoundly on the number of loci. The two-locus model predicts that genetic variability remains in the population due to stabilizing selection *per se*. On the other hand, in models with more than two loci the vast amount of genetic variability diminishes. The reason is that the optimum can be reached very closely by various homozygous genotypes (BÜRGER 2000, chapter VI) when there are more than two loci that control the trait. For the two-locus model and assuming a symmetric viability model (*e.g.* BODMER and FELSENSTEIN 1967; KARLIN and FELDMAN 1970), it has been shown that there are nine equilibria (BÜRGER 2000), seven of which can be stable but not simultaneously. Those seven equilibria split into four classes (BÜRGER and GIMELFARB 1999): they can be either polymorphic for both loci, one of them, or totally monomorphic. The equilibrium points of the two-locus, two-allele model can be depicted on a tetrahedron (KARLIN and FELDMAN 1970; BÜRGER 2000, page 23; see also Figure 3.1). The vertices correspond to the fixation of the labeled gamete, and frequencies are measured by the orthogonal distance from the opposite boundary face. Consequently, if an equilibrium point is located within the tetrahedron (internal equilibrium), then it is polymorphic for both of the loci (and for all four gametes), since the distance from each face is positive. On the other hand, an equilibrium point on one of the edges or vertices is monomorphic for at least one allele. The equilibria points on the vertices are monomorphic for the alleles in both of the loci; the equilibria points on the edges maintain two gametes, but they are monomorphic for one of the alleles. Of special interest in the present study are the equilibria that correspond to the fixation of one of the alleles for at least one locus, namely the equilibria on the edges and vertices. Throughout the article A_i denotes the i^{th} locus and A_{ij} the j^{th} allele of the i^{th} locus. Figure 3.1 illustrates a tetrahedron with the gametes $A_{11}A_{21}$, $A_{11}A_{22}$, $A_{12}A_{21}$, and $A_{12}A_{22}$ on the vertices. An equilibrium point K on the edge ($A_{11}A_{22}$, $A_{11}A_{21}$) corresponds to i) absence of gametes $A_{12}A_{21}$, and $A_{12}A_{22}$, and ii) frequencies of gametes $A_{11}A_{21}$, $A_{11}A_{21}$ equal to the distances KKa and KKb, respectively. WILLENSDORFER and BÜRGER (2003) fully explore the equilibrium properties of the two-locus, two-allele model of Gaussian selection under the assumption of a symmetric fitness function with respect to the double heteroygote. The analysis of WILLENSDORFER and BÜRGER (2003) is important because it provides the existence and stability criteria for the equilibrium points of the model. Let A_1 and A_2 denote the two loci with alleles A_{11}, A_{12} and A_{21}, A_{22}, respectively. The

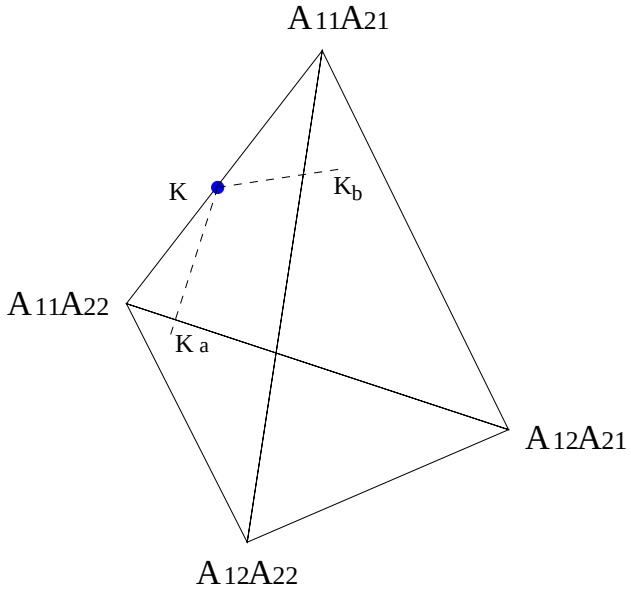

FIGURE 3.1: The tetrahedron that represents the state space for the two-locus two-allele model. Vertices correspond to fixation of the labeled gamete; an edge corresponds to the absence of the gametes that are not located on the edge; a face corresponds to the absence of the gamete on the opposite vertex, and an internal point corresponds to a polymorphic state for all the gametes

fitnesses of the nine possible genotypes are shown in Table 3.1. The genotypes $A_{11}A_{21}/A_{12}A_{22}$ and $A_{11}A_{22}/A_{12}A_{21}$ are equivalent. Let x_1, x_2, x_3, and x_4 represent the frequencies of the gametes $A_{11}A_{21}$, $A_{11}A_{22}$, $A_{12}A_{21}$, and $A_{12}A_{22}$, respectively. Furthermore, $W_i,\ i = 1,2,3,4$ is the marginal fitness of the gametes. Then, a classical result (*e.g.* KARLIN and FELDMAN 1970; WILLENSDORFER and BÜRGER 2003) gives the recursion relations for the frequencies in the next generation as:

$$\bar{W}x_i' = x_i W_i - \eta_i r D, \quad i = 1, 2, 3, 4, \tag{3.1}$$

where $\eta_1 = \eta_4 = 1$ and $\eta_2 = \eta_3 = -1$. D denotes the linkage disequilibrium and is defined as $D = x_1x_4 - x_2x_3$. The average fitness is $\bar{W} = 1 - d(x_1^2 + x_4^2) - a(x_2^2 + x_3^2) - 2b(x_1x_2 + x_2x_4) - 2c(x_1x_3 + x_2x_4)$. The system of Equation 3.1 cannot be solved explicitly (Reinhard Bürger, pers. communication). Notice that the model is deterministic, *i.e.* the stochastic effect of genetic drift is negligible.

WILLENSDORFER and BÜRGER (2003) parametrize the model so that the effect of gametes $A_{11}A_{21}$, $A_{11}A_{22}$, $A_{12}A_{21}$, and $A_{12}A_{21}$ are $-\gamma_1/2$, $\gamma_1/2$, $-\gamma_2/2$, and $\gamma_2/2$, $\gamma_1 \geq \gamma_2 \geq 0$; then, the optimal phenotype is 0 for the double heterozygote. Let a_i denote the fitness for the γ_1 phenotype, *i.e.* $\alpha_1 = \exp(-\gamma_1/\omega^2)$ under the Gaussian selection function. ω^2 quantifies the strength of selec-

Table 3.1: (A) Genotypic values and (B) fitnesses for the symmetric fitness model

A

	$A_{21}A_{21}$	$A_{21}A_{22}$	$A_{22}A_{22}$
$A_{11}A_{11}$	$-\gamma_1-\gamma_2$	$-\gamma_1$	$-\gamma_1+\gamma_2$
$A_{11}A_{12}$	$-\gamma_2$	0	γ_2
$A_{12}A_{12}$	$\gamma_1-\gamma_2$	γ_2	$\gamma_1+\gamma_2$

B

	$A_{21}A_{21}$	$A_{21}A_{22}$	$A_{22}A_{22}$
$A_{11}A_{11}$	$1-d$	$1-b$	$1-a$
$A_{11}A_{12}$	$1-c$	1	$1-c$
$A_{12}A_{12}$	$1-a$	$1-b$	$1-d$

tion (s in WILLENSDORFER and BÜRGER (2003)). Furthermore, let $\alpha_{12} = \exp(-2\gamma_1\gamma_2/\omega^2) = \exp(-2\sqrt{\ln\alpha_1 \ln\alpha_2})$. Then, the terms a,b,c,d in Table 3.1B can be represented as $a = 1-\alpha_1\alpha_2/\alpha_{12}$, $b = 1-\alpha_1$, $c = 1-\alpha_2$, $d = 1-\alpha_1\alpha_2\alpha_{12}$. Thus, α_1 and α_2 are sufficient to describe the fitness matrix. In the Results section, we illustrate that α_1, α_2, and the recombination rate r determine the equilibrium properties of the model. Explicit formulas are given by WILLENSDORFER and BÜRGER (2003) by linearization of Equation 3.1 at the equilibrium points (Equations 3.1, 3.2a, 3.2b, 3.8 in WILLENSDORFER and BÜRGER (2003)). Furthermore, we demonstrate that initial frequencies for the A_1 and A_2 locus determine to a large extent whether a new mutation in locus A_1 will be fixed. The symmetrical fitness model can be extended to an l-locus two-allele model. A l-dimension matrix is required to describe the genotypic values and the fitnesses of the genotypes. Then, similarly to the two-locus two-allele model, we assume that the optimal genotype is the l-tuple heterozygote, the optimum is at 0, and that symmetrical genotypes to the l-tuple heterozygote have symmetric genotypic values about the optimum. Then, it can be shown that for each locus the sum of the contributions of the two alleles is 0.

Even if the equilibrium properties of multi-locus models are not in the focus of the present study, they are relevant. They show that a selective sweep is not always achievable when a quantitative trait that is controlled by multiple loci is under Gaussian selection. As mentioned above, theoretical studies have shown that stable polymorphic equilibrium points are possible when the trait is controlled by few loci (< 4; BÜRGER 2000, chapter VI). Therefore, it is possible that a new mutation even if beneficial initially, will not fix in the population but will remain polymorphic as long as the optimum remains constant.

To our knowledge, the first effort that bridges quantitative trait evolution and selective sweeps was made by CHEVIN and HOSPITAL (2008). Their work was based on the seminal paper by LANDE (1983). Lande's model focuses on one locus of major effect on the trait, and treats the remaining loci of minor effects as genetic background for this locus. It is assumed that heritable background variation is maintained in a constant amount by polygenic mutation and recombination (LANDE 1975, 1983); also, the various loci that affect the trait are unlinked and there are

no epistatic interactions. In this model, the joint evolution of the allelic frequencies on the focal locus and the phenotype is determined by two difference equations, namely $\Delta q = \frac{q(1-q)\partial \bar{W}}{2\bar{W}\partial q}$ and $\Delta \bar{z} = \frac{h^2\sigma^2\partial \bar{W}}{\bar{W}\partial \bar{z}}$, where q denotes the frequency of the allele on the focal locus, $\bar{W}$ denotes the average fitness of the population, $\bar{z}$ the average genetic background value, and $h^2\sigma^2$ the amount of heritable genetic variability. Analyzing the system of these two equations, Lande inferred stable and unstable equilibrium points under various dominance schemes and fitness functions. CHEVIN and HOSPITAL (2008) used Lande's model to infer the deterministic trajectory of a beneficial mutation that affects a quantitative trait in the presence of background genetic variability. They studied both directional and stabilizing selection and showed that fixation needs longer time in their quantitative trait setup than in the classical one-locus model (*i.e.* when genetic variability in the background is absent). In contrast to CHEVIN and HOSPITAL (2008) the present simulation-based study assumes an explicit number of loci that determine the trait as this was done by BODMER and FELSENSTEIN (1967), KARLIN and FELDMAN (1970), BÜRGER (2000, chapter VI). Therefore, the assumption of constant variability in the genetic background is relaxed since the genetic background is modeled explicitly.

Of special interest in studies of multi-locus models is the role of linkage disequilibrium and the strength of selection. Previous work has shown that the number and the stability of the equilibria depend on the recombination rate between the loci and the strength of selection. These studies have been focused on two-locus systems and usually assume a large population size, so that genetic drift is negligible, and the frequencies of the gametes evolve deterministically. In the present study we analyze both the deterministic and the stochastic evolution of the multi-locus model by assuming a finite constant effective population size.

3.3 Methods

3.3.1 The model

We consider a diploid population of size N, and a quantitative trait under selection. The quantitative trait is controlled by l-diallelic loci with no epistatic interactions on the phenotype. The alleles for each locus are codominant. The alleles at the i^{th} locus are labeled as A_{i1} and A_{i2}. Allele A_{i1} contributes w_{i1} to the trait, and the contribution of A_{i2} equals to w_{i2} for each i. Without loss of generality the optimum for the trait is set to 0. The recombination fraction between alleles i and $i+1$, is $r_i \leq 0.5$. At time $t = t_0$, the number of A_{i1} alleles follows a binomial distribution with parameter $p_0(A_{i1})$ and N, and the loci are in linkage equilibrium ($D = 0$). The trait is assumed to be

under a Gaussian fitness function, *i.e.* if the phenotypic value of an individual is P, then its fitness is given by $W(P) = \exp(-P^2/\omega^2)$, and ω^2 determines how fast the fitness decreases away from the optimum. Here, the phenotypic value P is determined explicitly by the genotype. However, it is straitforward to include environmental noise, assuming that the environmental component is normally distributed with mean 0 and variance σ_E^2 (see Implementation). Individuals are considered to be hermaphrodites, *i.e.* an individual can represent both a male and a female; mating is random.

The population evolves forward in time from $t = t_0$ to the present $t = 0$ and generations do not overlap. For each allele A_{ij} an initial frequency $p_0(A_{ij})$ is assumed, and the mutation rate is 0 (see Implementation for extensions of the model). In each generation, the life cycle consists of (i) the zygote phase, (ii) viability selection, where individuals are selected as parents for the next generation according to their fitness value, (iii) recombination for each of the parents where gametes are formed, and (iv) random mating to form the zygotes of the next generation. In step (iv), N matings take place among N individuals. Each mating produces one diploid offspring and each individual can participate in multiple matings as a male or female. In each generation, at the zygote phase, the frequencies of the alleles of the locus of interest are recorded and the trajectories are stored. Notice that in this model selection and drift act simultaneously in step (ii), where a finite number of individuals is chosen as parents in the next generation. Also, random genetic drift acts in steps (iii) and (iv): from a pair of gametes only one recombinant is chosen to pass to the next generation.

The next step proceeds backward in time. Assume a sample of k individuals from the present-day population ($t = 0$). Given the trajectory of the A_{11} allele from the previous step, we implement coalescent simulations from $t = 0$ to the TMRCA of the neutral genomic region around the locus A_1. The backward in time simulations are based on the structured coalescent model (WAKELEY 2008; TESHIMA and INNAN 2009; EWING and HERMISSON 2010). That means that the population is subdivided into two genetic backgrounds: one class of lineages is linked to the A_{11} allele and the other is linked to the A_{12}. Given the trajectory of the A_{11} and A_{12} allele, the genealogical history of linked neutral regions is considered separately for the two classes, while recombination allows lineages to move between the two classes (as migration allows lineages to move in a structured population). With the backward process the genealogies of the genomic region around the locus A_1 are obtained. We assume that the genealogies of the genomic region around the locus A_1 are affected only by the locus A_1 and not by the remaining of loci. This simplification makes the backward simulations tractable, and allows us to use available simulation software (*e.g.* TESHIMA and INNAN 2009; EWING and HERMISSON 2010). However, as it is mentioned in following sections, this is correct only when selection is weak and the loci unlinked.

3.3.2 Summary statistics of the coalescent and SNP polymorphisms

Next, the genealogies are summarized. Summarizing the genealogies facilitates the inspection of their properties, and more importantly, the comparison to one-locus selective sweeps or to neutrality. Four summary statistics have been used. First, h the height of the coalescent tree which measures the scaled time from the present to the MRCA of the sample. Second, L the total length of the coalescent is calculated by summing up the lengths of all branches, and it is described by scaled time units as well. Third, we developed two summary statistics, b_L and b_N, which measure the balance of the coalescent when the root is placed at the node of the MRCA. b_L is based on the length of the subtrees on the right and on the left side of the MRCA; on the other hand, b_N uses the number of nodes on the right and on the left side (Equation 3.2).

$$b_L = \frac{4l_L l_R}{l^2}, \quad b_N = \frac{4n_L n_R}{n^2}. \tag{3.2}$$

l_L and l_R denote the total length of the left and right subtree of the MRCA, respectively, and l the total length of the coalescent. n_L and n_R is the number of nodes on the left and on the right side of the MRCA, respectively, and n is the total number of nodes (excluding the root), *i.e.* $n = 2k-2$, where k is the sample size. b_L and b_N take values in (0, 1]; when they equal 1 the coalescent trees are balanced perfectly, whereas smaller values denote some imbalance. The summaries of the genealogies are related to the perturbations of the coalescent due to the action of selection. It is well known that in the neighborhood of a beneficial mutation, directional selection reduces the height and the length of the coalescent, and increases its imbalance.

Furthermore, we used population genetics SNP summary statistics to describe the polymorphism patterns in a present-day sample, as we move along the sequence alignment away from the A_1 locus. Thus, we measure the level of polymorphism using the number of polymorphic sites. Tajima's D is used to summarize the site frequency spectrum. Additionally, we implemented the Depaulis and Veuille statistics (DEPAULIS and VEUILLE 1998), which calculate the number of haplotypes (K) and their divergence (H). Those summary statistics facilitate the comparison between polymorphism patterns that are created by the multi-locus model and the one-locus selective sweep. Similarly to the summaries of the genealogies, they can describe perturbations of the polymorphism patterns that are created by the action of recent selection; it is well known that the level of polymorphism and the number of haplotypes is reduced around the target of selection, the site frequency spectrum is shifted towards low- and high-frequency derived variants which cause negative values of Tajima's D, and the linkage disequilibrium increases on each side of the beneficial mutation (KIM and STEPHAN 2002; KIM and NIELSEN 2004; STEPHAN *et al.* 2006).

3.4 Implementation

Forward simulations have been implemented in a C++ software available from the address `http://bio.lmu.de/~pavlidis`. N diploid individuals are implemented. The number of loci l may be arbitrary. However, large (> 20) values of l may require extensive computational time. For each generation, N individuals are chosen as fathers, and N as mothers according to their fitness value. For each gender this is done by multinomial sampling with parameters N and $(F_1, F_2, \ldots, F_N)$, where F_i is the fitness of the i^{th} individual normalized by the average fitness of the population. Notice that the same individual is possible to be a mother and a father. Then, recombination occurs for each parent, and a recombinant chromosome is generated that will pass to the next generation. Random mating follows, where chromosomes from different parents merge and form the zygote. All measurements (frequencies of alleles, average fitness, average trait value etc) are calculated in the zygote step.

The code provides further extensions to the classical two- and l-locus models as this was described in Introduction. First, it allows for different optimum values for male and female individuals. Second, the optimum for the trait may change after time t_c (t_c follows either an exponential distribution, or it is predefined by the user), to a new value v_c which is either uniform or predefined by the user. Additionally, mutations can be assumed to occur for each locus. The environmental effect follows the Gaussian distribution $N(0, \sigma_E^2)$, or is absent. The effective population size is constant, but an extension to changing (stepwise) population size can be readily implemented.

3.5 Results

3.5.1 Trajectories of new variants

First, we study the two-locus two-allele symmetric model of WILLENSDORFER and BÜRGER (2003) and obtain the deterministic trajectory of a variant in the locus A_1. The goal of this analysis is to illustrate the role of the parameters of the model on the fixation of the A_{11} allele. Second, we introduce random genetic drift by simulating the evolution of a randomly mating population with effective population size $N = 10000$. Then, we relax the assumption of the symmetrical fitness matrix and finally we perform simulations of a five-locus two-allele model in order to get insight into the role of multiple loci.

Deterministic two-locus two-allele model with symmetrical fitness matrix: We implement the system described in Equation 3.1 and we record the frequency of the A_{11} allele for 10000

generations. The fitness matrix is symmetric in respect to the double heterozygous genotype (Table 3.1). The parameter values are drawn from uniform distributions whose boundaries are defined in Table 3.2. Following the analysis of WILLENSDORFER and BÜRGER (2003) the optimum value for the phenotype is set to 0. This facilitates the illustration of the results without the loss of generality. The initial frequencies for the gametes $A_{1i}A_{2j}$, $i, j = 1, 2$ are given as the product $p_0(A_{1i})p_0(A_{2j})$, and therefore the initial value of D is 0.

Table 3.2: The parameter values that were used for the simulations of the two-locus two-allele model

Parameter	Value min.	Value max.
r	0	0.5
$p_0(A_{11})$	0	0.2
$p_0(A_{21})$	0	1
ω^2	1	10
w_{11}	-2	2
w_{21}	-2	2

r: recombination fraction, $p_0(A_{11})$: initial frequency of the allele A_{11}, $p_0(A_{21})$: initial frequency of the allele A_{21}, ω^2: strength of selection, w_{11}: contribution of A_{11}, w_{21}: contribution of A_{21}.

Figure 3.2 illustrates a subset of the obtained trajectories at various levels of final frequencies. Notice that only 500 out of 10000 generations are shown because the frequencies remain constant. However, this cannot be generalized; there exist trajectories which approach the equilibrium frequency very slowly (not shown).

In Figure 3.2 we can see that, first, fixation of A_{11} allele is possible and this fixation may occur fast (within 10 generations). These trajectories are similar to the trajectories obtained from the classical selective sweep theory. Second, there is a subset of trajectories that remain polymorphic. Polymorphism is possible for various levels of equilibrium frequencies, depending on the initial conditions and parameters of the simulations. Furthermore, there is a class of trajectories that shows non-monotonic behavior. The frequency initially increases and then decreases to some equilibrium value. Figure 3.2B illustrates two non-monotonic trajectories. In the first one, the frequency approaches the value 0.5 in approximately ten generations, but then the allele disappears from the population. The second trajectory approaches fixation, and eventually it reaches the equilibrium frequency 0.5.

At the end of 10000 generations a continuum of frequencies in [0,1] is obtained, though with different probabilities. For example, frequencies in (0.5, 0.99) are rare (Figure 3.3). For the parameters of Table 3.2, the frequencies of the trajectories after 10000 generations can be summarized using the empirical cumulative distribution (Figure 3.3). Apparently, the vast majority of the fre-

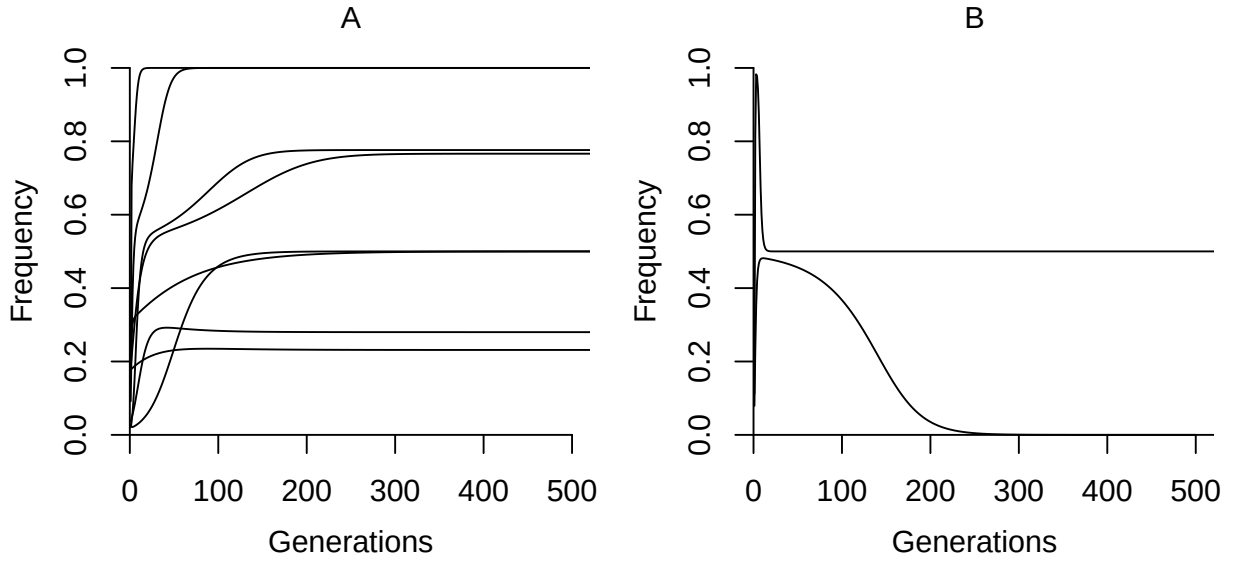

FIGURE 3.2: Trajectories obtained under the deterministic two-locus two-allele model with symmetrical fitness matrix. Only 500 out of 10000 generations are shown. The trajectories have approached their equilibrium value and the frequencies remain constant. A) monotonic trajectories, B) non-monotonic trajectories.

quencies correspond to 0, 0.5, or 1.

In order to identify the factors that determine the fixation of the A_{11} allele, we compare in pairs different sets of trajectories. For example, comparison of the trajectories that reach fixation with the trajectories that remain at frequency 0.5 gives insight into the parameter values that affect these two sets. In the next sections the following two comparisons are implemented: i) fixed trajectories (fixation class) versus trajectories that stay at equilibrium frequency 0.5 (polymorphic class), and ii) fixed trajectories versus trajectories where the allele A_{11} disappears from the population (extinction class). Throughout the text, the fixation class is defined as the set of trajectories whose equilibrium frequency is in the range (0.999, 1], the extinction class as the set of trajectories whose equilibrium frequency is in the range [0, 0.001), and the polymorphic class as the set of trajectories whose equilibrium frequency is in the range (0.499, 0.501), unless mentioned differently.

Fixation of the A_{11} allele corresponds to the equilibrium point at the vertices $A_{11}A_{21}$ or $A_{11}A_{22}$ on the tetrahedron of Figure 3.1, *i.e.* the monomorphic equilibria. Additionally, it corresponds to the absence of $A_{12}A_{21}$ and $A_{12}A_{22}$, *i.e.* equilibria at the edges of the tetrahedron. WILLENSDORFER and BÜRGER (2003) prove that two conditions are required for the stability of the monomorphic equilibria: $r \geq 1 - \alpha_1\alpha_2 \exp(\sqrt{\ln\alpha_1 \ln\alpha_2})$, where α_1 and α_2 are defined in the Introduction, and $\gamma_1 \leq 2\gamma_2$. Comparison of parameter values that result in fixation of the A_{11} allele with the pa-

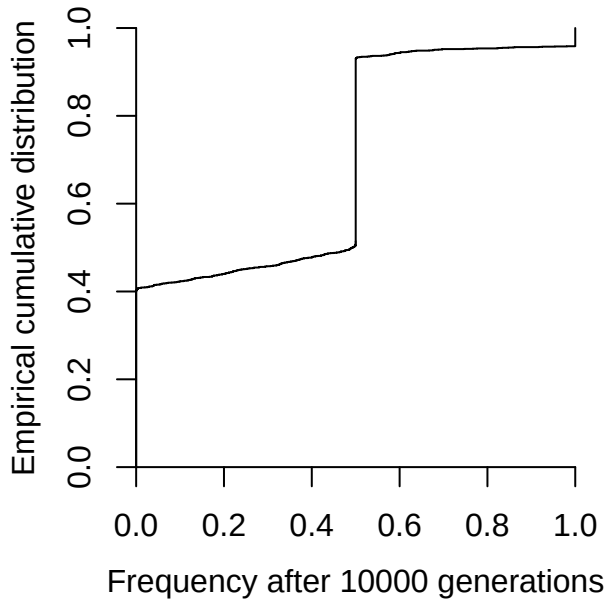

FIGURE 3.3: Empirical cumulative distribution for the frequencies obtained after 10000 simulated generations. A continuum of frequencies is obtained in [0,1]. The vast majority of frequencies for the A_{11} allele are either 0, 0.5, or 1.

rameter values that result in polymorphic equilibrium for the A_{11} allele shows that the parameters $c_1 = r - 1 + \alpha_1\alpha_2 \exp(\sqrt{\ln\alpha_1 \ln\alpha_2})$ and $c_2 = 2\gamma_2 - \gamma_1$ can separate the trajectories that fix from those that stay polymorphic. The role of c_1 is illustrated in Figure 3.4A. Given a set of simulated trajectories that are either polymorphic or result in fixation of the A_{11} allele, we plot the values for the parameter against the class of the trajectory (for this comparison class 1 means fixation, and class 0 means polymorphic). Importantly, in this plot the number of trajectories in class 1 and class 0 is equal. A similar plot is shown in Figure 3.4B for the parameter c_2, and in Figure 3.4C c_1 is plotted against c_2. In case that both c_1 and c_2 are positive (as required for the stability of the monomorphic equilibrium), then 98.4% of the trajectories belong to class 1. The stability conditions for the edge equilibria are more complicated (WILLENSDORFER and BÜRGER 2003). However, a simple condition for instability is that the edge equilibria cannot be stable when linkage is sufficiently tight.

In addition to c_1 and c_2, there are other parameters that provide information about the equilibrium state of the trajectory. Initial frequencies of the A_{21} allele close to the boundaries 0 or 1 yield fixation of the trajectory for the majority of the simulations, whereas intermediate initial frequencies lead to polymorphic equilibrium states (Figure 3.5B). For example, $\sim$76% of the trajectories that initiate at frequency < 0.1 for the A_{21} allele lead to the fixation of the A_{11} allele. Since the

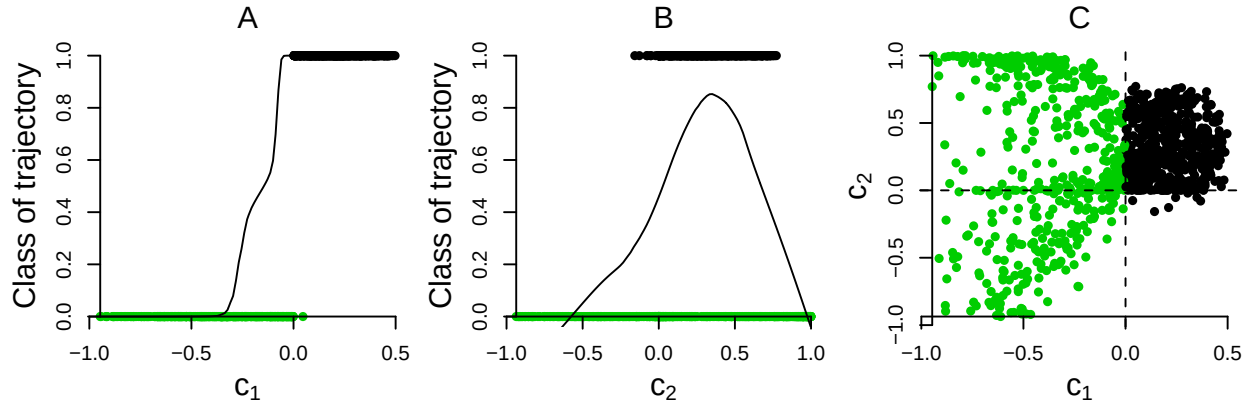

FIGURE 3.4: The role of parameter c_1 and c_2 on determining the class of the trajectory for the deterministic two-locus two-alleles model. Class 1 means trajectories that result in fixation of A_{11}, whereas class 2 includes the trajectories that stay polymorphic at frequency 0.5. The color code is as follows: green denotes class 0, black denote class 1. The line in A) and B) is the lowess smoother function for the data, and is a proxy for the probability of the class 1, given that its prior probability is 0.5. For example, in B when c_2 is 0.4, the probability of the class 1 approaches 0.8. A) The role of c_1 on predicting the class of the trajectory. $c_1$is crucial when the comparison is between monomorphic and polymorphic equilibria. In B) the role of c_2 is illustrated, whereas in C) c_1 is plotted against c_2. The dashed lines in C) denote the axes $x = 0$ and $y = 0$.

prior probability of each of the two classes is 0.5, the p-value for this event is $< 2 \times 10^{-16}$. Let the initial frequency of the A_{21} allele be close to 0 for the trajectories that result in fixation of allele A_{11} (*i.e.* black points in Figure 3.5B). Then, as illustrated in Figure 3.5C, these points are located at the proximity of the line $x = y$, *i.e.* the contributions of the A_{11} and A_{21} alleles to the phenotype are approximately equal. This means that the likelihood of the fixation of the trajectory is high when $w_{11} \approx w_{21}$, given that the initial frequency of A_{21} is low. Assuming that the initial frequency of A_{21} is low, then the majority of the genotypes for the A_2 locus will be $A_{22}A_{22}$, and a smaller proportion will be $A_{21}A_{22}$. If the contribution of the A_{21} allele is w_{22}, then the contribution of the A_{22} allele is $-w_{22}$, due to the symmetry of the model. Thus, initially the A_2 locus brings an individual $-2w_{22}$ units away from the optimum. Furthermore, since the initial frequency of the A_{11} allele is small, the majority of the genotypes at the A_1 locus will be $A_{12}A_{12}$, and a smaller proportion will be $A_{11}A_{12}$. From this initial state, there are two pathways for the population to move towards the optimum. The population will move either towards the genotype $A_{11}A_{22}/A_{11}A_{22}$ through the genotype $A_{11}A_{22}/A_{12}A_{22}$ or towards the genotype $A_{12}A_{21}/A_{12}A_{21}$ through the genotype $A_{12}A_{21}/A_{12}A_{22}$. Notice that if $w_{11} \approx w_{21}$, then the two previous final genotypes are optimal. Thus, in this case there is a competition between the two pathways, and consequently between the A_{11} allele and the A_{21} allele. Given that the initial frequency of the A_{11} allele is small, then fixation

of the A_{11} allele occurs when the frequency of the A_{21} is also very small; otherwise the A_{21} allele out-competes the A_{11} allele and the final state is the polymorphic equilibrium. This is also shown in Figure 3.5A. When the initial frequency of A_{11} is very small (as it is required for a classical selective sweep), then the probability of a trajectory that results in fixation of A_{11} is diminished. Furthermore, comparing the lowest frequencies for the A_{11} allele in class 0 and class 1, we observe that the lowest initial frequency of A_{11} in class 1 is at least one order of magnitude greater than in the class 0 (6.1×10^{-3} versus 1.1×10^{-5}). This means that classical selective sweeps (as described by MAYNARD SMITH and HAIGH (1974)) may be rare under the symmetrical fitness model compared to sweeps from standing genetic variation.

Another way to explain these results is the following. As mentioned above, assuming that the initial frequency of A_{21} is low, then the majority of the genotypes for the A_2 locus will be $A_{22}A_{22}$. Also, for the A_1 locus the majority of the genotypes will be $A_{12}A_{12}$. Given that the trajectory will result in fixation of the A_{11} allele, *i.e.* that the final state for the A_1 locus will be $A_{11}A_{11}$, then (under the symmetric model) it is required that $w_{11} \approx w_{21}$, so that the $A_{11}A_{11}$ genotype will cancel out the effect of the $A_{22}A_{22}$ genotype on the phenotype, and it will bring the individuals to the optimum. If the difference between w_{11} and w_{22} is large then the only optimum genotype is the double heterozygote, *i.e.* fixation of the A_{11} allele is unlikely. When the initial frequency of the A_{21} allele is close to 1, then the previous argument holds when $w_{11} \approx -w_{21}$ (blue points in Figure 3.5C).

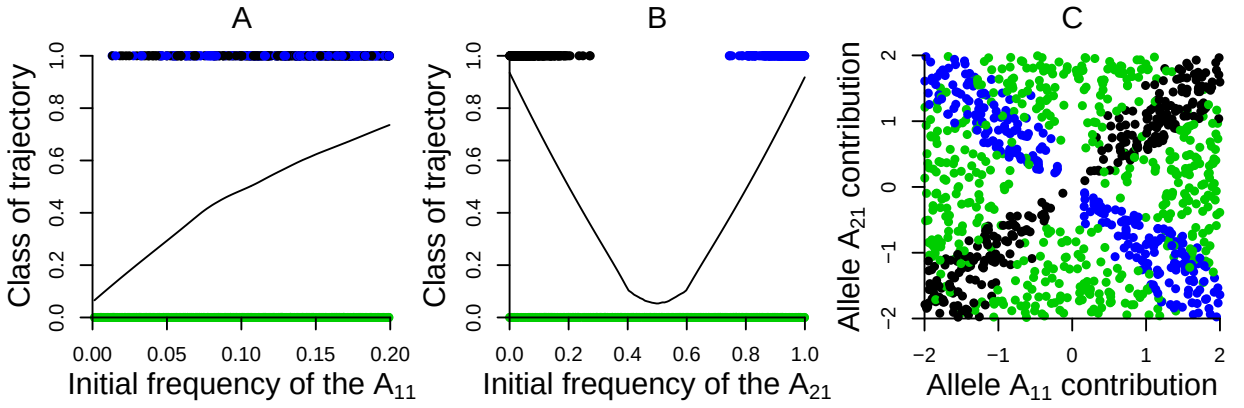

FIGURE 3.5: The role of parameters $p_0(A_{11})$ and $p_0(A_{21})$ on determining the class of the trajectory for the comparison of the fixation class versus the polymorphic class. The line in A) and B) represents the lowess smoothing function for the data. As shown in A), for very small values of $p_0(A_{11})$ the probability of obtaining a trajectory from the fixation class decreases considerably. In B) the initial frequency of A_{21}, $p_0(A_{21})$, shows a non-monotonic behavior: small and large values of $p_0(A_{21})$ make possible the fixation of the A_{11}. In C) we can see how the contributions of the alleles interact. When $w_{11} \approx w_{21}$ then it is possible to obtain trajectories that reach fixation.

Even though the parameters c_1 and c_2 can separate the trajectories that result in fixation from those that remain polymorphic, they cannot disentangle fixation from extinction of the A_{11} allele (Figure 3.6). This is because the obtained equilibria are either monomorphic or at the edges.

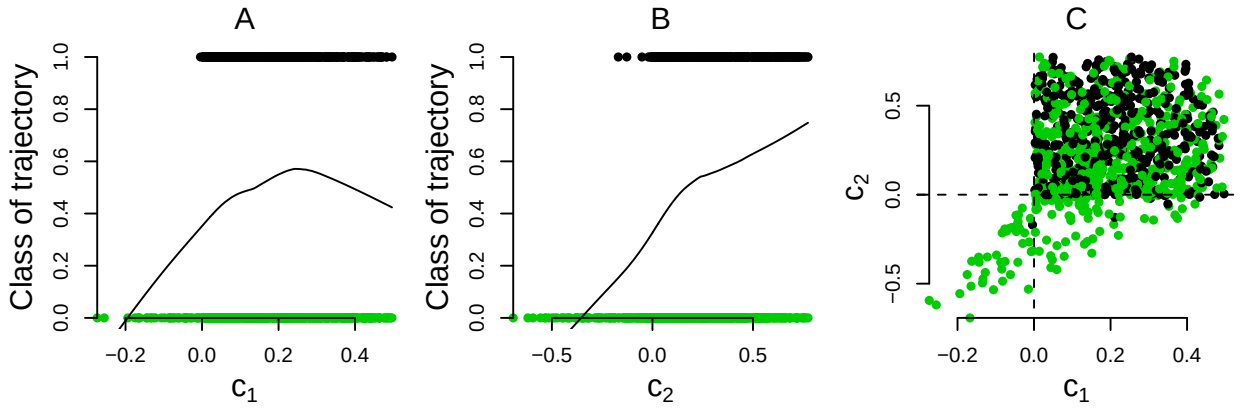

FIGURE 3.6: The role of parameter c_1 and c_2 on determining the class of the trajectory when class 1 and class 0 represent fixation and extinction of the A_{11} allele, respectively. The color code is as follows: green denotes class 0, black denote class 1. The line in A) and B) is the locally weighted scatterplot smoothing (lowess) function for the data (CLEVELAND 1979), and is a proxy for the probability of the class 1, given that its prior probability is 0.5. A) The role of c_1 on predicting the class of the trajectory. In B) the role of c_2 is illustrated, whereas in C) c_1 is plotted against c_2. The dashed lines in C) denote the axes $x = 0$ and $y = 0$. As expected, c_1 and c_2 cannot separate fixation from extinction for the A_{11} allele in the case of the symmetrical fitness model.

Figure 3.7 reveals that the initial frequency of the A_{21} allele is important for the equilibrium state of the trajectory. In particular, it shows that low and high initial frequencies of A_{21} may lead to fixation of A_{11}, whereas intermediate frequencies result in the extinction of A_{11}. Notice that in this comparison, the contributions w_{11} and w_{21} are located on the diagonals $x = y$ and $x = -y$ for both of the classes. This is because in equilibrium states the genotypes $A_{11}A_{22}/A_{11}A_{22}$ are optimal for the case of fixation, and the genotypes $A_{12}A_{21}/A_{12}A_{21}$ for the case of extinction of the A_{11} allele, *i.e.* homozygous states for both of the loci. The genotypes $A_{11}A_{21}/A_{11}A_{22}$ and $A_{12}A_{21}/A_{12}A_{22}$ are away from the optimum (given that w_{11} is not 0) and therefore it is improbable to dominate in the equilibrium state. Based on the previous arguments we may assume that the double homozygote dominates in the equilibrium state. Thus, it is required, in order for the phenotypic value of the individuals to be near the optimum, that the contributions w_{11} and w_{22} are in the proximity of the diagonal $y = x$ and $y = -x$ (Figure 3.7B).

Stochastic two-locus two-allele model with symmetrical fitness matrix: In this section we study the behavior of the stochastic model when the fitness matrix is symmetrical. The population

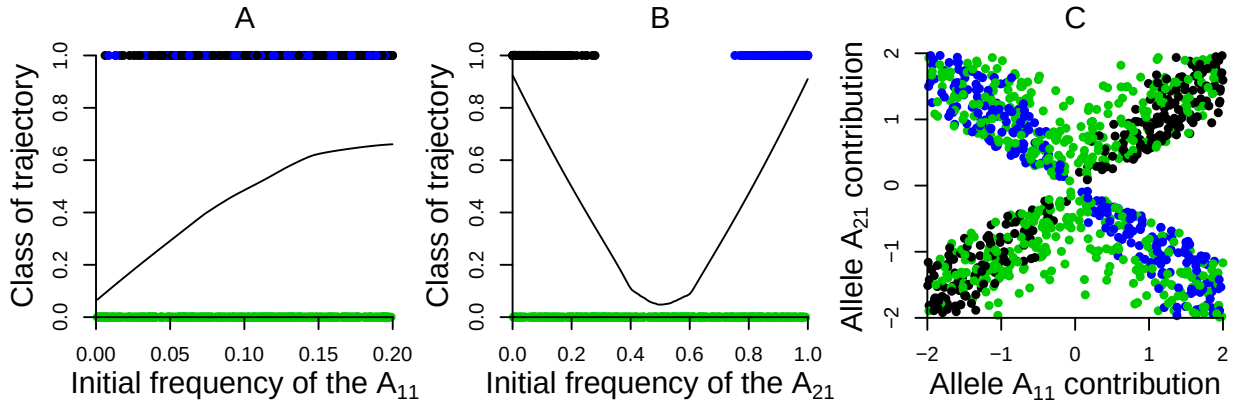

FIGURE 3.7: The role of parameters $p_0(A_{11})$, $p_0(A_{21})$ and allele contributions w_{11} and w_{21} on determining the class of the trajectory for the comparison the fixation class versus the extinction class. The line in A) and B) represents the lowess smoothing function for the data. As shown in A), for very small values of $p_0(A_{11})$ the probability of obtaining a trajectory from the fixation class decreases considerably. In B) the initial frequency of A_{21}, $p_0(A_{21})$, shows a non-monotonic behavior: small and large values of $p_0(A_{21})$ make possible the fixation of the A_{11}. In C) we can see how the contributions of the alleles interact. When $w_{11} \approx w_{21}$ then it is possible to obtain trajectories that reach fixation, but also trajectories where the A_{11} disappears from the populations.

size $N = 10000$. The simulation parameters are similar to the deterministic two-locus two-allele model with symmetrical fitness matrix. We use the average frequency of the last 500 generations, $\hat{f}_{500}$, to define the equilibrium frequency. This is because the frequency of the A_{11} does not remain constant but fluctuates due to random genetic drift. In Figure 3.8, we plot the empirical cumulative distribution of the $\hat{f}_{500}$. Comparing Figure 3.8 with Figure 3.3 we can see that the proportion of trajectories with the equilibrium frequency 0.5 is largely reduced in the stochastic model. This is expected as a consequence of random genetic drift, which drives the frequency of the trajectory towards its absorbing state.

The trajectories we obtain in the stochastic model are similar to those of the deterministic model. Figure 3.9 illustrates trajectories fluctuating at various equilibrium levels. In particular, Figure 3.9B shows non-monotonic trajectories, where the frequency of A_{11} approaches the value 0.5 but eventually it disappears from the population.

In order to determine the importance of various parameters we plot them against the class of the trajectory. As previously, three classes of trajectories are used in two comparison sets: i) trajectories that result in fixation versus trajectories that stay in a polymorphic equilibrium, and ii) trajectories that result in fixation versus trajectories that result in extinction. The definitions of the three trajectory classes (fixation, polymorphic, extinction) are given in the section **Deterministic**

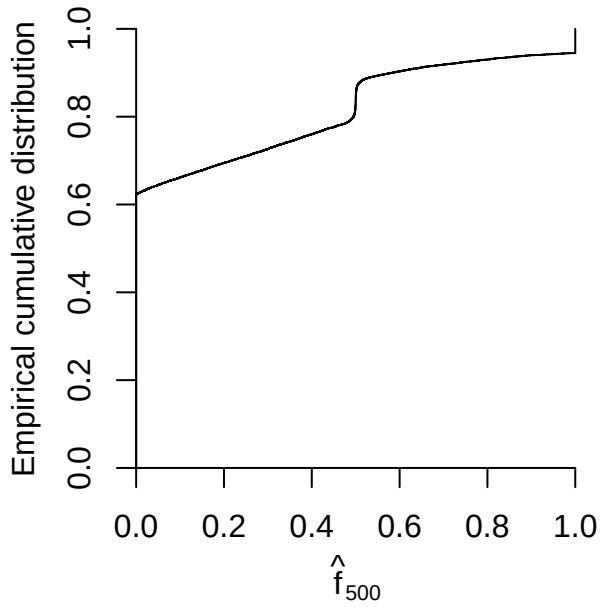

FIGURE 3.8: The empirical cumulative distribution for the equilibrium frequency of the stochastic two-locus two-allele model.

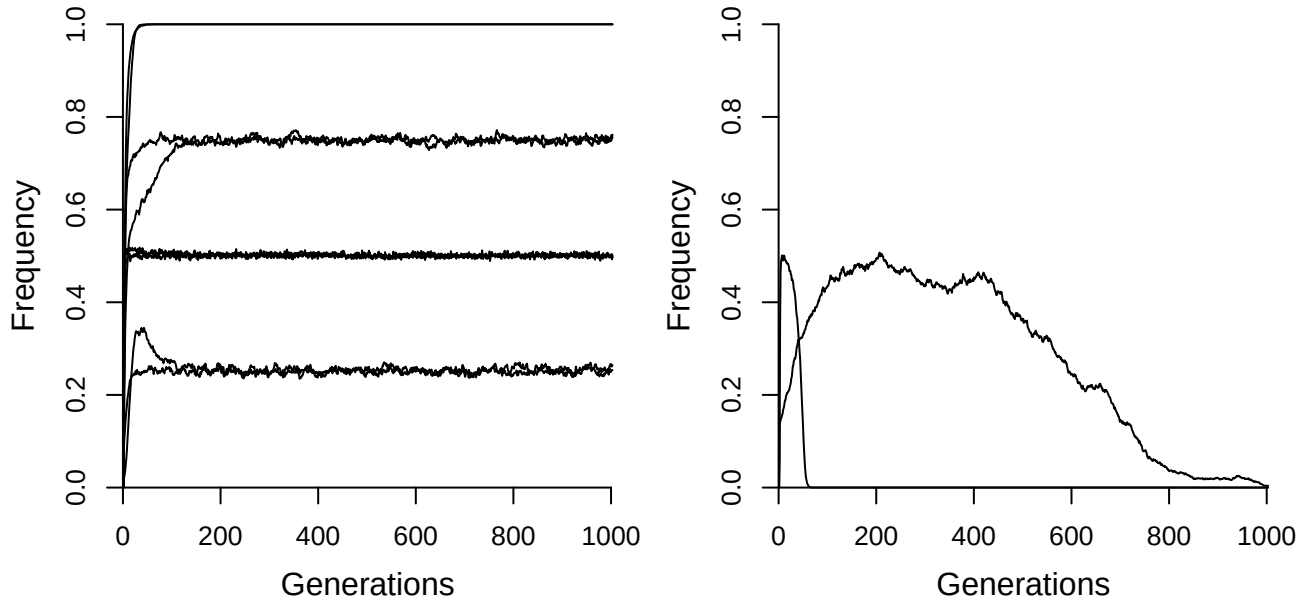

FIGURE 3.9: Examples of trajectories in the stochastic two-locus two-allele model. 1000 out of 10000 generations are shown. A) Trajectories at various equilibrium levels. B) Non-monotonic trajectories that approach the value 0.5 but eventually the allele A_{11} is getting extinct from the population.

two-locus two-allele model with symmetrical fitness matrix. For the first comparison set, the relation of six parameters with the class of the trajectory is depicted in Figure 3.10. We observe that the initial frequency $p_0(A_{11})$ of the A_{11} allele, the strength of selection ω^2, and the c_1 parameter reveal a linear relationship with the frequency of obtaining a trajectory of class 1. On the other hand, the initial frequency $p_0(A_{21})$ of the A_{21}, and the contribution of the alleles A_{11} and A_{21} are non-linear. In particular we observe that large absolute values for the contribution of A_{21} and small absolute values for the contribution of A_{11} favor the fixation of the A_{11} allele against a polymorphic equilibrium state (Figure 3.10).

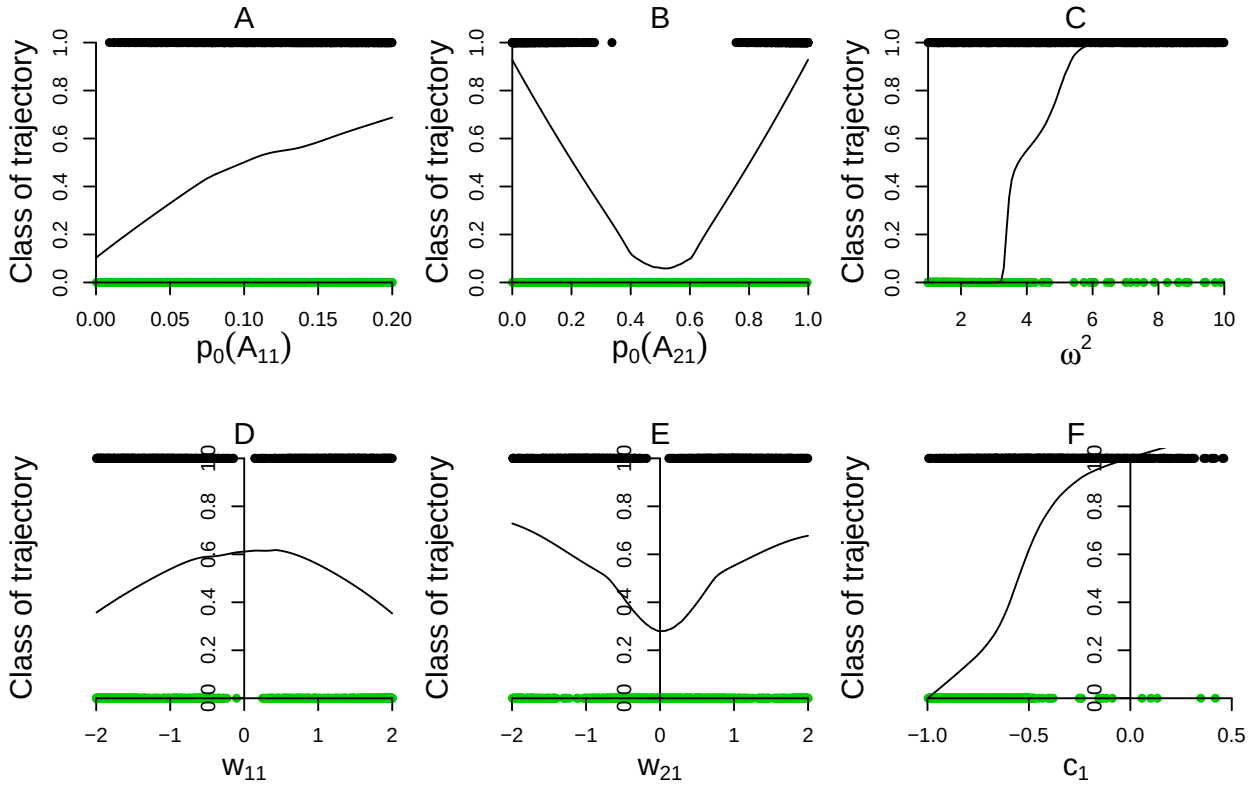

FIGURE 3.10: The relation of six parameters to the class of the trajectory. The line in each subfigure represents the lowess smoothing function of the data. A) The initial frequency of the A_{11} allele. B) The initial frequency of the A_{21} allele. C) The parameter ω^2 which defines the strength of selection. D) The contribution of the allele A_{11} to the genotypic value. E) The contribution of the A_{21} to the genotypic value. F) The parameter c_1. The red line represents the lowess smoother function for the data of the plot. Black points represent class 1 (trajectories that result in the fixation of A_{11}), whereas green points represent class 0 (trajectories that result in a polymorphic state for the A_{11}).

Comparing the plots from the stochastic simulations with the deterministic simulations (for the

symmetric fitness model) reveals that larger values of the parameter ω^2 (weaker selection) result in a higher frequency of fixed trajectories. However, the role of ω^2 is more crucial for the stochastic simulations. For small values of the ω^2 ($1 < \omega^2 < 2$) the frequency of the polymorphic trajectories is $\sim 88\%$ in the stochastic set, and $\sim 66\%$ in the deterministic set. On the other hand, for large values of the ω^2 ($9 < \omega^2 < 10$) the frequencies are 4.6% and 38%, respectively. This shows (see also Figure 3.10C) that especially in the stochastic model the frequency of polymorphic trajectories is very low when selection is not strong enough, whereas in the deterministic model the relation between the ω^2 and the class of the trajectory (fixed versus polymorphic) is not so crucial. The reason for this result is that in order to preserve the polymorphic state of the trajectory, selection has to overcome the effects of recombination and random genetic drift in the stochastic set, but only the effect of recombination must be overcome in the deterministic trajectory. Therefore, selection needs to be stronger in the stochastic simulations in order to result in a polymorphic equilibrium state. This is especially true for the polymorphic state around the frequency 0.5 (see Discussion). Regarding the relation between the contributions of each locus to the genotypic value, the results are similar to the deterministic model: the parameter values for the trajectories that fix are located in the proximity of the two diagonals $w_{11} = w_{11}$ or $w_{11} = -w_{21}$ depending on the initial frequency of the allele A_{21} (see the section **Deterministic two-locus two-allele model with symmetrical fitness matrix**).

The following results have been obtained for the comparison of the fixation class versus the extinction class in the stochastic simulation set. The roles of ω^2 and c_1 are not critical (Figure 3.11). This means that small and large values of the ω^2 have similar effects on the class of the trajectory. In this comparison both of the sets are associated with monomorphic (absorbing) states of the alleles. The strength of selection (at least for the values tested here) is not crucial, because maintaining either of the classes does not require strong selection. The importance of c_1 has been explained in the section **Deterministic two-locus two-allele model with symmetrical fitness matrix**. In brief, c_1 is not informative for disentangling the monomorphic equilibria. Interestingly, the roles of w_{11} and w_{21} have been inverted in this comparison. Values close to 0 are associated with class 0 for the w_{11} and with class 1 for the w_{21}, whereas the relation was inverted in the comparison of the fixation class versus the polymorphic class.

Deterministic two-locus two-allele model with generalized fitness matrix: In this section we relax the assumption of symmetry of the fitness matrix. The parameter space is given in Table 3.2. Essentially, the difference between this model and the symmetrical fitness model is that there is no restriction on the relations between the contributions of the alleles (see Introduction for the restrictions in the symmetrical fitness model). Thus, all four alleles A_{11}, A_{12}, A_{21}, and A_{22} may

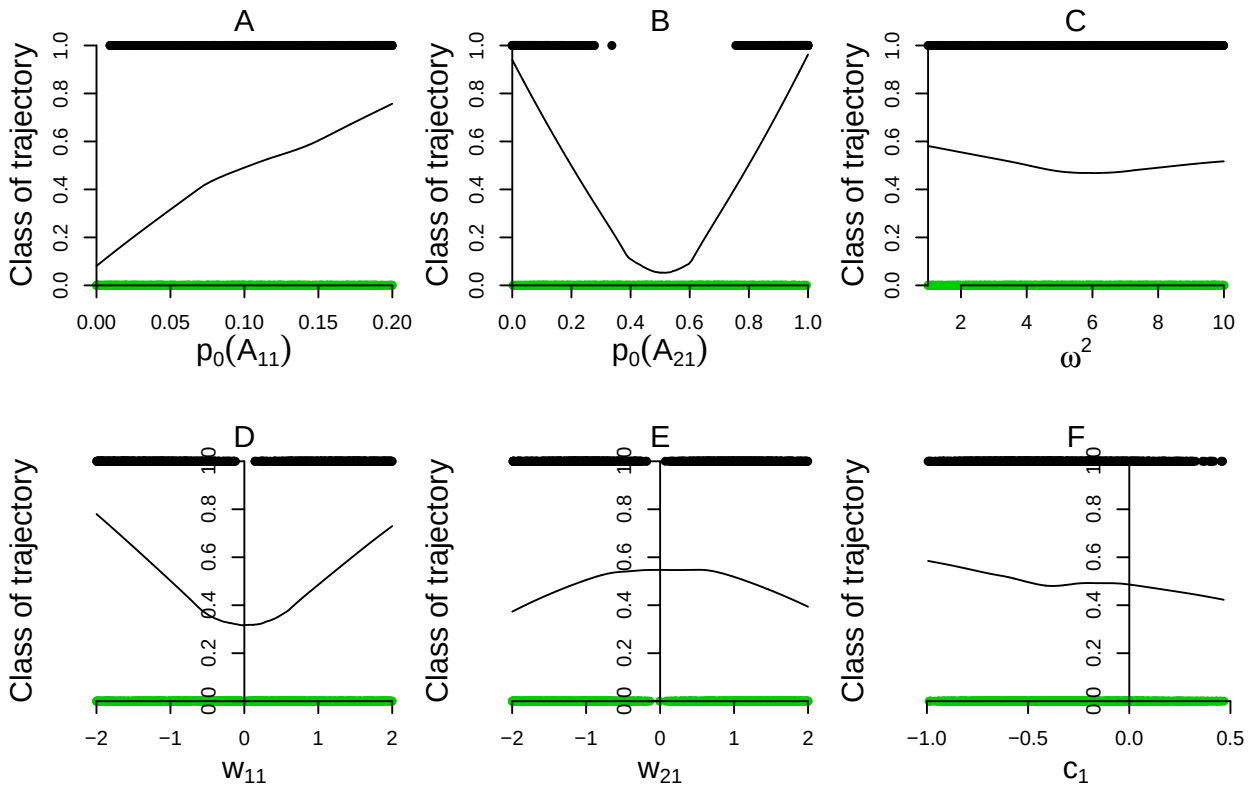

FIGURE 3.11: The relation of six parameters to the class of the trajectory for the comparison of the fixation class versus the extinction class. The line in each subfigure represents the lowess function of the data. A) The initial frequency of the A_{11} allele. B) The initial frequency of the A_{21} allele. C) The parameter ω^2 which defines the strength of selection. D) The contribution of the allele A_{11} to the genotypic value. E) The contribution of A_{21} to the genotypic value. F) The parameter c_1. The red line represents the lowess smoother function for the data of the plot. Black points represent class 1 (trajectories that result in the fixation of A_{11}), whereas green points represent class 0 (trajectories that result in the extinction of A_{11}).

assume any value in the parameter space [-2, 2].

The shape of the trajectories in this model is similar to the symmetrical fitness matrix model. The number of trajectories where the A_{11} allele disappears is similar for both models. However, in the case of the generalized fitness matrix model, the trajectories that result in the fixation of the A_{11} allele occur more often than in the symmetrical fitness model. This is shown by the comparison of the empirical distributions for the frequency of the trajectories after 10000 generations (Figure 3.3 and Figure 3.12). In the symmetrical fitness model (for the parameter values studied here) 4.15% of the trajectories are fixed, whereas 34.65% of the trajectories are fixed in the generalized fitness model. On the other hand, less trajectories stay at equilibrium frequency 0.5 (Figure 3.12). This could be expected because in the generalized fitness model, the double heterozygous genotype is not associated with the highest fitness.

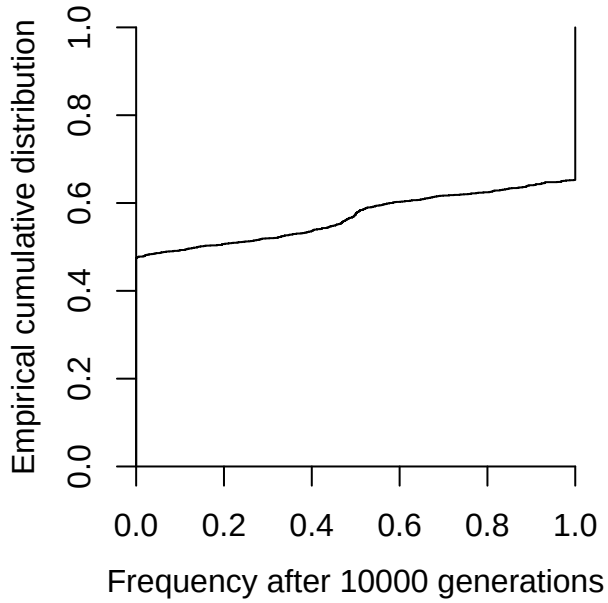

FIGURE 3.12: Empirical cumulative distribution of the frequencies of the trajectories after 10000 generations for the deterministic two-locus two-allele model assuming a generalized fitness matrix. In contrast to the symmetrical fitness model, where 4.15% of the trajectories fix, in the generalized fitness model 34.65% of the trajectories reach fixation. On the other hand, the percentage of trajectories that stay at equilibrium frequency 0.5 is smaller.

The role of parameters c_1 and c_2 in disentangling fixation from the polymorphic class is not so clear as in the symmetrical fitness model. As shown in Figure 3.13, increasing values for c_1 result in an increasing probability for the fixation of the allele A_{11}. However, this is not as clear-cut as in the symmetrical fitness matrix model (Figure 3.4C and Figure 3.13C). Furthermore, the initial

frequencies of A_{21} and A_{11} do not have an impact on the fixation of the trajectory. The lowest frequency of the A_{11} allele observed is 4.3×10^{-5}, much lower than the minimum frequency of the A_{11} allele in the symmetrical fitness model (6.1×10^{-3}). Additionally, the patterns observed in Figure 3.13C regarding the contributions of the alleles to the phenotype in the symmetrical fitness model are not observed in the generalized fitness model. In summary, the results indicate that in the generalized fitness models classical selective sweeps from rare variants may occur more often than in the symmetrical fitness model. However, there is no simple relation between the parameters that determine the fate of the trajectory. The results are similar when comparing the fixation class versus the extinction class.

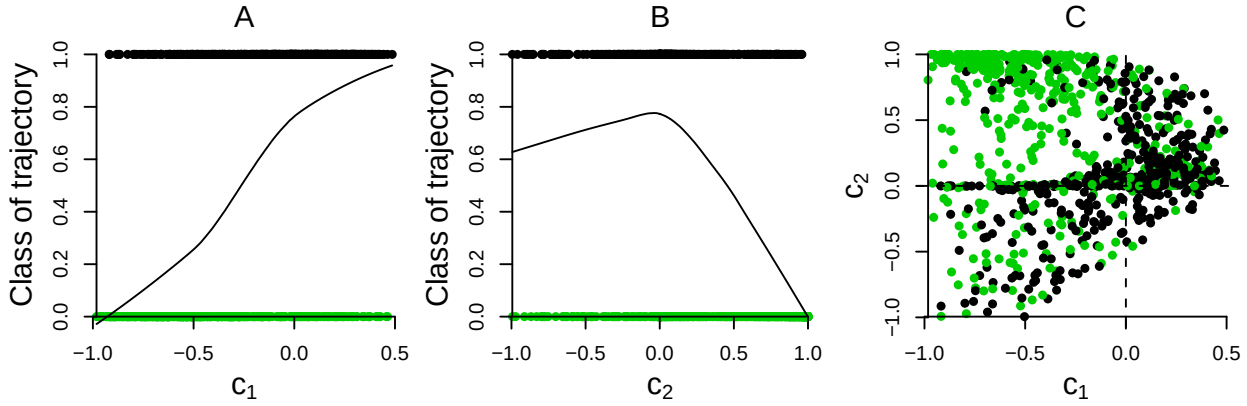

FIGURE 3.13: The roles of parameters c_1 and c_2 in the deterministic two-locus two-allele model with generalized fitness matrix for the comparison of the fixation class versus the polymorphic class. The line in A) and B) is the lowess smoothing function of the data. c_1 is informative for the class of the trajectory (in A). On the other hand, c_2 (in B) and the combination c_1 *and* c_2 (in C) appear to be less informative.

An informative quantity for disentangling trajectories in which A_{11} is getting fixed from those in which it stays polymorphic or disappears is the mean trait value in the beginning of the evolutionary trajectory. For mean initial trait values close to the optimum value 0, trajectories result in either extinction or polymorphic equilibrium for the A_{11} allele (Figure 3.14). On the other hand, when the initial mean trait is far from the optimum, then fixing the A_{11} allele becomes more probable. When the mean value for the trait under selection is far from the optimum, then the allele A_{11} can be beneficial. On the other hand, when the population is already at the optimum or close to it, then the A_{11} allele will not be favored in general.

Stochastic five-locus two-allele model with symmetric fitness matrix: In this section the five-locus two-allele models is analyzed. The parameter space that we have used is a direct extension of the two-locus model. Therefore, the contribution w_{i1} of each A_{i1} allele, $i = 1, \ldots, 5$ is in

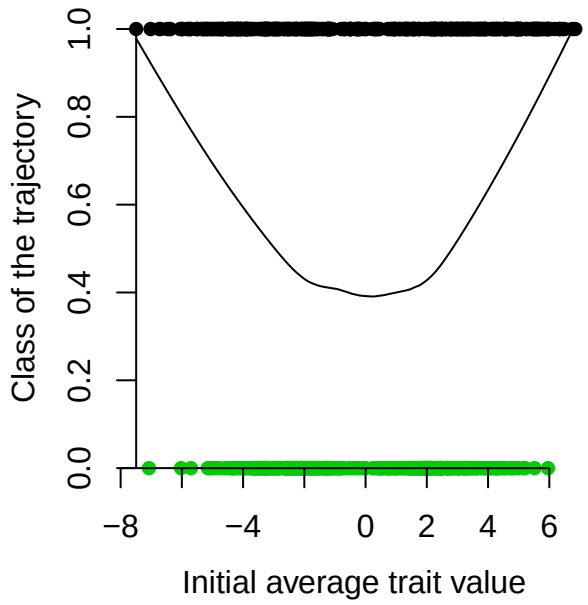

FIGURE 3.14: The effect of the initial trait value on the fate of the trajectory. The line represents the lowess function of the data. When the initial mean trait value is close to the optimum then the A_{11} allele fixes less often than when the mean trait value is far from the optimum. Results are similar when the comparison is between fixation versus polymorphic trajectories or fixation versus extinction trajectories.

[-2, 2] and the contribution w_{i2} of the A_{i2} allele is equal to w_{i1}, as it is required for the symmetrical fitness model. All the remaining parameters are similar to the two-locus two-allele model. The empirical cumulative distribution of the $\hat{f}_{500}$ is shown in Figure 3.15. In the five-locus model the proportion of the equilibrium trajectories is reduced compared to the two-locus model, and monomorphic states are obtained more frequently than the two-locus model.

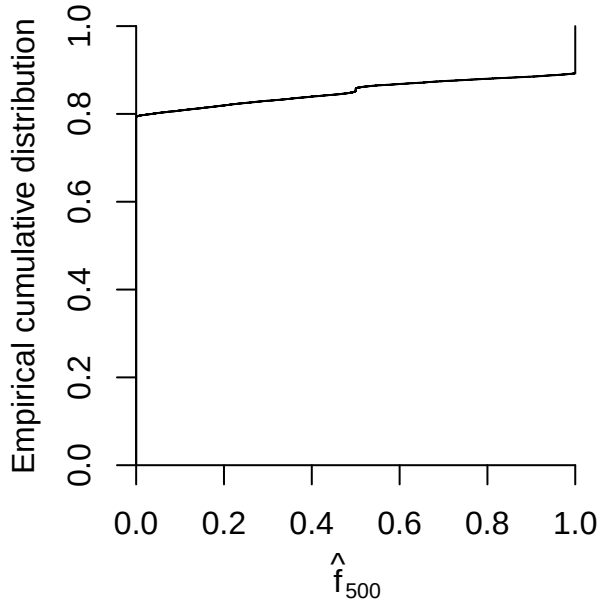

FIGURE 3.15: The empirical cumulative distribution of the five-locus two-allele model with symmetrical fitness matrix. Compared to the two-locus model, the five-locus model shows an increased proportion of monomorphic trajectories.

In contrast to the two-locus two-allele model the pairwise relations between the contributions of the alleles do not show the patterns that were observed in Figure 3.10. This is because more than two loci determine the phenotype. Therefore, pairwise comparisons may reveal no information about the relations of the contributions that are required for the fixation of the A_{11} allele (Figure 3.16). Furthermore, as Figure 3.16 shows, the frequency of fixed trajectories depends on the initial frequency of the A_{11} allele. However, the dependency is not so strong as in the two-locus model. The effect of ω^2 is similar to the two-locus model. Strong selection is required in order to maintain the polymorphic state. Comparing the results from the two-locus and five-locus models we conclude that fixation of the A_{11} allele occurs more frequently in the five-locus model than in the two-locus model. This is because heterozygote states are maintained efficiently only when strong selection is applied.

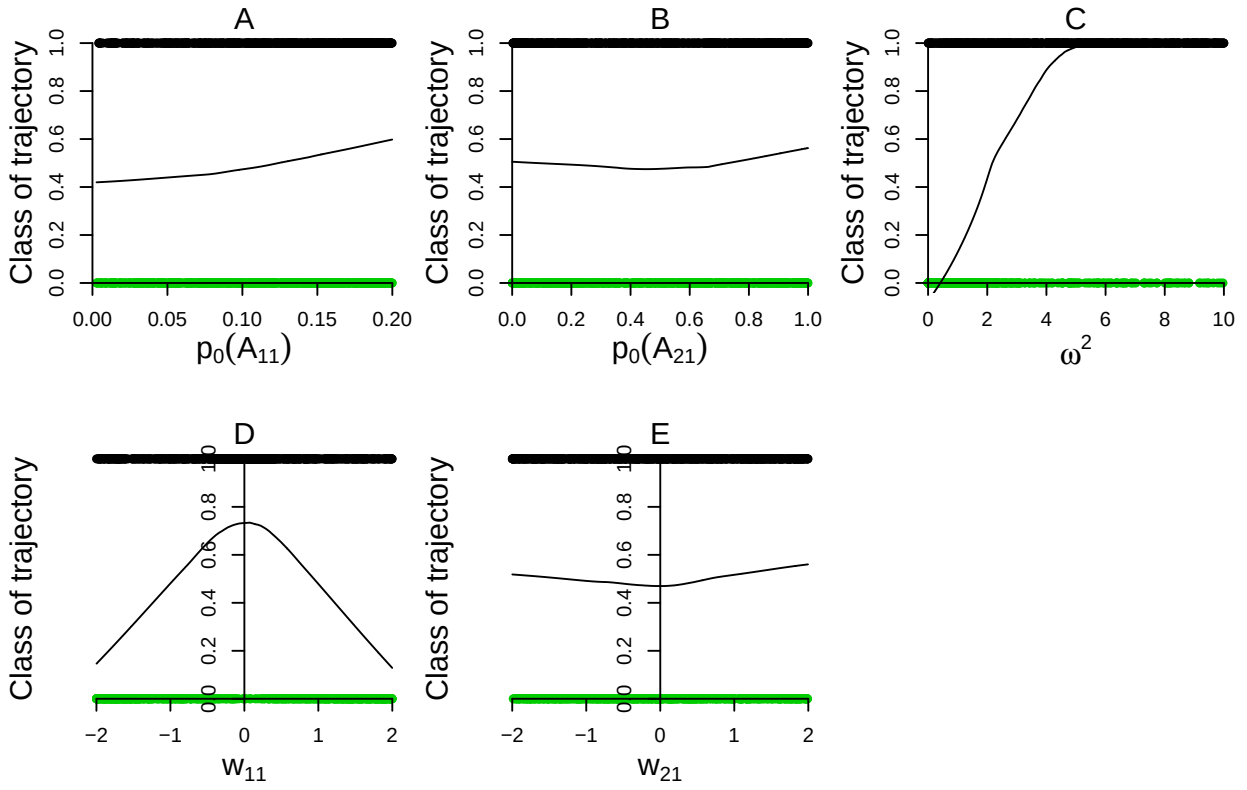

FIGURE 3.16: The role of various model parameters in determining the class of the trajectory for the comparison of the fixation class versus the polymorphic class. In contrast to the two-locus two-allele model the initial frequency $p_0(\mathrm{A}_{11})$ is not so critical for the class of the trajectory. The ω^2 is related linearly to the class of the trajectory, whereas the w_{11} shows a non-monotonic behavior, similar to the two-locus two-allele model.

3.5.2 Coalescent simulations conditioning on the trajectory of the A_{11} allele

In this section we perform coalescent simulations in order to obtain i) the genealogies and ii) the neutral polymorphism patterns in the neighborhood of the A_1 locus. The results in this section are approximate because of two reasons. First, conditioning on the frequency of one allele implies that the coalescent rates of all the genotypes that carry this allele are equal. However, in the case of multiple locus models this is not true. For example, the coalescent rates of the A_{11} allele is different when it is located on the gametes $A_{11}A_{21}$ and $A_{11}A_{22}$ since the dynamics of these two gametes is different (see Equation 3.1). This is also shown in Figure 3.17, where a random pair of trajectories for the $A_{11}A_{21}$ and $A_{11}A_{22}$ gametes are drawn for the deterministic two-allele two-locus model with symmetrical fitness matrix. The growth rate of these trajectories and their equilibrium frequencies are different. Therefore, the coalescent rate of the A_{11} allele is depending on the gamete that carries it.

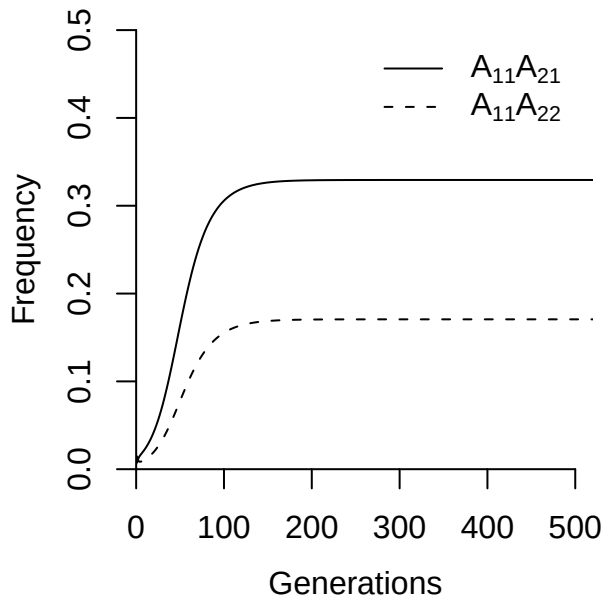

FIGURE 3.17: The growth rates for the $A_{11}A_{21}$ and $A_{11}A_{22}$ trajectories are different. The $A_{11}A_{21}$ trajectory grows faster than the $A_{11}A_{22}$ trajectory. Therefore, the coalescent rate for the allele A_{11} is different when it is located on the gamete $A_{11}A_{21}$ or on $A_{22}A_{22}$.

The second reason is that both loci A_1 and A_2 affect the dynamics of a neutral locus, when the recombination fraction is smaller than 0.5. Thus, simulating the genealogy of a neutral site would require tracking the frequencies of all the gametes backward in time instead of tracking the frequency of a single allele. Such an analysis is beyond the goals of the present article. Combining

the two previous arguments indicates that the results of this section are valid mainly for cases of weak selection and not too tightly linked loci.

Properties of the coalescent: Given a trajectory, coalescent simulations require to specify the time that the backward process is considered. For example, given a trajectory from Figure 3.9 the genealogies will be strikingly different if the backward process initiates 100 or 5000 generations after the onset of the A_{11} allele. Thus, an arbitrary time point is required, which represents the beginning of the backward simulation process. Here, we have used 100 generations after the trajectory has reached its equilibrium frequency. This time point is temporally close to the onset of the A_{11} allele. Therefore, the signature of the trajectory on the neutral polymorphisms is still present on the data. Backward simulations have been performed using a modified version of the software mbs (TESHIMA and INNAN 2009). Our mbs algorithm implements the infinite site model, in contrast to the original software, and it calculates and outputs statistics related to the coalescent trees, such as the height, the total length, and the balance of the coalescent (see also Equation 3.2). For the coalescent simulations we have used parameters related to human data. Assuming that the mutation rate $\mu = 10^{-9}$ per nucleotide per year (*e.g.* ZHAO *et al.* 2006), then $\theta = 4N\mu = 0.001$ per bp per generation. The ratio $\rho/\theta = 1$. The effective population size $N = 10000$ and remains constant. Simulations are performed for a 0.5-Mb genomic segment. The A_1 locus is located on the middle of the simulated segment. The sample size is 50. For a given equilibrium frequency-bin (see below), we have chosen randomly one trajectory whose initial frequency is below 0.001. This is done in order to resemble closely a selective event of a new variant. For a given trajectory, 1000 coalescent simulations are performed. Finally, the summary statistics for the coalescent trees are computed at the recombination breakpoints for each simulation, and the results are binned. For example, if the positions $x_1 = 103989$, $x_2 = 103995$, $x_3 = 105000$ are breakpoints (*i.e.* the genealogy may change), and the bin size is set to 100, then x_1 and x_2 will be in the same bin (1039), whereas the x_3 will be in a different bin (1050). The results from the same bin are averaged over the whole set of simulations. We repeat this process for four sets of trajectories in which the equilibrium frequency is (i) 1, (ii) 0.9 to 1, (iii) 0.3 to 0.4, and (iv) trajectories that show the non-monotonic behavior. The results presented here are obtained from the analysis of the two-locus two-allele stochastic model with symmetrical fitness matrix. The results for the remaining models are similar because the shape of the trajectories is similar for the various equilibrium frequencies (Figure 3.18). For trajectories that result in fixation (Figure 3.18A), the signatures of selective sweeps emerge in the proximity of the locus under selection: coalescent trees are shorter in length and height, and simultaneously they are imbalanced in the proximity of the A_1 locus. For trajectories that result in polymorphic equilibrium the signatures

are weaker. For example when the equilibrium frequency is between 0.9 and 1, the total length of the coalescent is smaller, and the tree imbalance as measured by the b_N statistic is higher. However, the height of the coalescent tree is similar to the neutral expectation (red solid line in Figure 3.18). Interestingly, the imbalance of the coalescent is higher for trajectories in the bin [0.9, 1) than the trajectories that result in fixation of the A_{11} allele in the proximity of the A_1 locus. The explanation is as follows. When fixation of the A_{11} allele has occurred, *all* genealogical lines coalesce recently in the proximity of the locus A_1. Thus a short tree is generated which is not imbalanced because no line has escaped the coalescence. Imbalance is generated further from the locus A_1, because recombination breaks the link between a neutral site and the A_1. On the other hand, trajectories in the bin [0.9, 1) generate imbalanced genealogies very close to the locus A_1 because a large fraction of the present-day lines carry the allele A_{11} (and coalesce recently in the past), whereas a small fraction of the present-day lines carry the allele A_{21} and coalesce further in the past.

Properties of the polymorphism patterns: We have used classical population genetics summary statistics to describe the properties of the polymorphisms in the proximity of the A_{11} allele. We use the same simulation sets as in the previous section. A sliding-window approach with window-length 5kb and offset 1kb has been implemented. The length of the genomic fragment and the position of the A_1 locus are provided in the previous section. The summary statistics are described in Methods. For each window the mean value of each summary statistic is calculated over the simulated datasets. Figure 3.19 illustrates the results.

Tajima's D is negative over the whole region for frequencies > 0.9. For fixed trajectories, Tajima's D becomes less negative closely to the A_1. For trajectories close to fixation, Tajima's D obtains its most negative value exactly at the location of the locus A_1. Comparing Figure 3.19 with Figure 3.18 we can associate Tajima's D with the b_N statistic. The number of polymorphic sites follows, as expected, the statistic L, which measures the total length of the coalescent. The number of haplotypes K and the haplotypic diversity H are also informative about the locus A_1.

3.6 Discussion

In this study, we explore selective sweeps in multi-locus two-allele models. Selection applies to the phenotypic values through a Gaussian fitness function. The Gaussian function seems an appropriate choice for many quantitative traits (ENDLER 1986; WILLENSDORFER and BÜRGER 2003), because it naturally formalizes the concept of an optimum value. Furthermore, it is flexible enough to allow for modeling both stabilizing and directional selection. Stabilizing selection is modeled by assuming that the optimal genotypic value is located between the extreme genotypic

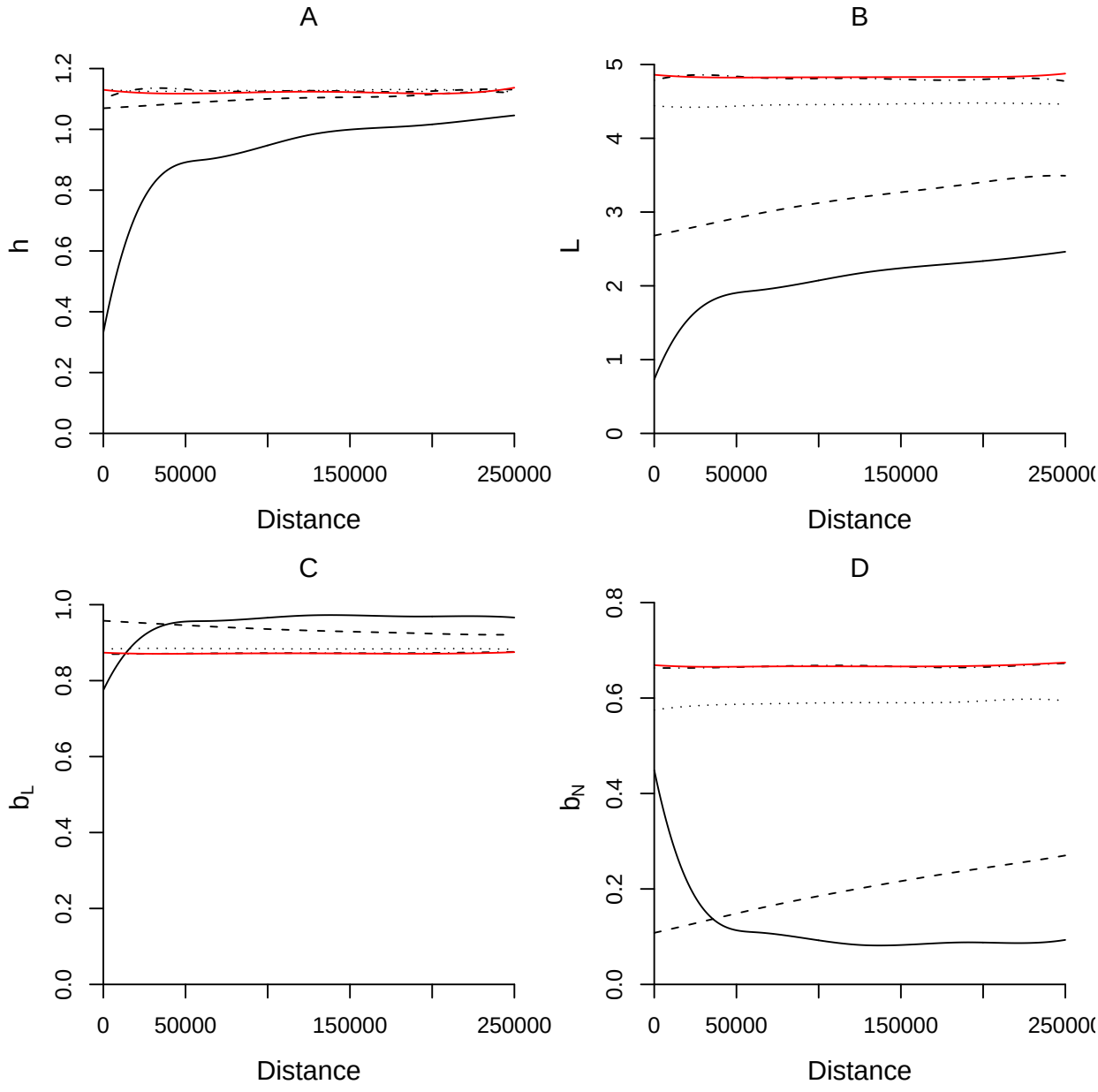

FIGURE 3.18: Summary statistics for the coalescent trees as a function of the distance from the locus A_1. The x-axis denotes the distance from the A_1. The solid line refers to the equilibrium frequency 1 (fixation), the dashed line refers to the equilibrium frequency in [0.9, 1), the dotted line refers to the frequency [0.3, 0.4) and the red line to neutral simulations with the same parameter values. Notice that the results for the non-monotonic trajectories overlap completely with the neutral curves. In A) the height of the tree is shown. B) shows the total length of the coalescent, and C) and D) the balancing statistics b_L and b_N, respectively.

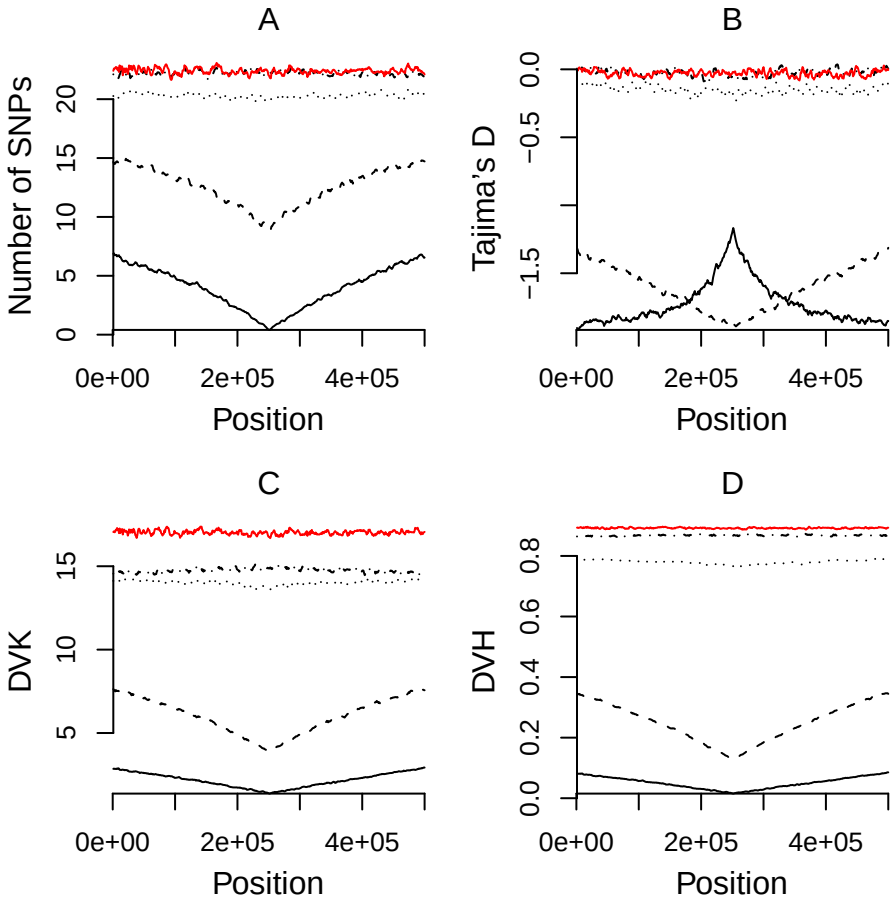

FIGURE 3.19: Sliding window analysis of a 500-kb genomic region flanking the locus A_1. For each window the average value of the statistic is calculated over 1000 simulated datasets. The window length is 5kb and the step length 1kb. The solid line refers to the equilibrium frequency 1 (fixation), the dashed line refers to the equilibrium frequency in [0.9, 1), the dotted line refers to the frequency [0.3, 0.4), the dotted-dashed line to the non-monotonic trajectories that reach frequency 0.5 but eventually go to extinction, and the red line to neutral simulations with the same parameter values. A) Average number of polymorphic sites. B) Tajima's *D*. C) The number of haplotypes calculated by the Depaulis and Veuille *K* (DVK). D) The haplotype diversity measured by the Depaulis and Veuille *H* (DVH).

values that an individual may obtain. Often, in this case, the optimal genotype is heterozygous for one or more loci. A classic example of stabilizing selection is the human birth weight. Babies of low weight have impaired thermoregulation and are more susceptible to infectious disease, whereas babies of large body weight are more difficult to deliver. Directional selection can be modeled by assuming that the optimum is more extreme than the genotypic values that the individual may obtain. Therefore, the allele frequencies shift towards the direction of fixation of the most extreme genotype favored by selection.

Previous studies (KARLIN and FELDMAN 1970; BODMER and FELSENSTEIN 1967) suggest that multiple equilibrium points exist in two-locus two-allele models with a Gaussian fitness function. Furthermore, conditions are provided for their existence and stability. However, the trajectories of the alleles towards the equilibrium points have not been explored. This study focuses on the trajectory of an allele, which initially is in low frequency and at its equilibrium points. In agreement with WILLENSDORFER and BÜRGER (2003) multiple equilibrium points in [0, 1] have been illustrated in this study depending on the initial values of the model parameters.

An important result of this study shows that selective sweeps that initiate from very low frequency of A_{11} allele are very rare in the two-locus two-allele model with symmetrical fitness matrix. Multiple conditions need to be satisfied in order to achieve fixation. First, the contribution of one of the alleles in the second locus (*e.g.* A_{21}) should be approximately equal to the contribution of the A_{11} allele. Second, the frequency of the A_{21} allele needs to be very low. In this regime the population is initially dominated by the A_{12} and A_{22} alleles which drive the population far from its optimum value (since they have similar contributions due to the symmetry of the model). Thus, the A_{11} competes with the A_{21} allele; since their contribution is similar their initial frequencies may determine the fate of the trajectory of the A_{11} allele. In fact this result suggests that in the two-allele two-locus model a selective sweep becomes possible when the second locus is nearly monomorphic, *i.e.* when the model resembles the one-locus two-allele model. Since fixation of A_{11} becomes more probable as the value of initial frequency increases, a model of sweeps from standing genetic variation may be more suitable.

Relaxing the assumption for the symmetry of the fitness matrix, we show that the fixation of the A_{11} allele becomes more likely. This is because the optimum genotype does not correspond necessarily to the heterozygote states. Thus, when the fitness matrix is generalized, the fitness may be either stabilizing or directional. For example, if the contributions of the A_{11}, A_{12}, A_{21}, and A_{22} alleles are -1, 1, 1, 1, and the optimum genotypic value is at 0, then the A_{11} allele is clearly beneficial: the second locus contributes $+2$ to the genotypic value, and only the $A_{11}A_{11}$ genotype may bring the population to the optimum by contributing -2. Figure 3.12 shows that in

the generalized fitness model a larger proportion of trajectories result in fixation compared to the symmetrical fitness model.

Assuming an effective population size $N = 10000$ we explore the effects of random genetic drift. Fluctuation of the frequencies of the trajectories are limited because the population size is not too small. However, genetic drift increases the proportion of the trajectories that reach monomorphic states (Figure 3.9). This is expected because genetic drift pushes the model towards its absorbing states. Therefore, selection needs to be strong enough in order to maintain the polymorphic state of the trajectory. This is illustrated clearly in Figure 3.10, where the ω^2 value is small for the vast majority of trajectories that are polymorphic at equilibrium.

When more than two loci are modeled then the proportion of trajectories that reach fixation increases. This is in agreement with the results of BÜRGER (2000), who shows that when the trait is determined by more than four loci then the monomorphic equilibrium points become more likely. Intuitively, the proportion of trajectories that reach a monomorphic equilibrium state increases because the optimum may be reached not only by the heterozygote genotypes but also by various combinations of homozygotes. Assume the five-locus model with symmetrical fitness matrix, and further assume that the contributions of the A_{i1} alleles are 2, 3, 1, 2, 2. Then, multiple configuration may bring the population to the optimum. For example, the homozygotes for the genotype $A_{11}A_{21}A_{32}A_{42}A_{52}$ is at the optimum. The same is true for the $A_{12}A_{22}A_{31}A_{41}A_{51}$. Of course, the five-locus heterozygote is at the optimum as well. Thus, the population can reach its optimum by fixing appropriate combinations of alleles. Additionally, Figure 3.16 shows that fixation of the A_{11} allele depends on the initial frequency of the A_{11}. However, this dependency is not as strong as in the two-locus model. Therefore, fixation of the 11 allele is possible even for small initial frequencies as it is required for classical selective sweeps.

Conditioning on the trajectory of the A_{11} allele, coalescent simulations have been implemented. As mentioned previously, this is correct only for weak selection and when the loci are not linked. However, a first approximation is useful in order study the genealogy properties and the patterns of neutral polymorphism around the A_1 locus. When the A_{11} allele fixes in the population, then the genealogies around the A_1 locus are similar to the classical selective sweep (given that the initial frequency of the A_{11} is small). The coalescent trees are on average imbalanced and short in the proximity of the A_1, as expected in a classical selective sweep model. The imbalance becomes larger for a certain genomic region as we move away from the focal locus, and then it reverts to neutral levels. The length of the region where the signature of selection is visible depends on the recombination fraction. For the set of simulations in this article, the recombination fraction is small. Thus the genomic regions where the signature of selection appears is large (> 250 kb).

When the trajectories do not reach fixation, then a part or all of the signatures of a selective sweep become invisible, depending on the equilibrium frequency of the trajectory. For example, when the equilibrium frequency is between 0.9 and 1, then the height of the coalescent tree equals the neutral expectations, because ancestral alleles (A_{12}) exist in the present day sample. For smaller equilibrium frequencies (*e.g.* 0.3 to 0.4) both the coalescent summaries and the polymorphism summaries resemble the neutral expectations.

Depending on the simulation parameters a large number of trajectories is maintained at some equilibrium value and does not reach fixation. For those trajectories analysis of incomplete sweeps (VOIGHT *et al.* 2006; TANG *et al.* 2007; SABETI *et al.* 2002) may be useful. There is, however, an essential difference between incomplete sweeps and sweeps in multi-locus models that were studied in this article. Incomplete sweeps are on the way to fixation, however the sweeps studied here remain at equilibrium frequency. Therefore, the signatures of selection will be visible only in the cases that the equilibrium frequency has been achieved recently. If the trajectory remained at the equilibrium level for too long, then the signatures of selection will fade away due to recombination.

The results indicate that detection of selection from polymorphism patterns in multi-locus models may be hard. When the focal allele (A_{11}) fixes in the population, then the statistical tools that are used to detect sweeps in one-locus two-allele models may be useful (*e.g.* KIM and STEPHAN 2002; NIELSEN *et al.* 2005; PAVLIDIS *et al.* 2010). Also, this is true for trajectories close to fixation. Even if the patterns appear to be different than those of fixed trajectories (Figure 3.19), the direction of perturbations is similar to the classical sweep models, and therefore the same statistical tools may be used. However, for smaller equilibrium frequencies some or all the signatures of selection studied in this article disappear.

In multi-locus two-allele models a class of trajectories that is absent from one-locus two-allele models comprises of non-monotonic trajectories. These trajectories approach fast a certain frequency, but eventually they decline either to extinction or to some other frequency. The difference between the maximum frequency and the equilibrium frequency may be quite large. In the simulated datasets, we observed differences even larger than 0.5. However, the polymorphism patterns and the coalescent patterns seem to be very similar to the neutral expectations. Thus, those trajectories may be completely invisible using the summary statistics studied in this paper. Summarizing the results, it may be claimed that the statistical tools that have been developed to detect selective sweeps may detect a small proportion of the multi-locus selection cases: only those cases that result in fixed trajectories or equilibrium trajectories close to fixation. Tools that are used for detecting incomplete sweeps may be useful when the trajectory has reached its equilibrium frequency very recently. For trajectories that have reached their equilibrium frequency further in the past, we

expect that recombination will destroy the signatures of selection. In fact, the results imply that positive or stabilizing selection may occur in a much higher rate than previous studies which analyze selective sweeps report (*e.g.* LI and STEPHAN 2006). However, the majority of the cases remains undetectable since both the coalescent trees (as summarized here) and the polymorphism summary statistic do not deviate from neutrality.

To our knowledge the only study of selective sweeps in quantitative traits was done by CHEVIN and HOSPITAL (2008). CHEVIN and HOSPITAL (2008) assume an infinite number of unlinked and independent loci that control the trait. Moreover they assume that the variability in the genetic background remains constant during the trajectory of the new allele, and that the effect of the focal locus on the trait value is small compared to the effect of the genetic background. These assumptions enable them to solve the trajectory of a new allele analytically for linear, exponential, and Gaussian fitness functions. Furthermore, CHEVIN and HOSPITAL (2008) focus mainly on the trajectories that reach fixation. On the other hand the present article focuses on finite locus models. Considering a finite number of loci makes the model intractable mathematically. Therefore, computer simulations were employed to study the trajectory of a new allele. The contribution of alleles may be arbitrary as well as the recombination fraction between the loci. In this article we provide information about the role of various parameters on the fixation of the trajectories, but also we study extensively the trajectories that remain polymorphic. Polymorphic trajectories are possibly absent in CHEVIN and HOSPITAL (2008) due to the large number of loci that control the trait. The results of CHEVIN and HOSPITAL (2008) indicate that trajectories of new alleles evolve slightly slower than classical selective sweeps, and selective sweeps of their model look slightly older than the classical one-locus selective sweep. This is true for the multi-locus model studied here as well, however to a less extent (results not shown). The present study may be considered complementary to the study of CHEVIN and HOSPITAL (2008) for finite multi-locus models, providing information about the trajectories of new alleles and the polymorphism patterns generated by selective sweeps in multi-locus models. This information is essential for the development of software which will be able to detect selective sweeps in multi-locus models.

General Discussion

The availability of large scale data for population genetics studies has offered a possibility for the precise identification of the footprints of hitchhiking events in the genome. Until now this kind of analysis has been only performed in a few model species. However, advancing DNA sequencing technology will allow the generation of population genetic data also in non-model organisms. The accumulation of such data in a multitude of species and populations allows us to address such questions as (i) are some genes that are involved in certain functions more subjected to adaptive evolution than others? (ii) is positive selection more frequent in populations that have to adapt to the new environment? and (iii) what is the rate at which adaptive substitutions occur? Simultaneously, it becomes apparent that statistical methods based on the classical selective sweep model may be inadequate to capture more complicated selective events.

The genes that were identified by selection mapping in natural populations of *D. melanogaster*, mice and plants appear to fall into three functional categories: genes of sensory pathways (*i.e.* genes involved in the development of the eye, skin or hairs), genes determining body size, and defense/immunity genes. Although the number of genes detected so far is small, the emerging pattern confirms the working hypothesis that most genes identified on the basis of selective sweeps play a role in ecological adaptation. Among these genes only *ph-p* (encoding polyhomeotic proximal protein, a part of a universal transcription repressor Polycomb group) does not have a specific function related to the environment. On the other hand, it is remarkable that genes involved in temperature adaptation and energy metabolism have not yet been identified in flies by the selective sweep method. For the genes that experienced positive selection in human, additional categories and sub-categories can be defined. For example, the genes responding to the selection pressures during the transition to novel food sources with the advent of agriculture form a new category (including LCT). Furthermore, olfactory and pigmentation genes are important sub-categories of the genes involved in sensory perception (NIELSEN *et al.* 2007). It should be noted, however, that the identification of a specific gene or function might not always be possible. There is accumulating evidence that selection also affects non-coding portions of the genome (*e.g.* ANDOLFATTO 2005;

BUSH and LAHN 2005). As the biological role of these non-coding regions is still poorly understood, the assessment of the functional consequences of positive selection on such loci poses an additional challenge.

In both flies and humans the signatures of selection are to some extent population-specific and thus suggestive of local adaptation. VOIGHT *et al.* (2006) found the strongest signals of selection in human populations in Africa(Yoruba). WILLIAMSON *et al.* (2007), however, detected more evidence for sweeps in Chinese and European-American populations than in the African-American population. These contradictory results may be due to the fact that the power to detect selective sweeps is lower in the African-American sample. In *D. melanogaster*, in five of the six cases discussed above both African and non-African samples were analyzed and in four of them the sweep originated in Africa. This result is not consistent with the hypothesis that the novel environments encountered by flies imposed new selective pressures, which in turn led to an increased rate of local selective sweeps. Whether this result is a consequence of a lack of power is unclear at present. Nonetheless, it is consistent with the analysis of LI and STEPHAN (2006) who found no difference in the rate of adaptive substitution between African and European populations in an X chromosome wide analysis. This issue needs to be revisited as soon as more data are available.

The estimated rates of adaptive substitutions obtained by LI and STEPHAN (2006) agree surprisingly well with earlier estimates based on DNA sequence divergence between *D. simulans* and *D. yakuba* (SMITH and EYRE-WALKER 2002). However, the latter study estimates the rate of adaptive substitutions over a long time period and also takes weak selection into account. As LI and STEPHAN (2006) only estimate the rates of relatively young and strong selection events this might indicate an acceleration of adaptive evolution in recent times.

This study aimed to make a contribution to solving the above-mentioned general questions of molecular evolution. The goals of this study were: (i) implementation of a method that is able to detect accurately selective sweeps in natural populations that have experienced past demographic changes; (ii) application of the methods to real data; (iii) extension of selective sweeps in multi-locus models. To achieve these goals, first, the combination of two algorithms, the *SweepFinder* and the ω-statistic that use SFS and LD information, respectively, was applied to disentangle selective sweeps from neutrality. Then, the *SweepFinder* algorithm and the *CLR* test were applied on the subgenomic region of African population of *D. melanogaster* that included the *HDAC6* gene. Finally, we studied selective sweeps in multi-locus models assuming a finite number of loci that control the trait. Regarding the functional characterization of genes that are involved in adaptive evolution, Chapter 1 provides a modification of the *SweepFinder* algorithm that is able to detect the target of selection accurately (median distance from the target is < 1 kb). Then, coupling the

selective sweep mapping with gene ontology analysis enables the functional characterization of the targets of positive selection. Furthermore, Chapter 1 describes a machine learning approach which outperforms current methods and is able to detect selective sweeps in non-equilibrium populations. These populations have experienced significant demographic changes in the recent past such as population bottlenecks or founder events. Detecting selective sweeps in such populations gives insight into the adaptation of populations in new environments. The third question aims at studying the rate of adaptive substitutions in the genome. In Chapter 3 I show that if selection operates on multi-locus models then adaptive substitutions do not always occur; the population may remain polymorphic for one or more loci. Furthermore, the term 'adaptive substitution' obtains a relative meaning in multi-locus models. A certain substitution can be adaptive only in a specific genetic background, whereas it may be deleterious in other genetic backgrounds.

Several approaches and findings that are presented in the thesis show aspects of novelty. To begin with, machine learning methods are introduced for first time in the population genetics field. The machine learning framework uses information from both the neutral datasets and the datasets with selection. This increases the power of detecting selective sweeps and, importantly, reveals the demographic scenarios when the separation between selection and demography is not possible. Furthermore, we developed modifications of existing algorithms so that the precision of the algorithm to detect the location of a beneficial mutation increases under certain demographic regimes. Secondly, we used a variable significance threshold for the neutrality tests because our analysis showed that the values of the tests can be biased at the edges of the subgenomic region. Additionally, we implemented an ascertainment bias correction when more than one populations are involved in the initial choice of the subgenomic region. Particularly in Chapter 2, the initial choice of the subgenomic region was based on a previous analysis of the European population of *D. melanogaster* (LI and STEPHAN 2006). We introduced a simulation approach that corrects this kind of ascertainment bias. Finally, we study selective sweeps in multi-locus models. We demonstrated that selective events in multi-locus models may remain undetectable using current approaches. To our knowledge the only study of selective sweeps in quantitative traits was done by CHEVIN and HOSPITAL (2008). However, their approach assumes that the variability in the genetic background remains constant and that the number of loci in the genetic background is infinite. In contrast, our study assumes a finite number of loci that control the trait. This allows to relax the assumption of CHEVIN and HOSPITAL (2008) regarding the variability of the genetic background. Furthermore, we analyze the coalescent trees by implementing two summary statistics which measure the imbalance of genealogies in a genomic region. We place emphasis on the trajectories that reach a polymorphic equilibrium, because these trajectories are absent in classical

selective sweep models. Results show that the detection of selective sweeps in multi-locus models is challenging unless the trajectory of the focal allele goes to fixation. In summary, our results provide information about the trajectories of new alleles and the polymorphism patterns generated by selective sweeps in multi-locus models. This information is essential for the development of software which will be able to detect selective sweeps in multi-locus models.

The ability to map target genes of selection is of a great practical importance since it may open up new opportunities for studying adaptation and understand genetic diseases and mechanisms of immunity in humans. However, in order to make progress in these directions it is important to perform functional analysis of the genes under selection. Functions of many of the genes identified by selective sweep mapping are not clear. In most of the cases we have only a vague notion of which allele was under recent selection and why. Additional studies, such as QTL analysis, gene regulatory network and pathway analysis, that will relate the selection mapping to specific phenotypes are important research directions for the future.

Bibliography

AKEY, J. M., 2009 Constructing genomic maps of positive selection in humans: where do we go from here? Genome Research **19**: 711–722.

AKEY, J. M., M. A. EBERLE, M. J. RIEDER, C. S. CARLSON, M. D. SHRIVER, *et al.*, 2004 Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biology **2**: e286.

ANDOLFATTO, P., 2005 Adaptive evolution of non-coding DNA in *Drosophila*. Nature **437**: 1149–1152.

ANDOLFATTO, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Research **17**: 1755–1762.

BALDING, D. J., and R. A. NICHOLS, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica **96**: 3–12.

BARTON, N., 1998 The effect of hitch-hiking on neutral genealogies. Genetical Research **72**: 123–133.

BEAUMONT, M. A., 2003 Estimation of population growth or decline in genetically monitored populations. Genetics **164**: 1139–1160.

BEAUMONT, M. A., and D. J. BALDING, 2004 Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology **13**: 969–980.

BEISSWANGER, S., and W. STEPHAN, 2008 Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated *polyhomeotic* genes in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America **105**: 5447–5452.

BEISSWANGER, S., W. STEPHAN, and D. DE LORENZO, 2006 Evidence for a selective sweep in the wapl region of *Drosophila melanogaster*. Genetics **172**: 265–274.

BODMER, W. F., and J. FELSENSTEIN, 1967 Linkage and selection: theoretical analysis of the deterministic two locus random mating model. Genetics **57**: 237–265.

BOYAULT, C., B. GILQUIN, Y. ZHANG, V. RYBIN, E. GARMAN, *et al.*, 2006 HDAC6-p97/VCP controlled polyubiquitin chain turnover. The EMBO Journal **25**: 3357–3366.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY, and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140**: 783–796.

BULMER, M. G., 1973 The maintenance of the genetic variability of polygenic characters by heterozygous advantage. Genetical Research **22**: 9–12.

BÜRGER, R., 1998 Mathematical properties of mutation-selection models. Genetica **102**: 279–298.

BÜRGER, R., 2000 *The mathematical theory of selection, recombination, and mutation*. John Wiley, Hoboken.

BÜRGER, R., 2002 On a genetic model of intraspecific competition and stabilizing selection. American Naturalist **160**: 661–682.

BÜRGER, R., and A. GIMELFARB, 1999 Genetic variation maintained in multilocus models of additive quantitative traits under stabilizing selection. Genetics **152**: 807–820.

BÜRGER, R., and A. GIMELFARB, 2004 The effects of intraspecific competition and stabilizing selection on a polygenic trait. Genetics **167**: 1425–1443.

BUSH, E. C., and B. T. LAHN, 2005 Selective constraint on noncoding regions of hominid genomes. PLoS Computational Biology **1**: e73.

CHADEAU-HYAM, M., C. J. HOGGART, P. F. O'REILLY, J. C. WHITTAKER, M. D. IORIO, *et al.*, 2008 Fregene: simulation of realistic sequence-level data in populations and ascertained samples. BMC Bioinformatics **9**: 364.

CHEVIN, L.-M., S. BILLIARD, and F. HOSPITAL, 2008 Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation. Genetics **180**: 301–316.

CHEVIN, L.-M., and F. HOSPITAL, 2008 Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics **180**: 1645–1660.

CHINTAPALLI, V. R., J. WANG, and J. A. T. DOW, 2007 Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nature Genetics **39**: 715–720.

CLEVELAND, W., 1979 Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association **74**: 829–836.

COMERON, J. M., M. KREITMAN, and M. AGUAD, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151**: 239–249.

COOP, G., and R. C. GRIFFITHS, 2004 Ancestral inference on gene trees under selection. Theoretical Population Biology **66**: 219–232.

CROW, J., and M. KIMURA, 1970 *An Introduction to Population Genetics Theory*. Burgess Publishing Company, Minneapolis, Minneapolis, Minnesota.

DEPAULIS, F., and M. VEUILLE, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular Biology and Evolution **15**: 1788–1790.

DOW, J. A. T., and S. A. DAVIES, 2006 The Malpighian tubule: rapid insights from post-genomic biology. Journal of Insect Physiology **52**: 365–378.

DUDA, R. O., P. E. HART, and D. G. STORK, 2000 *Pattern Classification (2nd Edition)*. Wiley-Interscience, Hoboken.

ENDLER, J., 1986 *Natural selection in the wild*. Princeton University Press, Princeton.

EWENS, W. J., 2004 *Mathematical Population Genetics I. Theoretical Introduction*. Springer-Verlag, New York, New York, 2nd edition.

EWING, G., and J. HERMISSON, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064–2065.

EXCOFFIER, L., A. ESTOUP, and J.-M. CORNUET, 2005 Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics **169**: 1727–1738.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155**: 1405–1413.

FISHER, R., 1930 *The genetical theory of natural selection*. Clarendon Press, Oxford, UK.

FU, Y. X., 1995 Statistical properties of segregating sites. Theoretical Population Biology **48**: 172–197.

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

FUREY, T. S., N. CRISTIANINI, N. DUFFY, D. W. BEDNARSKI, M. SCHUMMER, *et al.*, 2000 Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16**: 906–914.

GAVRILETS, S., and A. HASTINGS, 1994 Maintenance of multilocus variability under strong stabilizing selection. Journal of Mathematical Biology **32**: 287–302.

GERHARDT, H., 1994 The evolution of vocalization in frogs and toads. Annual Review of Ecology and Systematics **25**: 293–324.

GILLESPIE, J. H., 1984 Pleiotropic overdominance and the maintenance of genetic variation in polygenic characters. Genetics **107**: 321–330.

GILLESPIE, J. H., and M. TURELLI, 1989 Genotype-environment interactions and the maintenance of polygenic variation. Genetics **121**: 129–138.

GLINKA, S., L. OMETTO, S. MOUSSET, W. STEPHAN, and D. DE LORENZO, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165**: 1269–1278.

HAHN, M. W., 2008 Toward a selection theory of molecular evolution. Evolution: International Journal of Organic Evolution **62**: 255–265.

HAN, J., and M. KAMBER, 2000 *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, San Francisco.

HARR, B., M. KAUER, and C. SCHLÖTTERER, 2002 Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences of the United States of America **99**: 12949–12954.

HASTIE, T., R. TIBSHIRANI, and J. H. FRIEDMAN, 2001 *The Elements of Statistical Learning*. Springer, New York.

HERMISSON, J., and P. S. PENNINGS, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics **169**: 2335–2352.

HERNANDEZ, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. Bioinformatics **24**: 2786–2787.

HEY, J., and R. NIELSEN, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the National Academy of Sciences of the United States of America **104**: 2785–2790.

HOGGART, C. J., M. CHADEAU-HYAM, T. G. CLARK, R. LAMPARIELLO, J. C. WHITTAKER, *et al.*, 2007 Sequence-level population simulations over large genomic regions. Genetics **177**: 1725–1731.

HUDSON, R., 1990 Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology **7**: 44.

HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

HUDSON, R. R., M. KREITMAN, and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116**: 153–159.

HUTTER, S., H. LI, S. BEISSWANGER, D. DE LORENZO, and W. STEPHAN, 2007 Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics **177**: 469–480.

INTERNATIONAL HAPMAP CONSORTIUM, 2003 The International HapMap Project. Nature **426**: 789–796.

INTERNATIONAL HAPMAP CONSORTIUM, 2005 A haplotype map of the human genome. Nature **437**: 1299–1320.

INTERNATIONAL HAPMAP CONSORTIUM, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature **449**: 851–861.

Drosophila 12 GENOMES CONSORTIUM, 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. Nature **450**: 203–218.

JENSEN, J. D., V. L. BAUER DUMONT, A. B. ASHMORE, A. GUTIERREZ, and C. F. AQUADRO, 2007a Patterns of sequence variability and divergence at the diminutive gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. Genetics **177**: 1071–1085.

JENSEN, J. D., Y. KIM, V. BAUER DUMONT, C. F. AQUADRO, and C. D. BUSTAMANTE, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics **170**: 1401–1410.

JENSEN, J. D., K. R. THORNTON, and P. ANDOLFATTO, 2008 An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. PLoS Genetics **4**: e1000198.

JENSEN, J. D., K. R. THORNTON, C. D. BUSTAMANTE, and C. F. AQUADRO, 2007b On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics **176**: 2371–2379.

KAPLAN, N. L., R. R. HUDSON, and C. H. LANGLEY, 1989 The hitchhiking effect revisited. Genetics **123**: 887–899.

KARLIN, S., and M. W. FELDMAN, 1970 Linkage and selection: two locus symmetric viability model. Theoretical Population Biology **1**: 39–71.

KAUER, M. O., D. DIERINGER, and C. SCHLÖTTERER, 2003 A microsatellite variability screen for positive selection associated with the out of Africa habitat expansion of *Drosophila melanogaster*. Genetics **165**: 1137–1148.

KELLY, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146**: 1197–1206.

KHOCHBIN, S., A. VERDEL, C. LEMERCIER, and D. SEIGNEURIN-BERNY, 2001 Functional significance of histone deacetylase diversity. Current Opinion in Genetics & Development **11**: 162–166.

KIM, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics **172**: 1967–1978.

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.

KINGMAN, J., 1982 The coalescent. Stochastic processes and their applications **13**: 235–248.

KIRBY, D. A., and W. STEPHAN, 1996 Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics **144**: 635–645.

LANDE, R., 1975 The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genetical Research **26**: 221–235.

LANDE, R., 1983 The response to selection on major and minor mutations affecting a metrical trait. Heredity **50**: 47–65.

LANGLEY, C. H., Y. N. TOBARI, and K. I. KOJIMA, 1974 Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics **78**: 921–936.

LI, H., and W. STEPHAN, 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics **171**: 377–384.

LI, H., and W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genetics **2**: e166.

LIN, K., H. LI, C. SCHLÖTTERER, and A. FUTSCHIK, 2010 Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics : Epub ahead of print.

LUDWIG, M. Z., C. BERGMAN, N. H. PATEL, and M. KREITMAN, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**: 564–567.

MACPHERSON, J. M., G. SELLA, J. C. DAVIS, and D. A. PETROV, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics **177**: 2083–2099.

MATTHIAS, P., M. YOSHIDA, and S. KHOCHBIN, 2008 HDAC6 a new cellular stress surveillance factor. Cell Cycle **7**: 7–10.

MAURICIO, R., and M. RAUSHER, 1997 Experimental manipulation of putative selective agents provides evidence for the role of natural enemies in the evolution of plant defense. Evolution **51**: 1435–1444.

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

MCVEAN, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics **175**: 1395–1406.

MYERS, S., C. FEFFERMAN, and N. PATTERSON, 2008 Can one learn history from the allelic spectrum? Theoretical Population Biology **73**: 342–348.

NAIR, S., D. NASH, D. SUDIMACK, A. JAIDEE, M. BARENDS, *et al.*, 2007 Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Molecular Biology and Evolution **24**: 562–573.

NIELSEN, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154**: 931–942.

NIELSEN, R., I. HELLMANN, M. HUBISZ, C. BUSTAMANTE, and A. G. CLARK, 2007 Recent and ongoing selection in the human genome. Nature Reviews Genetics **8**: 857–868.

NIELSEN, R., M. J. HUBISZ, I. HELLMANN, D. TORGERSON, A. M. ANDRÉS, *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. Genome Research **19**: 838–849.

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

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

ORENGO, D. J., and M. AGUADÉ, 2004 Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics **167**: 1759–1766.

ORENGO, D. J., and M. AGUADÉ, 2007 Genome scans of variation and adaptive change: extended analysis of a candidate locus close to the phantom gene region in *Drosophila melanogaster*. Molecular Biology and Evolution **24**: 1122–1129.

ORENGO, D. J., and M. AGUADÉ, 2010 Uncovering the footprint of positive selection on the X chromosome of *Drosophila melanogaster*. Molecular Biology and Evolution **27**: 153–160.

PANDEY, U. B., Z. NIE, Y. BATLEVI, B. A. MCCRAY, G. P. RITSON, *et al.*, 2007 HDAC6 rescues neurodegeneration and provides an essential link between autophagy and the UPS. Nature **447**: 859–863.

PAVLIDIS, P., S. HUTTER, and W. STEPHAN, 2008 A population genomic approach to map recent positive selection in model species. Molecular Ecology **17**: 3585–3598.

PAVLIDIS, P., J. D. JENSEN, and W. STEPHAN, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics **185**: 907–922.

PERRETT, D. I., K. A. MAY, and S. YOSHIKAWA, 1994 Facial shape and judgements of female attractiveness. Nature **368**: 239–242.

PFAFFELHUBER, P., A. LEHNERT, and W. STEPHAN, 2008 Linkage disequilibrium under genetic hitchhiking in finite populations. Genetics **179**: 527–537.

PICKRELL, J. K., G. COOP, J. NOVEMBRE, S. KUDARAVALLI, J. Z. LI, *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Research **19**: 826–837.

POOL, J. E., V. BAUER DUMONT, J. L. MUELLER, and C. F. AQUADRO, 2006 A scan of molecular variation leads to the narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. Genetics **172**: 1093–1105.

PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160**: 1179–1189.

PRZEWORSKI, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164**: 1667–1676.

RAMOS-ONSINS, S. E., S. MOUSSET, T. MITCHELL-OLDS, and W. STEPHAN, 2007 Population genetic inference using a fixed number of segregating sites: a reassessment. Genetical Research **89**: 231–244.

RIEBLER, A., L. HELD, and W. STEPHAN, 2008 Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics **178**: 1817–1829.

ROBERTSON, A., 1956 The effect of selection against extreme deviants based on deviation or on homozygosis. Journal of Genetics **54**: 236–248. 10.1007/BF02982779.

ROZAS, J., J. C. SNCHEZ-DELBARRIO, X. MESSEGUER, and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19**: 2496–2497.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER, *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419**: 832–837.

SABETI, P. C., P. VARILLY, B. FRY, J. LOHMUELLER, E. HOSTETTER, *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449**: 913–918.

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

SMITH, N. G., and A. EYRE-WALKER, 2002 Adaptive protein evolution in *Drosophila*. Nature **415**: 1022–1024.

SPENCER, C. C. A., and G. COOP, 2004 SelSim: a program to simulate population genetic data with natural selection and recombination. Bioinformatics **20**: 3673–3675.

STEPHAN, W., 1995 An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. Molecular Biology and Evolution **12**: 959–962.

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. H. E. WIEHE, and M. W. LENZ, 1992 The effect of strongly selected substitutions on neutral polymorphism - Analytical results based on diffusion theory. Theoretical Population Biology **41**: 237–254.

SVETEC, N., P. PAVLIDIS, and W. STEPHAN, 2009 Recent strong positive selection on *Drosophila melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by population genomic analysis. Molecular Biology and Evolution **26**: 1549–1556.

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

TANG, K., K. R. THORNTON, and M. STONEKING, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biology **5**: e171.

TESHIMA, K. M., and H. INNAN, 2009 mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. BMC Bioinformatics **10**: 166.

THORNTON, K., and P. ANDOLFATTO, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics **172**: 1607–1619.

THORNTON, K. R., and J. D. JENSEN, 2007 Controlling the false-positive rate in multilocus genome scans for selection. Genetics **175**: 737–750.

TUFTO, J., 2000 Quantitative genetic models for the balance between migration and stabilizing selection. Genetical Research **76**: 285–293.

TURELLI, M., 1984 Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theoretical Population Biology **25**: 138–193.

VAPNIK, V. N., 1995 *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

VERDEL, A., S. CURTET, M. P. BROCARD, S. ROUSSEAUX, C. LEMERCIER, *et al.*, 2000 Active maintenance of mHDA2/mHDAC6 histone-deacetylase in the cytoplasm. Current Biology **10**: 747–749.

VIGNIERI, S. N., J. G. LARSON, and H. E. HOEKSTRA, 2010 The selective advantage of crypsis in mice. Evolution **64**: 2153–2158.

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**: e72.

WAKELEY, J., 2008 *Coalescent theory*. Roberts & Co. Publishers, Greenwood Village, Colorado.

WIEHE, T. H. E., and W. STEPHAN, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Molecular Biology and Evolution **10**: 842–854.

WILLENSDORFER, M., and R. BÜRGER, 2003 The two-locus model of Gaussian stabilizing selection. Theoretical Population Biology **64**: 101–117.

WILLIAMSON, S. H., M. J. HUBISZ, A. G. CLARK, B. A. PAYSEUR, C. D. BUSTAMANTE, *et al.*, 2007 Localizing recent adaptive evolution in the human genome. PLoS Genetics **3**: e90.

YANG, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution **24**: 1586–1591.

ZHAO, Z., N. YU, Y.-X. FU, and W.-H. LI, 2006 Nucleotide variation and haplotype diversity in a 10-kb noncoding region in three continental human populations. Genetics **174**: 399–409.

ZIVKOVIC, D., and T. WIEHE, 2008 Second-order moments of segregating sites under variable population size. Genetics **180**: 341–357.

Appendix

Table S1

# SNP	$i-1$	i	$i+1$
...			
$i-1$	-	$Z_{i-1,i}$	$Z_{i-1,i}+Z_{i,i+1}$
i	-	-	$Z_{i,i+1}$
$i+1$	-	-	-
...			

TABLE S1: The matrix used for the pre-calculation of the ω-statistic for all possible configurations. A cell $Z_{i,j}$, $i<j$ represents the sum of all pairwise linkage disequilibrium comparisons (r^2) for the sites that belong to the window $[i,j]$. We have implemented a recursive algorithm in order to calculate this matrix. In detail, the calculation starts from the cell $Z_{i,i+1}$ *i.e.* the cells next to the main diagonal and proceeds upwards to the cell $Z_{i-1,i+1}$. Then $Z_{i-1,i+1}=Z_{i-1,i}+Z_{i,i+1}$. $Z_{i,i+1}=r^2_{i,i+1}$ and $Z_{i-1,i+1}$ has been calculated in the previous cycle. Then, using this matrix it is trivial to calculate the components of the ω-statistic for any configuration. When the left and right sub-regions are defined by $[i,k]$ and $[k+1,j]$, respectively, then the numerator is the sum $Z_{i,k}+Z_{k+1,j}$ weighted by the number of calculations $[\binom{k-i+1}{2}+\binom{j-k}{2}]^{-1}$, whereas the denominator is $Z_{i,j}-Z_{i,k}-Z_{k+1,j}$ weighted by $[(k-i+1)(j-k)]^{-1}$.

Figure S1

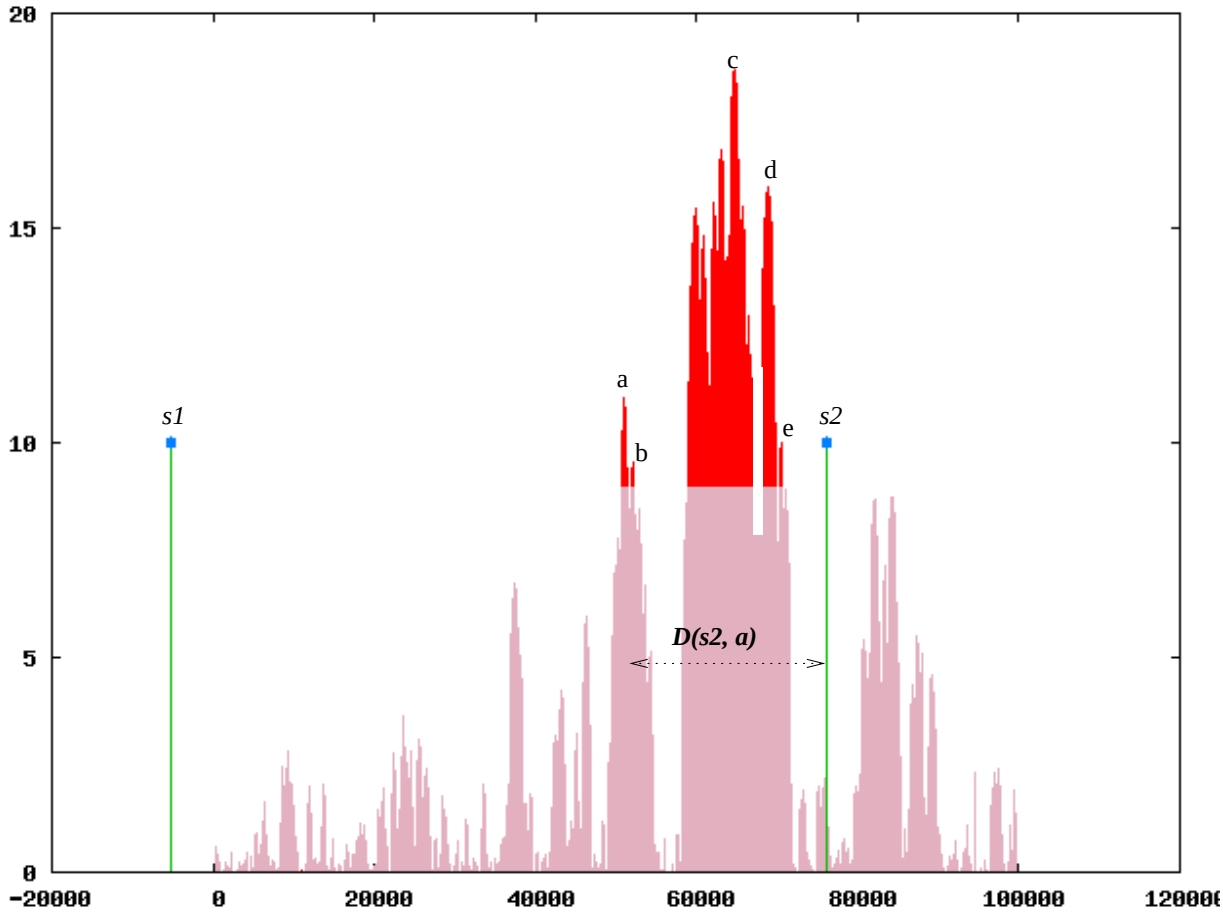

FIGURE S1: The distance between a peak of the landscape of the statistic and the selective sweep locations. In the history of the population two selective sweeps have occurred recently, at different time points and different locations on the chromosome. The selective sweep locations are illustrated as *s1* and *s2* (vertical lines). Given a user defined threshold, the landscape of the statistic is split in two regions, *i.e.* above and below the threshold. A peak is defined as the maximum point in an isolated (by the threshold) region. Thus, 5 peaks (a to e) have been formed. The distance $D(s2, a)$ of the 'a' peak measures the distance between this peak and *s2* which is the closest sweep location from this peak.

Figure S2

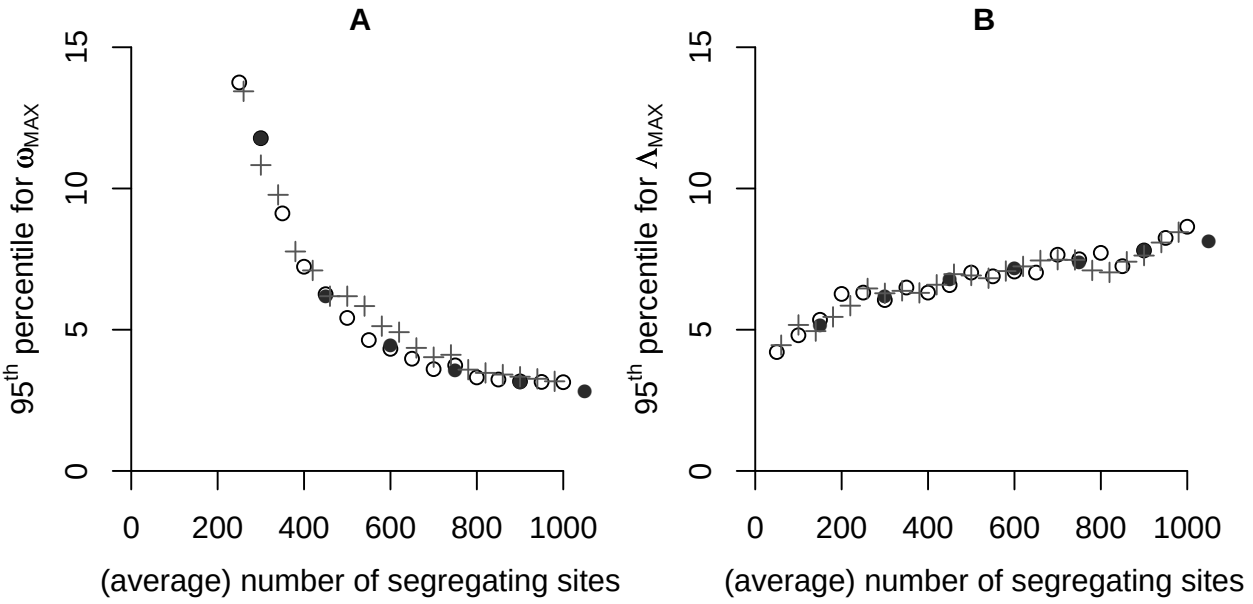

FIGURE S2: The 95^{th} percentile for A) the ω-statistic and B) *SweepFinder* based on the $F\theta$ (filled circles), the FS (open circles) and the $F\theta S$ procedures (crosses). Equilibrium neutrality simulations have been performed for a 50-kb genomic segment and 12 sequences ($h_n \approx 3$). Recombination rate is $0.05/\text{bp}$. For a given number of segregating sites (x-axis) simulations were performed by (i) fixing the number of segregating sites S_n (open circles), (ii) using $\theta_{\text{NEU}} = \theta_{\text{W}} = \frac{S_n}{h_n}$ (filled circles). In this case simulations generate on average S_n segregating sites. (iii) Under the $F\theta S$ process (crosses) we used the same $\theta_{\text{NEU}} = \theta_{\text{W}} = \frac{S_n}{h_n}$ but only the realizations that produced S_n segregating sites (see **Equilibrium selection versus equilibrium neutrality**).

Figure S3

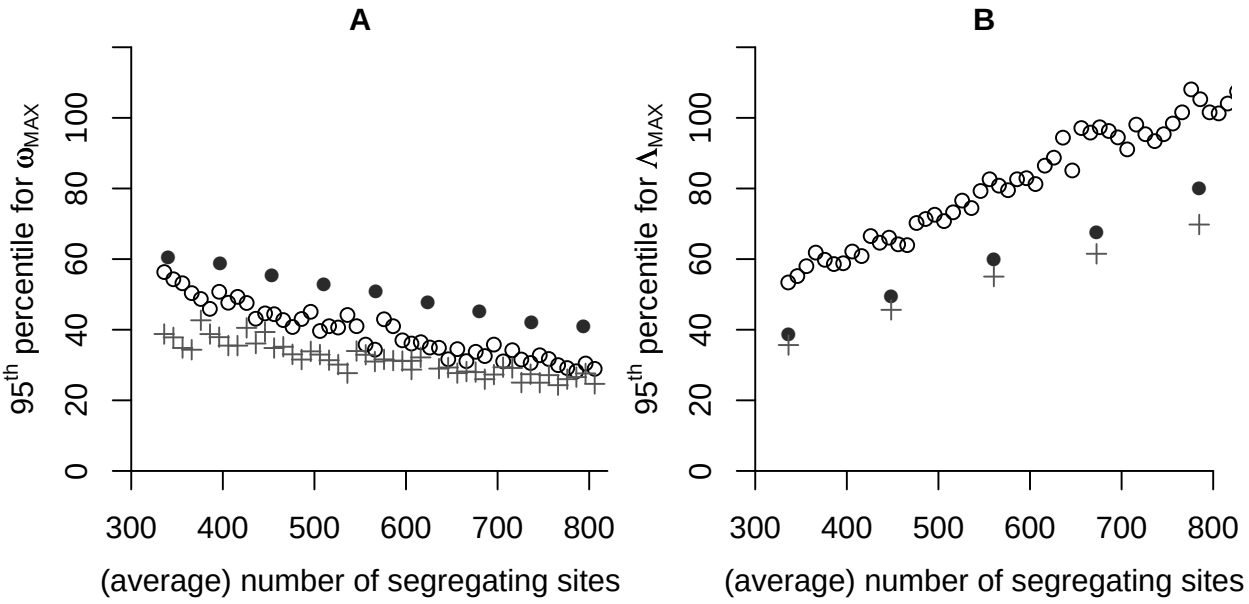

FIGURE S3: The 95^{th} percentile for A) the ω_{MAX} and B) Λ_{MAX} based on the $F\theta$ (full circles), the FS procedure (open circles) and the $F\theta S$ approach (crosses). Bottleneck simulations have been performed for a 50-kb genomic segment and 12 sequences ($h_n \approx 3$). We have used the demographic scenario inferred by LI and STEPHAN (2006) that describes the history of the European population of *D. melanogaster*. Recombination rate is $0.05/\mathrm{bp}$. For a given number of segregating sites (x-axis) simulations were performed by (i) fixing the number of segregating sites S_n (open circles), (ii) using $\theta_{\mathrm{NEU}} = \widehat{\theta} = \frac{2S_n}{E(T_c)}$, where $E(T_c)$ is the expected total length of the coalescent of n sequences (ZIVKOVIC and WIEHE 2008) (filled circles). In this case simulations generate on average S_n segregating sites. (iii) Under the $F\theta S$ process (crosses) we used the same $\theta_{\mathrm{NEU}} = \widehat{\theta}$, but only the realizations that produced S_n segregating sites (see **Equilibrium selection versus equilibrium neutrality**).

Figure S4

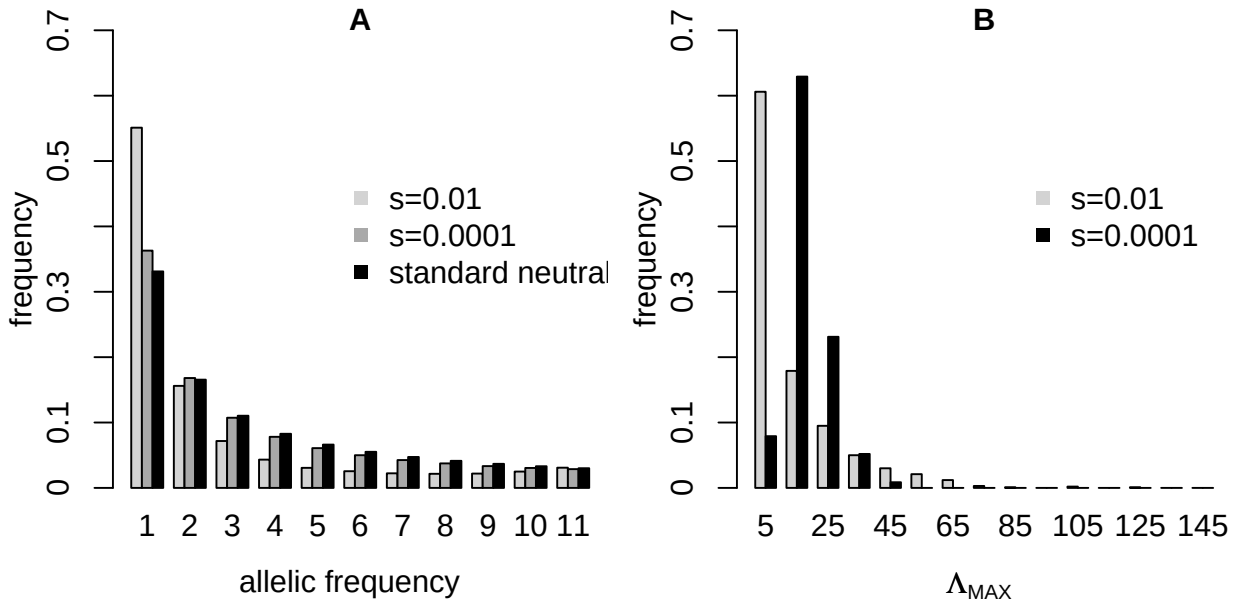

FIGURE S4: Comparisons between recurrent selective sweeps when $\frac{H_{RHH}}{H_{NEU}} = 0.25$ and $s = 0.01, 0.0001$. A) The SFS of the RHH model when $s = 0.0001$ is similar to that of the standard neutral SFS whereas a large excess of singletons appears when $s = 0.01$. B) When the SFS of the data itself is used in the *SweepFinder* calculations then the model with $s = 0.0001$ shows higher values of Λ_{MAX}. This is because the genomic regions affected by positive selection are smaller for smaller s values and a large fraction of the genome remains still unaffected by positive selection.

Figure S5

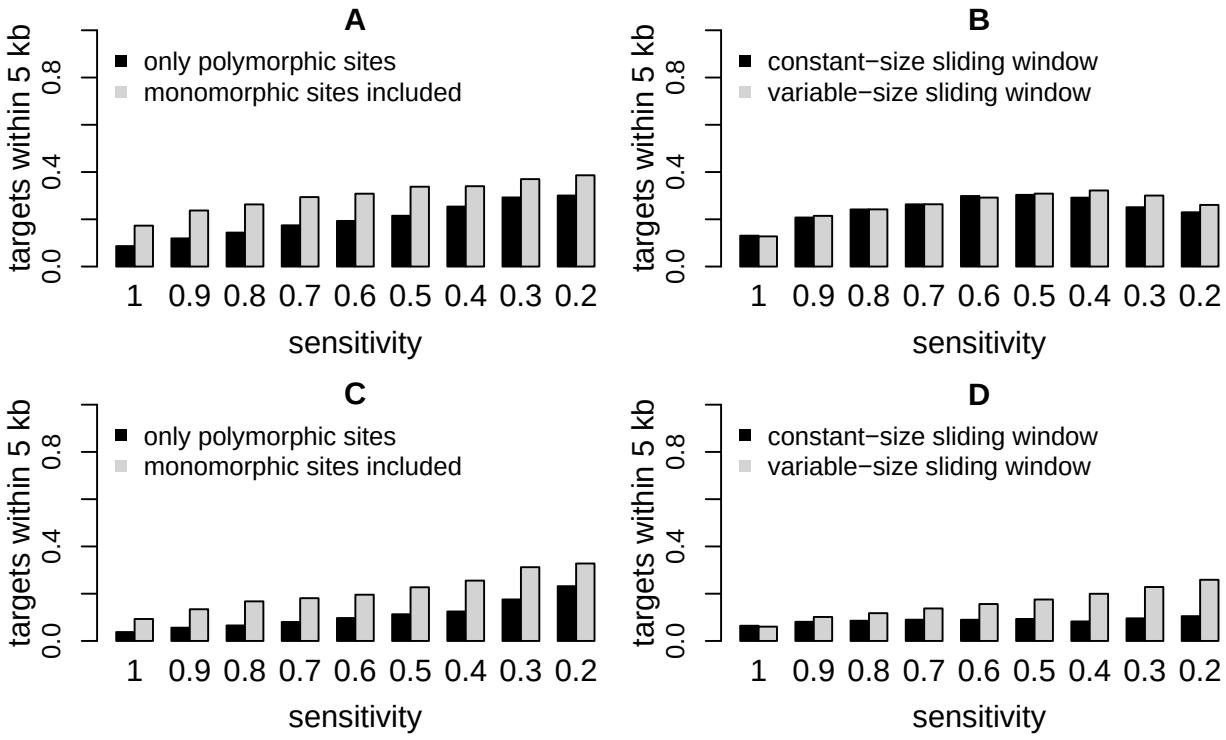

FIGURE S5: The fraction of predicted targets within 5 kb from the true location of the selective sweep for a recurrent selective sweep scenario where $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.25$ (above A) and B)) and $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.50$ (below C) and D)). A) and C): Comparison of the precision of *SweepFinder* when only polymorphic sites are used (dark bars) and a fraction of monomorphic sites is embedded (light bars). B) and D): Comparison between the variable-size sliding window approach and the constant-size sliding window approach. The precision of the two approaches is similar for low threshold values (high sensitivity, low specificity). However, for higher cutoff values the variable-size sliding window method is slightly more precise. Simulations assume a 100-kb genomic fragment. Selective sweeps have occurred uniformly within this region or within its flanking regions following a homogeneous Poisson distribution in time. The selection coefficient is $s = 0.01$, $\theta = 0.008/\mathrm{bp}$, and $\rho = 0.08/\mathrm{bp}$.

Figure S6

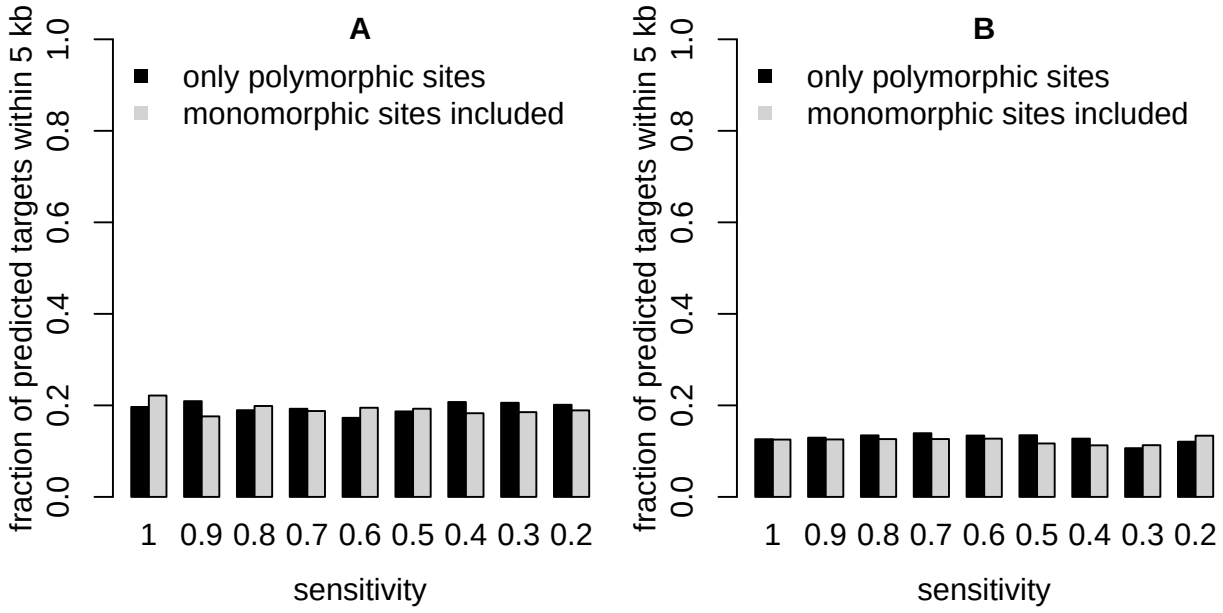

FIGURE S6: The expected fraction of peaks whose distance from the randomized 'target of selection' is smaller than 5 kb. The target randomization was performed A) in the SHH model with $\alpha = 2500$, $\theta = 0.005$/bp, and $\rho = 0.05$/bp and B) in the RHH model with $\frac{H_{\mathrm{RHH}}}{H_{\mathrm{NEU}}} = 0.25$ and selection coefficient $s = 0.01$, $\theta = 0.008$/bp, and $\rho = 0.08$/bp. In the SHH model the length of the simulated region is 50 kb and a selective sweep has occurred in the middle of this region. Then, this target was distributed uniformly between 0 and 50 kb. In the RHH model the length of the simulated region is 100 kb and selective sweeps have occurred inside or outside the region. These targets of selection are distributed uniformly. The graphs denote that even if the target is a random point on the genomic region some peaks will be in the proximity of the target as they are located in the same region.

Figure S7

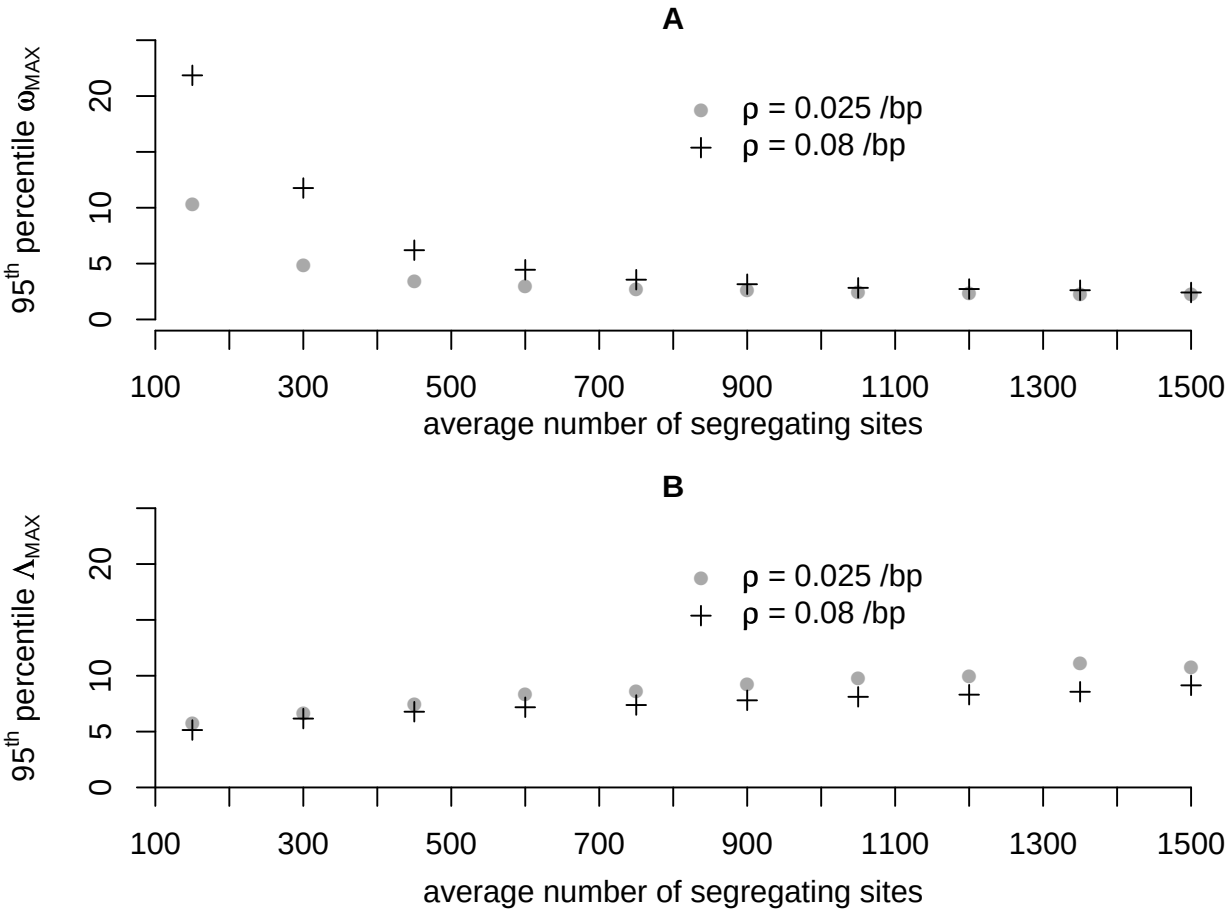

FIGURE S7: Variable values of recombination rate affect A) the ω_{MAX} and B) Λ_{MAX} values. Simulations have been performed for various levels of recombination rate for a constant-size neutral population. All remaining parameters are equal among all simulations. For each value of S_n on the x-axis we have used θ that on average generates S_n segregating sites ($\theta = \frac{S_n}{h_n}$). The ω-statistic is primarily affected for lower numbers of polymorphic sites, while the opposite is true for the *SweepFinder*. The demographic model used in the simulations represents the standard neutral model.

Figure S8

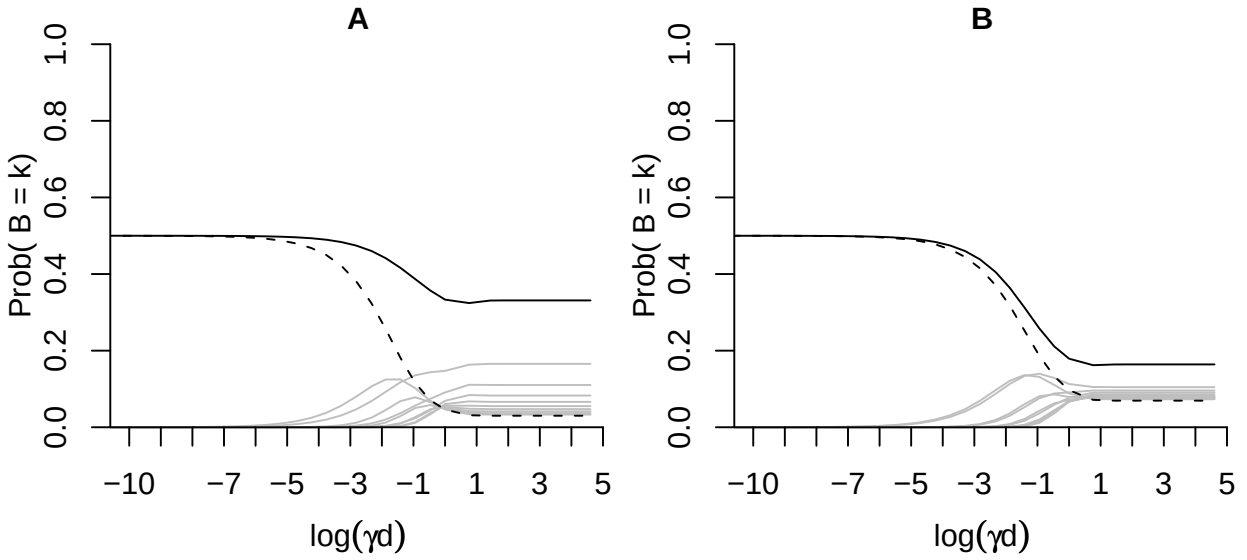

FIGURE S8: The likelihood curves for each polymorphism class: A) under an equilibrium selection model ($\rho = 0.05$/bp, $\theta = 0.005$/bp), and B) under the *TS* ($\rho = 0.08$/bp, $\theta = 0.008$/bp) model. For both scenarios a selective sweep has occurred in the middle of a 50-kb region and the selection intensity $\alpha = 2500$. The x-axis denotes the value of parameter $\gamma = \frac{r}{s}\log(2N)$ (log scale) multiplied by the distance *d* from the center of the sweep. If we assume a constant recombination rate *r* and selection coefficient *s*, γ represents the distance from the location of the selective sweep *x*. The likelihood curve for the singletons (class '1') is depicted by the black solid line, whereas the class '11' (out of 12 sequences) is represented by a black dashed line. Gray lines illustrate the likelihood curves for the classes 2-10. For both A) and B) the class '1' and the class '11' contribute to the likelihood close to the sweep. Conversely, classes 2-10 contribute at larger distances from *x*. The major difference between A) and B) is that the singleton contribution is lower in B) than A) at larger distances. This is because the frequency of singletons is greater in B) than in A).

Curriculum vitae

CONTACT INFORMATION

Address
LMU Biozentrum
Department Biologie II
Grohaderner Str. 2
82152 Planegg-Martinsried
Germany

Email
pavlidis@bio.lmu.de
pavlidisp@gmail.com
Web
bio.lmu.de/~pavlidis/

PERSONAL INFORMATION

Nationality	*Greek*
Date of Birth	*15 July 1980*
Marital Status	*Married, two children*

EDUCATION

PhD Student in Evolutionary Biology — **2007-present**
Ludwig-Maximilians-Universität — *Munich, Germany*

PhD Student in Computer Science and Mathematics — **2005–2007**
University of Tartu — *Tartu, Estonia*

M.Sc. of Biology — **2003–2005**
University of Crete — *Heraklion, Greece*

Diploma in Agricultural Biotechnology — **1998–2003**
Agricultural University of Athens — *Athens, Greece*

WORKSHOPS & RESEARCH ACTIVITIES

Training in the group of Prof. Jan Komorowski on microarray classification methods — **Jun. 2005**
University of Uppsala — *Uppsala, Swedem*

Evolutionary Biology Course in Guarda — **Jun. 2006**
University of Basel — *Basel, Switzerland*

Summer Institute in Statistical Genetics — **Sep. 2007**
University of Liege — *Liege, Belgium*

SELECTED PRESENTATIONS IN SYMPOSIA & CONFERENCES

Detection of positive selection and demographic inference using SNP data, *Jul. 2010*
P. Pavlidis, S. Laurent, J.D. Jensen, and W. Stephan, *SMBE 2010, Lyon, France*

Detecting the signature of positive selection in genomes, *May 2009*
P. Pavlidis, and W. Stephan, *VW 2009, Münster, Germany*

Evolutionary conservation of *BDNF* gene coexpression network suggests potential regulators of *BDNF* gene expression, *Jun. 2009*
T. Aid-Pavlidis, P.Pavlidis, T.Timmusk, *SMBE 2009, Barcelona, Spain*

Evolution of Behavioral Strategies in Group Selection, *Apr. 2003*
C. Bouyioukos P. Pavlidis, and J. Sourdis, *RECOMB, Berlin, Germany*

The problem of genomic rearrangements, *May 2000*
P. Pavlidis, C. Bouyioukos, L. Salichos, and J. Sourdis, *EEBE, Skiathos, Greece*

HONORS & AWARDS

Award from Greek National Scholarship's Foundation for the M.Sc. program in Plant Molecular Biology and Biotechnology
2004, University of Crete, *Heraklion, Greece*

Award from Greek National Scholarship's Foundation for the M.Sc. program in Plant Molecular Biology and Biotechnology
2003, University of Crete, *Heraklion, Greece*

Award from Greek National Scholarship's Foundation, 2nd in Panhellenic Entrance Exams
1998, *Athens, Greece*

SCHOLARSHIPS

Fellowship from the Volkswagen Stiftung, for Ph.D. studies
2007–2010, Ludwig-Maximilians-University, *Munich, Germany*

Scholarship from the Estonian government for Ph.D. studies
2005–2007, University of Tartu, *Tartu, Estonia*

INVITED TALKS

The problem of detecting selective sweeps in bottlenecked populations
May 2010, HCMR, *Heraklion, Greece*

Combining SFS and LD to detect selective sweeps: a machine learning approach
Apr. 2010, University of Vanderbilt, *Nashville, TN, USA*

Detecting selective sweeps in natural populations of *D. melanogaster*
Apr. 2010, MPI for Evolutionary Anthropology, *Leipzig, Germany*

A novel approach for detecting selective sweeps in whole genome data, using machine learning techniques
Jun. 2009, University of Zürich, *Zürich, Switzerland*

Detecting sweeps in whole genome data
Apr. 2009, Evolutionary Genomics: the impact of next generation sequencing technologies, *Wittenberg, Germany*

PUBLICATIONS

Svetec N, Werzner A, Wilches R, Pavlidis P, Álvarez-Castro J, Broman K, Metzler D, Stephan W. 2011, **Identification of X-linked quantitative trait loci affecting cold tolerance in** *Drosophila melanogaster* **and fine-mapping by selective sweep analysis.** *Molecular Ecology, 20:530–544*

Pavlidis P, Jensen JD, Stephan W. 2010, **Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations.** *Genetics, 185:907–922.*

Pavlidis P*, Laurent S*, and Stephan W. 2010, **msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis.** *Molecular Ecology Resources, 10:723–727,* *authors have contributed equally

Tsuchimatsu T, Suwabe K, Shimizu-Inatsugi R, Isokawa S, Pavlidis P, Städler T, Suzuki G, Takayama S, Watanabe M, Shimizu KK. 2010, **Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene.** *Nature, 464:1342–1346*

Aid-Pavlidis T*, Pavlidis P*, Timmusk T. 2009, **Meta-coexpression conservation analysis of microarray data for studying the regulation of a single gene - BDNF gene case study.** *BMC Genomics, 10:420,* *authors have contributed equally

Svetec N, Pavlidis P, Stephan W. 2009, **Recent strong positive selection on** *Drosophila melanogaster HDAC6***, a gene encoding a stress surveillance factor, as revealed by population genomic analysis.** *Molecular Biology and Evolution 26:1549–1556*

Johnsen JM, Teschke M, Pavlidis P, McGee BM, Tautz D, Ginsburg D, Baines JF. 2009, **Selection on cis-regulatory variation at** *B4galnt2* **and its influence on von Willebrand Factor in house mice.** *Molecular Biology and Evolution 26:567–578*

Pavlidis P, Hutter S, Stephan W. 2008, **A population genomic approach to map recent positive selection in model species.** *Molecular Ecology 17:3585–3598*

Pavlidis P, Poirazi P. 2006, **Individualized markers optimize class prediction of microarray data.** *BMC Bioinformatics, 7:345-358*