UNIVERSITY OF CRETE

MASTER THESIS

Localizing selective sweeps using higher order Site Frequency Spectra: the 2-D SFS.

Author:
Maria Malliarou

Supervisors:
Pavlos Pavlidis
Christophoros Nikolaou
Emmanouil Ladoukakis

A thesis submitted in fulfillment of the requirements for the degree of Master in Bioinformatics

in the

Medical School, University of Crete, Greece

July, 2020

Πανεπιστήμιο Κρήτης

Μεταπτυχιακή Εργασία

Εντοπισμός επιλεκτικής συμπαράσυρσης χρησιμοποιώντας μεγαλυτερης διάστασης φάσμα αλληλικών συχνοτήτων**: 2D-SFS**

Συγγραφέας:
Μαρία Μαλλιαρού

Επιβλέποντες:
Παύλος Παυλίδης
Χριστόφορος Νικολάου
Εμμανουήλ Λαδουκάκης

Η παρούσα πτυχιακή εργασία πραγματοποιηθηκε για τις απαιτήσεις του μεταπτυχιακού προγράμματος σπουδών Βιοπληροφορικής

στο

Τμήμα Ιατρικής, Πανεπιστήμιο Κρήτης, Ελλάδα

Ιούλιος 2020

UNIVERSITY OF CRETE

Abstract

Department of Medicine

Master in Bioinformatics

Localizing selective sweeps using higher order Site Frequency Spectra: the 2-D SFS.

by Maria Malliarou

When an allele is favored by natural selection, its frequency may increase in the population and linked neutral variants will increase their frequency as well a phenomenon called selective sweep. As the distance from the beneficial mutation increases, recombination will allow neutral variants to escape the so-called 'hitchhiking effect' of the beneficial mutation, thus, generating characteristic patterns of neutral variants, locally, around the target of natural selection. These patterns of neutral variants have been exploited in the last 20 years by computational methods that aim at localizing the action of positive selection on genomes obtained from natural populations. It is well-known that certain demographic models pose severe challenges on the detection of selective sweeps because they generate patterns of neutral variants that resemble those of a selective sweep. Considerable effort has been devoted to understanding the patterns of neutral variants generated by demographic models, solely. However, we still have no description of a selective sweep model in a population that has experienced demographic changes during its evolutionary history. Thus, even though neutrality is well-described, this is not the case for the model of positive selection. Here, we present a novel methodology for incorporating demography on the selection model of a selective sweep. We demonstrate that certain demographic models may change dramatically the well-known patterns of positive selection. To facilitate the detection of positive selection in non-isolated natural populations that have experienced demographic changes we implemented a selective sweep detection software, called SweeD-sim that extends Sweep Detector (SweeD) software. Moreover, to incorporate specific Linkage Disequilibrium patterns of genetic hitchhiking we have expanded our approach in two dimensions, creating a so called Two Dimensional Site Frequency Spectrum (2D-SFS) in hope of more significant results .

Keywords: selective sweeps, simulations, kernel density estimation, 2D-SFS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

Περίληψη

Τμήμα Ιατρικής

Μεταπτυχιακό Βιοπληροφορικής

Εντοπισμός επιλεκτικής συμπαράσυρσης χρησιμοποιώντας μεγαλυτερης διάστασης φάσμα αλληλικών συχνοτήτων**: 2D-SFS**

Μαρία Μαλλιαρού

Όταν ένα αλληλόμορφο επιλέγεται μέσω φυσικής επιλογής, η συχνότητά του μπορεί να αυξηθεί στο πλυθυσμό και συνδεδεμένες με αυτό, ουδέτερες μεταλλάξεις, θα αυξηθούν σε συχνότητα, ένα φαινόμενο που ονομάζεται επιλεκτική συμπαράσυρση. Καθώς η απόσταση από την ευεργετική μετάλλαξη αυξάνεται, ο ανασυνδυασμός επιτρέπει ουδέτερες παραλλαγές να ξεφύγουν από το λεγόμενο «φαινόμενο ωτοστόπ» της ωφέλιμης μεταλλαγής, δημιουργώντας έτσι χαρακτηριστικά μοτίβα ουδέτερων παραλλαγών, τοπικά, γύρω από τον στόχο της φυσικής επιλογής. Αυτά τα μοτίβα ουδέτερων πολυμορφισμών έχουν αξιοποιηθεί τα τελευταία 20 χρόνια με υπολογιστικές μεθόδους που στοχεύουν στον εντοπισμό της δράσης της θετικής επιλογής σε γονιδιώματα που λαμβάνονται από φυσικούς πληθυσμούς. Είναι γνωστό ότι ορισμένα δημογραφικά μοντέλα θέτουν σοβαρές προκλήσεις στην ανίχνευση επιλεκτικής συμπαράσυρσης επειδή δημιουργούν μοτίβα ουδέτερων πολυμορφισμών που μοιάζουν με εκείνα μίας επιλεκτικης συμπαράσυρσης. Έχει καταβληθεί σημαντική προσπάθεια για την κατανόηση των προτύπων ουδέτερων παραλλαγών που δημιουργούνται αποκλειστικά από δημογραφικά μοντέλα. Ωστόσο, δεν υπάρχει ακόμη περιγραφή ενός επιλεκτικού μοντέλου σάρωσης σε έναν πληθυσμό που έχει βιώσει δημογραφικές αλλαγές κατά την εξελικτική του ιστορία. Έτσι, παρόλο που η ουδετερότητα έχει μελετηθεί εκτενώς, αυτό δεν ισχύει για το μοντέλο της θετικής επιλογής. Εδώ, παρουσιάζουμε μια νέα μεθοδολογία για την ενσωμάτωση της δημογραφικής ιστορίας ενός πλυθυσμού στο μοντέλο επιλογής μιας επιλεκτικής σάρωσης. Δείχνουμε ότι ορισμένα δημογραφικά μοντέλα μπορεί να αλλάξουν δραματικά τα γνωστά πρότυπα θετικής επιλογής. Για να διευκολύνουμε την ανίχνευση της θετικής επιλογής σε μη απομονωμένους φυσικούς πληθυσμούς που έχουν βιώσει δημογραφικές αλλαγές, εφαρμόσαμε ένα λογισμικό ανίχνευσης επιλογής μέσω προσομοιώσεων (SweeD-sim) που επεκτείνει το ήδη υπάρχον εργαλείο Sweep Detector(SweeD). Επιπλέον, για να ενσωματώσουμε συγκεκριμένα μοτίβα ανισορροπίας σύνδεσης της γενετικής συμπαράσυρσης, έχουμε επεκτείνει την προσέγγισή μας σε δύο διαστάσεις, δημιουργώντας ένα λεγόμενο δισδιάστατο φάσμα αλληλικών συχνοτήτων (2D-SFS) με στόχο στατιστικά σημαντικότερα αποτελέσματα.

Λέξεις κλειδιά**:** επιλεκτική συμπαράσυρση, προσομοιώσεις, εκτίμηση πυκνότητας με πυρήνες, δισδιάτατο φασμα αλληλικών συχνοτήτων

Acknowledgements

This work has been carried out at the Institute of Computer Science (ICS) a part of Foundation of Research and Technology Hellas (FORTH), Crete, Greece. I would like to express my sincere gratitude to my main supervisor, Dr. Pavlos Pavlidis not only for his contribution and insightful comments to this project but mostly for his patience and believing in my abilities in times where I thought about pursuing other things. I am also grateful to all my lab members, especially to Antonios Kioukis and Aggelos Koropoulis for giving me useful information about computer science, for their constant motivation and making the working environment extremely fun to be around.

Contents

Chapter 1

Introduction

1.1 Selective Sweep Theory

When a beneficial mutation emerges in a population, its frequency tends to increase in individuals. Consequently, neutral or weakly selected variants close to the mutation under selection will also have higher frequencies, a process that Maynard Smith and Haigh (1974) (1) called genetic hitchhiking. As a result of this process, hitchhiking events can reduce genetic variation near the site of selection in a genome, thereby inducing a selective sweep. Selection against recurrent deleterious mutations also reduces variation at linked loci (Charlesworth et al., 1993 (2)) a mechanism called "background selection" which causes the continuous removal of linked sequences along with deleterious mutations.

The fundamentals of the hitchhiking model which were analyzed by Maynard Smith and Haigh (1974) (1) are shown in Figure 1.1. Initially, when a beneficial allele arises by mutation, three different haplotypes are present in the population: two of them are polymorphic at the neutral locus (with alleles A and a) and monomorphic at a selected locus nearby, while the third haplotype carries the beneficial allele at the selected locus and the neutral allele A at the other locus. After fixation of the beneficial allele, only one haplotype exists in the population. If no recombination event has occurred between the neutral and selected loci (lower left side of the panel); in this case, variation at the neutral locus is eliminated at the time of fixation through the hitchhiking effect. If recombination has occurred during the fixation process of the beneficial allele, the neutral locus remains polymorphic, and thus two haplotypes are present in the population (lower right side). After fixation of the beneficial allele, the neutral locus remains polymorphic as long as it can escape hitchhiking. The chance of this happening increases with the recombination rate and with the time available for recombination to occur. The latter is proportional to the selection coefficient of the beneficial allele.

Although genetic hitch-hiking was introduced in 1974, it was in the late 1980's where patterns of reduced variation were found in regions of the genome with low recombination rates in Drosophila Species (Aguadé et al., 1989; (4), Stephan and Langley, 1989 (5)). Also in 1991, Berry et al. (6) showed lack of polymorphism in the cubitus interruptus locus located on the nonrecombinant fourth chromosome of Drosophila natural lines, indicating recent positive sweeps. Begun and Aquadro's work in 1992 (7) demonstrated a correlation between levels of DNA variation and recombination rates in the *D.melanogaster* genome whereas average divergence to *D.simulans* was hardly affected by recombination. All of the aforementioned studies led to more extended models of genetic hitch-hiking by Kaplan et al. (1989) (8) who created a three-phase simulation model to study the effect of selective sweeps on genealogies, Wiehe and Stephan (1993) (9) who used diffusion equations methods, Barton (1998) (10) who improved the theory concerning the effects of a single

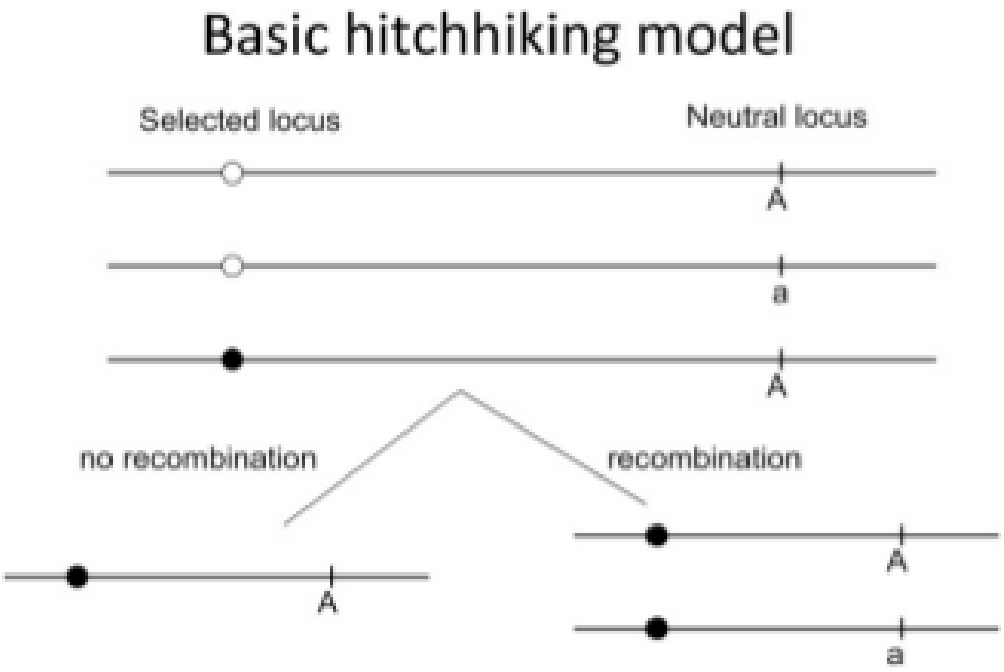


FIGURE (1.1) Basic hitchhiking model. The upper part of the figure shows the three haplotypes present in a population when a beneficial mutation (filled circle) occurs at the selected locus. The wild type allele at the selected locus is indicated by an open circle. At the neutral locus two alleles A and a are present. The haplotypes after the fixation of the beneficial allele are depicted in the lower part of the figure. If no recombination occurs during the fixation process one haplotype is present (left side).With recombination the neutral locus stays polymorphic and two haplotypes remain (right side).(Stephan 2019) (3)

sweep and Gillespie (2000) (11) who called this effect "genetic draft" and studied it as stochastic process. The authors have determined that in large populations diversity vanishes in recombining regions at the site of selection immediately after the fixation of the beneficial allele and increases as a function of the ratio of the recombination rate (between the neutral and selected sites) and the selection coefficient.

1.2 Signatures of Selective Sweeps

1.2.1 Site Frequency Spectrum

The Site Frequency Spectrum (SFS) is a count of the number of mutations that exist in a frequency of $x_i = \frac{i}{n}$ for $i = 1, 2, ..., n-1$, in a sample of size n. In other words, it represents a summary of the allele frequencies of the various mutations in the sample. In a standard neutral model (i.e., a model with random mating, constant population size, no population subdivision, etc), the expected value of x_i is proportional to $\frac{1}{i}$. As shown by Braverman et al. (1995) (12) and Fay and Wu (2000) (13), a selective sweep will increase the fraction of mutations segregating at low frequencies in the sample. Neutral variants that are initially linked to the beneficial mutation, increase in frequency, whereas variants that are initially not linked to the beneficial

mutation, decrease in frequency during the fixation of the beneficial mutation. Figure 1.2 illustrates the shift of the SFS after a selective sweep and the corresponding polymorphic table.

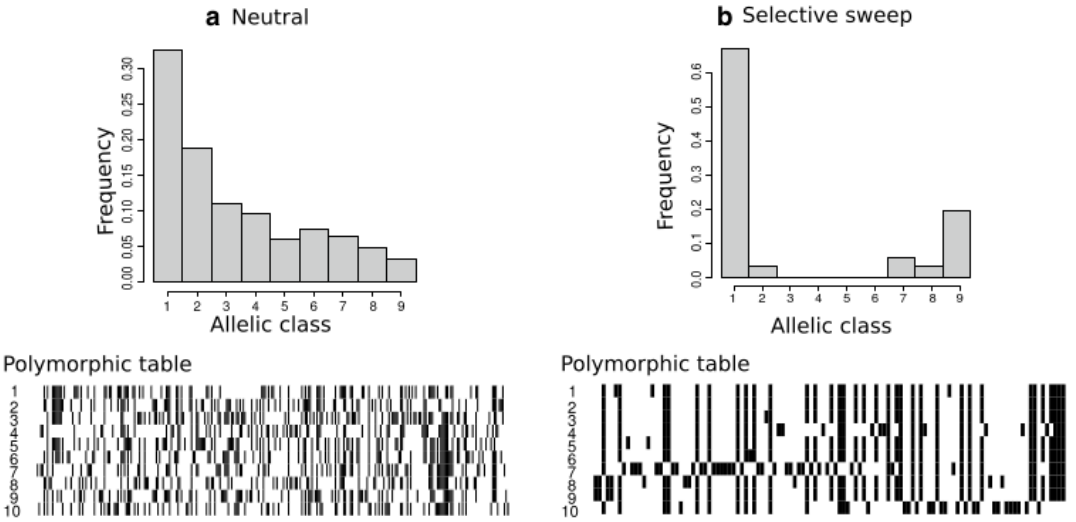


FIGURE (1.2) The SFS signature of a selective sweep compared to the neutral SFS. In the polymorphic table, black cells denote derived alleles, whereas the white cells denote ancestral alleles. Each column in the polymorphic table represents a SNP. Monomorphic sites have been excluded. a Neutral SFS and its respective polymorphic table. b SFS after a selective sweep and its respective polymorphic table (14)

Given this information, Kim and Stephan (2002) (15) developed a composite-likelihood ratio (CLR) test to detect local reductions of nucleotide diversity along a recombining chromosome, and predict the strength and location of the target of selection, using the SFS signature of a selective sweep. In a likelihood ratio test we compare two statistical models usually by maximizing the parameters in the first model and keep the second. The CLR test compares the probability of the observed polymorphism data under the standard neutral model (i.e., constant population size) with the probability of the data under the model of a selective sweep. Under neutrality, the expected number of sites where the derived variant is in the frequency interval [p, p + dp] in the population is given by

$$\phi_0(p)dp = \frac{\theta}{p}dp \qquad (16)$$

where θ is given from the formula $\theta = 4N\mu$ with N being the effective number of individuals in a diploid population and μ the mutation rate per generation. Based on the model proposed by Fay and Wu (2002) (13) after a hitchhiking event this distribution is transformed approximately to

$$\phi_0(p) = \begin{cases} \frac{\theta}{p} - \frac{\theta}{C}, & \text{for } 0 < p < C. \\ \frac{\theta}{C}, & \text{for } 1 - C < p < 1\,. \end{cases}$$

, where C is given approximately by $1 - \epsilon^{\frac{r}{s}}$, where ϵ is the frequency of the beneficial allele when it begins to increase deterministically and approximately is estimated by the formula $\epsilon = \frac{1}{a}$ based on simulations, $a = 2Ns$, s is the selection coefficient, r is the recombination fraction between the neutral locus and the selected locus. The probability of observing a site where k derived alleles are found in a sample of size n is given by

$$P_{n,k} = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \phi(p) dp, \ (\ k = 1, ..., n-1)$$

and

$$P_{n,0} = 1 - (P_{n,1} + ... + P_{n,n-1})$$

where $\phi(.) = \phi_0(.)$ under the neutral model and $\phi(.) = \phi_1(.)$ under the hitchhiking model. The likelihood of all data under the model of genetic hitchhiking is obtained by multiplying the probabilities for all nucleotide sites under consideration. Then, the authors calculated the maximum composite likelihood under the neutral model (L_0) and that under the hitchhiking model (L_1) and the likelihood ratio is given by L_1/L_0. Since both functions have a lot of free parameters, the authors chose only s as a free variable that maximizes the CLR and the others are either specified or obtained empirically by the data. Although the Kim and Stephan CLR test was the first test to detect selective sweeps, it was only efficient for small genome fragments and since it was based on Fay and Wu's model, it was sensitive to assumptions regarding mutation rates and rates of recombination. Moreover, since the neutral model was derived by an equilibrium neutral population, i.e., a population with constant population size, Jensen et al. (2005) (17) showed that the test is not robust to undetected population structure or a recent bottleneck, with some parameter combinations resulting in a false positive rate of nearly 90%.

In 2005, Nielsen et al. (18) developed a similar method for detecting selective sweeps in whole-genome data comparing a model under neutrality and a selective sweep model by introducing two modifications to the Fay and Wu's CRL test. For the neutral model (the denominator of the CLR test), instead of assuming standard neutrality and obtaining the SFS based on Kimura (1971), they used the empirical SFS derived from the background pattern of variation of their data. For their selection model, they estimated the probability p_e of each ancestral state escaping a selective sweep through recombination onto the selected background. This approximated probability offered by studies like Maynard Smith and Haigh 1974 (1); Barton 1998 (10); Kim and Stephan 2002 (15)) of the descendant neutral copy at the end of a sweep is given by

$$p_e = 1 - e^{-\alpha d}$$

where d is the distance of from the location of the sweep to the sampled SNP locus and α is a parameter that depends on the rate of recombination, the effective population size, and the selection coefficient of the selected mutation. Then under specific assumptions, the probability $p_e(k)$, that k where $0 < k < n$, out of n gene copies sampled for a locus escaped the sweep, is binomially distributed with parameters p_e and n :

$$P_e(k) = \binom{n}{k} p_e^k (1-p_e)^{n-k}$$

If k lineages escape the sweep, the ancestral sample right before the sweep contains H = minn, k + 1 lineages. If the distribution of allele frequencies in a sample of size n, in the absence of a selective sweep, is given by p = $(p_1, p_2, ..., p_{n-1})$)then the probability of observing j mutant lineages in an ancestral sample of size H is given by

$$p_{j,H} = \sum_{i=j}^{n-1} p_i \frac{\binom{i}{j}\binom{n-i}{H-j}}{\binom{n}{H}}$$

If there are j mutant lineages in an ancestral sample of size k + 1, the probability that the most recent common ancestor of the lineages that did not escape the selective sweep is of the mutant type, is j / (k + 1). This implies that the probability of observing a mutant allele of frequency B out of n in the sample after a selective sweep, is

$$p_B^* = p_e(n)p_B + \sum_{k=0}^{n-1} p_e(k)(p_{B+1-n+k,k+1}\frac{B+1-n+k}{k+1} + p_{B,k+1}\frac{k+1-B}{k+1})$$

With this expression, they calculated the composite likelihood for a set of SNP data assuming a selective sweep of intensity α at a certain location in the genome where by maximizing for α and p for all possible locations in the genome. Since this modification does not rely on the θ parameter which can vary among a chromosome, it is feasible to perform the test on a larger scale. Computationally, their tool, SweepFinder, can process small and moderate sample sizes efficiently but not a large number of sequences. Yet again, their model does not take into account demographic scenarios like a bottleneck, population subdivision or gene-flow with other populations.

In 2013, Pavlidis et al. (19), released SweeD, a computationally advanced test based on the SweepFinder algorithm which provides the user the option to employ a user-specified demographic model for the theoretical calculation of the expected neutral SFS. Their implementation can analyze data of a larger scale which increases sweep detection accuracy.

1.2.2 Linkage Disequilibrium

Levels of linkage disequilibrium (LD), the correlation among alleles from different loci, tend to increase in regions under selection although Przeworski (2001) (20) showed that this phase may be relatively short. Upon fixation of the beneficial mutation, elevated levels of LD emerge on each side of the selected site, whereas a decreased LD level is observed between sites found on different sides of the selected site. The high LD levels on the different sides of the selected locus are due to the fact that a single recombination event allows existing polymorphisms on the same side of the sweep to escape the sweep. On the other hand, polymorphisms that reside on different sides of the selected locus need a minimum of two recombination events in order to escape the sweep. Given that recombination events are independent, the level of LD between SNPs that are located on different sides of the positively selected mutation decreases.

1.3 Selective Sweeps and Coalescence Theory

Coalescence (Hudson, 1983 (21); Kingman, 1982 (22)) is a stochastic process to generate genealogies from a population by tracing randomly sampled alleles backwards in time. Consider a population of N individuals that reproduce under the neutral Wright - Fisher model which means that each generation is discrete and is formed by randomly sampling N parents with replacement from the current generation.

The number of offsprings of a specific individual is binomially distributed where the probability of being chosen is 1/N whereas the probability of not being chosen is 1 - 1 / N. Going backwards in time, lineages coalesce whenever two or more individuals are produced from the same parent until the most recent common ancestor (MRCA) is found. This means that in a sample of k individuals where k(k-1)/ 2 pairs could coalesce the probability of one of these coalesces in the previous generation is given by

$$P(coalescence) = \frac{k(k-1)}{2}\frac{1}{2N}$$

As a result, expected coalescence time for k alleles is exponentially distributed with a mean 4N and coalescence rate of $\frac{k(k-1)}{4N}$ for diploid populations and a mean $2N$ and coalescence rate of $\frac{k(k-1)}{2N}$ for haploid population. Since the assumptions of the Wright Fisher model are violated in the case of an loci undergoing a selective sweep Hudson and Kaplan (1986) showed how conditioning on the allelic types of a sample alters the coalescent process, in a way that is similar to the effect of geographic structure and migration. Two lineages with the same allelic type may coalesce, but two lineages with different types must wait for mutation to change the type of one or the other. This idea led Hudson et al. (1988) (23) to apply this idea to a locus under selection, showing that rates of coalescence and mutation in the ancestral process depend on the frequencies of the two alleles. Consider a locus A with two alleles A_1 and A_2 where A_2 is selectively favored. If in a given time t in the past the allelic frequency of A_2 is x(t) and there are i ancestral lineages then the rate of coalescence between any pair of them is $\frac{1}{x(t)}$, and the total rate is

$$\frac{\binom{i}{2}}{x(t)} = \frac{i!2}{(i-2)!x(t)}$$

If x(t) = 1, the rate is the same as in the standard neutral coalescent. However, if $x(t) < 1$, then the rate is greater than in the standard neutral coalescent. The reason for this is that, when x(t) is smaller, there are fewer possible parents of the i lineages, so the probability of a common ancestor in a single generation is larger. Then let's consider a locus B close to the selected locus A at a distance m. Each of the members of a sample of size n taken at the B locus will be linked either to an A_1 allele or to an A_2 allele at the selected locus, and the same is true of the ancestral lineages of the sample.This linkage that makes the ancestry at the B locus differ from the standard neutral coalescent. This means that if i B-locus lineages are linked to A_2 alleles, then the rate of coalescence between each pair is $\frac{1}{x(t)}$ and the total rate is identical to the total rate for the A locus as given. This is true to m values not being too large,minimising the possibility that both recombination and coalescence occur in a single generation. The more we move away, the probability of B-locus lineages escaping the sweep is increasing. As we go backwards in time the frequency of A_2 decreases from 1 down to 1 / (2N) which means that for the B-locus alleles that are linked to A_2, the rates of coalescence will increase as x(t) decreases since it depends on 1 / x(t). As a result we can conclude that coalescent events occur at higher rates in the presence of a sweep than they do in its absence producing short trees whereas in neutrality we have short branches that get longer as we move closely to the MRCA.

1.4 Demography in selective sweep detection

In chapter 1.2, we discussed signatures that are strong indicators of a selective sweep which can be generated by the fixation of a favoured allele, coalescent rates and recombination events. Nevertheless, we should take into account that demographic events can generate similar patterns of polymorphisms that resemble the signatures of genetic hitchhiking. For example, let's assume a population of large effective size that experienced a severe bottleneck (decrease of the population size) that its ancestral state was also of large population size. As we mentioned above, since the coalescent rate is inversely proportional to the size, the probability of observing a larger number of coalescent events in a short period of time is increased although in cases where the bottleneck is not so severe,lineages can escape the bottleneck, taking them more time to coalesce. In a recombining chromosome, genomic regions that have witnessed a massive amount of coalescent events during the bottleneck phase may alternate with genomic regions with lineages that have escaped the bottleneck phase which subsequently means that we observe similar SNP patterns to those generated by as selective sweep making the detection process impractical because of large false positive rates. In order to minimise these rates, the methods we described above, take advantage of the fact that while the effects of a selective sweep is observed only to small regions close to the selection site, neutral demographic changes generate genome-wide patterns. Initially, they estimate an average, genome-wide SFS calling it background SFS followed by detecting regions that fit the selection model but not the background SFS. This approach assumes that SFS does not variate in a recombining genome which according to Becquet (2003) (24) is not the case. Instead he showed that a bottleneck in the presence of recombination results in increased heterogeneity in variability patterns along a chromosome, reminiscent of the effects of selection. There is a need, therefore, to understand selective sweep models in the presence of demographic changes. Such models have not been developed yet, mainly due to the mathematical challenges of the problem. In addition, neutral demographic models of natural populations may be composed of multiple populations that exchange migrants. The implementation of a general model that will model selective sweeps in the presence of demographic changes is therefore extremely challenging.

In this manuscript, we describe an alternative approach, in which the likelihood of a selection model in the presence of past demographic changes is approximated by simulations and kernel density estimation functions. With these simulations, we hope to extract models of selection that best describe populations that have been strongly affected by both natural selection and past demographic events.

Chapter 2

Methods

The basic concept of this project is to create a tool that can facilitate the detection of positive selection in populations that have experienced demographic changes. For this purpose, we use the general concept of the aforementioned Composite Likelihood Ratio but, instead of applying mathematical equations as the previous approaches, we perform simulations in order to extract specific allelic frequencies patterns. Based on a simulation with specific demographic and selection parameters, we then calculate the probability of the occurrence of an allelic class based on its distance from a candidate selective region and with those probabilities we perform the CLR test in samples under consideration. Based on the results, we need to determine whether our sample has been under the influence of positive selection or not. We describe the method in detail, in the following sections.

2.1 Data Generation

The first step of our tool is the creation of a specific data set with similar mutation, recombination and demographic parameters as the sample under consideration. There are several tools cited in the literature that can accurately estimate all the above parameters. **(add some literature if possible BOTTLENECK, GENEPOP etc)** To generate our data, we used the mssel, an extension of ms (25), a Monte Carlo computer program written in C, that generates samples drawn from a population evolving according to a Wright-Fisher neutral model. The program assumes an infinite-sites model of mutation, and allows recombination, gene conversion, symmetric migration among sub-populations, and a variety of demographic histories. For each sample, the program generates a random genealogical history of a segment of a chromosome. Conditional on the genealogy of a sample, mutations are randomly placed on the genealogy according to the usual assumption that the number of mutations on a branch is Poisson distributed with mean given by the product of the mutation rate and the branch length. The times between nodes in the genealogy are approximated by continuous (exponential) distributions. For the analysis we constructed specific scenarios where each bottleneck model is characterized by a reduction in population size at some point in time and a recovery to the original population size (backwards in time) and for each demographic model we generated 1000 datasets. The mutation parameter of all models was set to $4N\mu = 2000$ and selection is supposed to act in the middle of the region. We also constructed neutral data sets with the same parameters in order to evaluate our method and estimate true and false positive rates.

2.2 Selective Sweep Model

After obtaining the simulated data set under selection, the next step is to calculate the probabilities of the allelic classes X_i (i = 1,2,...,N-1) given its distance (D) from the selected loci ($P(X_i|D)$. This conditional probability can be described mathematically by Bayes theorem and can be stated as the following equation:

$$P(X_i|D) = \frac{P(D|X_i)P(X_i)}{P(D)}$$

where $P(D|X_i)$ is the probability to observe specifically the allelic class X_i D units away from the selection site, P(X_i) is the probability of the class independently of its location and P(D) is the probability of observing any allelic class in the specific location and can be further analysed as

$$P(D) = P(D|X_1)P(X_1) + P(D|X_2)P(X_2) + ... + P(D|X_{N-1})P(X_{N-1})$$

All the components of Bayes' theorem can be inferred from the simulated data set where for each allelic class we obtain the positions of their occurrences and their total frequencies (SFS). Each position is transformed based on its distance from the selection site. If we find n occurrences of an allelic class, let ($D_1, D_2, ...D_n$) be the sampled distances drawn from some distribution with an unknown density . This probability density function (PDF) can later be used to specify the probability of any given variable falling within a particular range of values and is given by the integral of this variable's PDF over that range. Since the distribution of allelic positions is not known, we use a kernel density estimation (KDE) approach which is a non parametric way to estimate the PDF of a random variable. A KDE for the function f is

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

where K is a non negative function called the kernel and h is a smoothing parameter called bandwidth. The selection of kernel has limited impact on optimal PDF estimation so for the purposes of this study, we use the Gaussian kernel which is defined as

$$K(x) = 2\pi^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}$$

and the Epenchinkov kernel which is defined as

$$K(x) = \frac{3}{4}(1 - x^2)$$

The bandwidth of the kernel can strongly influence the estimate obtained from a KDE, this is why bandwidth selection is a crucial step of the process. In most kernel density estimations, the bandwidth is being computed by Silverman's rule of thumb approximations where:

$$h = \widehat{\sigma} n^{-\frac{1}{5}}$$

if a Gaussian kernel is used and

$$h = 2.34\widehat{\sigma} n^{-\frac{1}{5}}$$

for the Epenchinkov kernel, where $\widehat{\sigma}$ is the standard deviation of the sample and n is the sample size. After estimating the PDF of the distances of all allelic classes and the site frequency spectrum, we can calculate the conditional probability from Bayes's Law that best describes a data set that is influenced by both selection and bottleneck as well.

2.3 Analysis in two dimensions

For the next part, we wanted to incorporated the signature of LD into our model. The novelty of our method is that instead of using each allelic class independently as 1, 2, ... N-1, we consider the occurrence of two consecutive SNP's as an single allelic class. By creating $\frac{N(N-1)}{2} - 1$ unique allelic pairs we calculated their frequencies and referred them as 2D-SFS (1 - 1, 1 - 2, ..., (N - 1) - (N - 1)). The distance of these pairs was the distance of the closest SNP to the candidate selection site. Afterwards, for each pair of SNP's we followed the same procedure as described before, estimating their probability densities function.

2.4 Composite Likelihood Ratio Test

Although we have already mentioned the fundamentals of a CLR test, we will describe it in this section in accordance with the way we perform it in a sample. In this study, we have two hypotheses: samples are under selection indicating signatures of selective sweeping or samples are under neutrality referred as L_0 and L_1 respectively. Let's assume a data set with n individuals. At first we need to remove all monomorphic sites. Next, depending on the dimension, we calculate the site frequency spectrum in 1D or 2D, where it denotes the frequencies of the neutral model. For the sweep model we estimate the probability density function from a simulation file the same way we described in a previous section, thus obtaining the probabilities of the selection model. If our data set has X polymorphic sites, we randomly select possible selection sites and perform the CLR in a small fraction of the genome containing x_l from the left side and x_r SNP's. We transform the positions of these SNP's in accordance with the distance from the candidate selection site and then perform the CLR test where:

$$CLR = \frac{\prod_{i=1}^{x_l} P(i|selection) \prod_{j=1}^{x_r} P(j|selection)}{\prod_{i=1}^{x_l} P(i|neutrality) \prod_{j=1}^{x_r} P(j|neutrality)}$$

where $P(\cdot|selection)$ is the probability of a site under selection derived from the simulation data set and $P(\cdot|neutrality)$ is the probability of a site under neutrality which is derived from the SFS of the sample. Since the product of numbers less than 1 can be extremely small, we estimate the logarithm of the CLR test, transforming the above equation to

$$log(CLR) = \frac{\sum_{i=1}^{x_l} log(P(i|selection)) + \sum_{j=1}^{x_r} log(P(j|selection))}{\sum_{i=1}^{x_l} log(P(i|neutrality)) + \sum_{j=1}^{x_r} log(P(j|neutrality))}$$

The length of the fraction is a very crucial parameter for our calculations, because it gives information for the strength of selection. That being the case, for each candidate region, we also check various length parameters that maximize the CLR. The

region with the highest value of the CLR is considered the best candidate selection site.

2.5 SweeD-Sim

Althouh the initial method was implemented in Python the promising outcome from our results has led us to implement the entire procedure into a fully functioning software tool in C, called SweeD-Sim.SweeD-Sim is more time efficient and can create simulation and obtain the probabilities faster than Python. With SweeD-Sim, the user can test if his sample contains regions that have been under selection , using patterns of simulated data. More specifically, the user can specify the demographic parameters of his samples, and then SweeD-Sim will create a simulated data set, obtain the probabilities under selective sweep with kernel density estimation and perform the CLR test for a number of candidate loci across the genome. SweeD-Sim maximizes the log likelihood depending on the distance from the selection site, thus determining the length of the selective sweep. With the grid parameter, SweeD-Sim takes a specified number of candidate regions for which computes their CLR results. The region with the highest CLR is returned, among with the scores of every other region as well.

2.6 Specificity and Sensitivity Analysis

After the implementation of our method, we wanted to determine whether our tool can efficiently manage to distinguish samples under selection and samples under neutrality with the same bottleneck history. For this purpose, we used the ms and mssel simulation tools to create test samples for the evaluation. We examined the results in two ways. First, we wanted to record if in the selection data, where we specified the selection site, the SweeD-Sim tool could accurately predict this specific loci. Next, we wanted to record the specificity and sensitivity of our method. For this purpose we constructed a receiver operating characteristic (ROC) curve, a graphical plot that illustrates the diagnostic ability of a binary classifier system (here Neutrality and Selection) as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) meaning that samples are under selection and SweeD-Sim defined them under selection as well, against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity where the false-positive rate can be calculated as $1-$ specificity. Moreover, we can calculate the area under the curve (AUC) from the ROC curve, which represents the probability that our method will rank a random sample under selection more highly than a random sample under neutrality. In general, an area of 1 represents a perfect test; an area of .5 represents a worthless test. A scientific guide for classifying the accuracy of a test is usually the following:

.90-1 = excellent (A) .80-.90 = good (B) .70-.80 = fair (C) .60-.70 = poor (D) .50-.60 = fail (F)

2.7 Comparison with SweeD

For the last part of this analysis, we scanned the same samples under selection and neutrality with SweeD in order to distinguish where our method can outperform SweeD in data sets where demography strongly affects the distribution of SNP's. As

we mentioned in the introduction, SweeD uses a mathematical equation to compute the CLR test and is Since SweeD uses the one dimensional SFS, the results in this comparison have only incorporated the 1D analysis .

Chapter 3

Results

3.1 Site Frequency Spectrum Patterns

For the first part of our analysis we wanted to check the frequencies of allelic classes in our simulated data in different bottleneck scenarios. The results can be shown in Figure 3.1. As expected, in the scenario of the mild bottleneck, as we move closer to the selection site, frequencies of the outer classes tend to be increased, a result of recombination and selective sweep. In the case of the bottleneck scenario with extreme severity we see very strange patterns and the neutral model has no difference to the selection one.

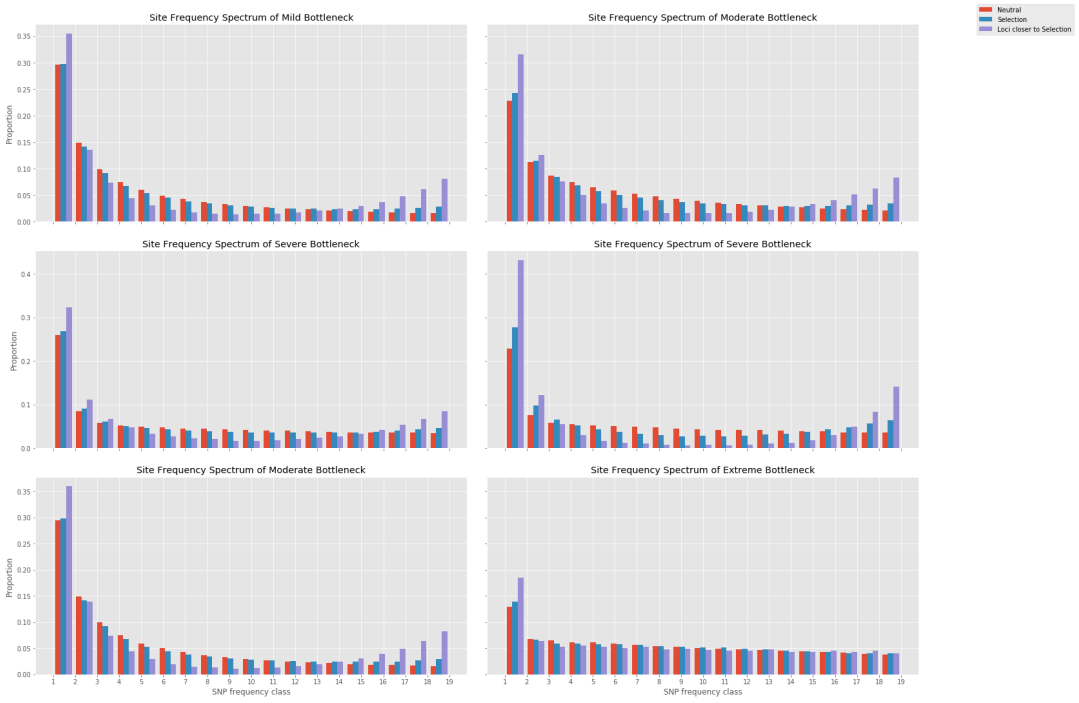


FIGURE (3.1) Site Frequency Spectrum of different Bottleneck Scenarios of samples under neutrality and under selection. Both neutral and selection samples have similar patterns, while loci closer to selection tends to show a preference to low and high allelic classes.

More over, the SFS of both neutral and selection models share similar patterns, thus confirming that probabilities for the neutral model on the CLR test can be estimated from the data set.

3.2 Composite Likelihood Ratios in Different Bottleneck Scenarios

We performed our implementation to both neutral and selection samples with the same demographic parameters in one and two dimension respectively. For this part we performed the CLR test to data set with mild bottleneck (Model 1), moderate to severe bottleneck(Models 24, 36) and a data set with severe bottleneck (Model 60) for a specific number of SNP's surrounding the candidate selection loci. In the following figure, we demonstrate the results in the form of boxplots in figure 3.2.

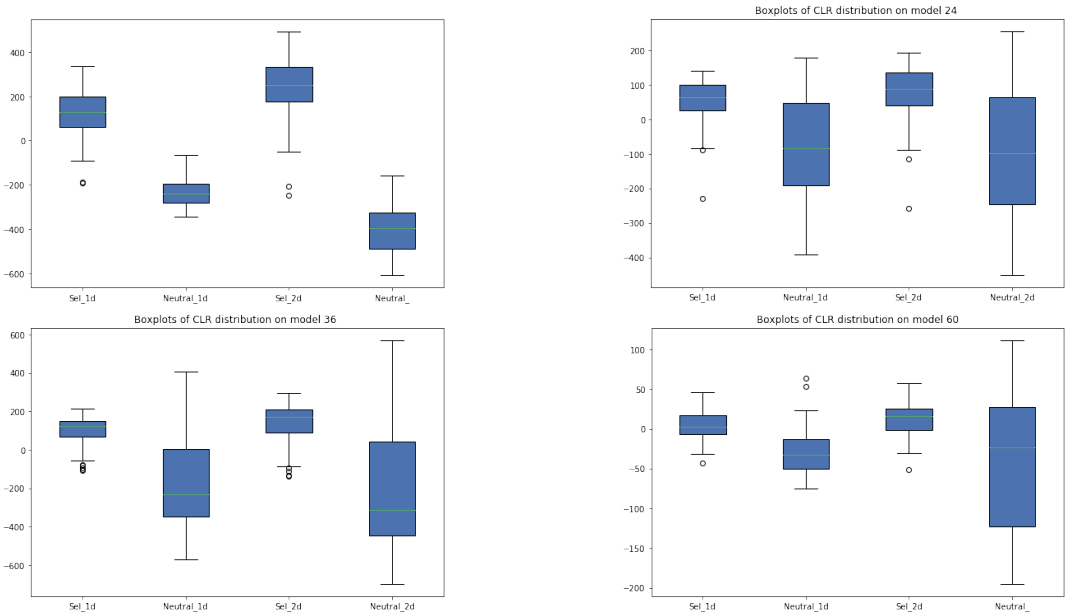


FIGURE (3.2) Boxplots of CLR test results in loci of specific length in 1D and 2D in 150 samples.

For the models of mild and moderate bottlenecks we have a clear differentiation in both 1D and 2D for selection and neutrality with a really small overlap. For the model of severe bottleneck the approach in 1D also shows difference in selection and neutrality while in 2D we have some overlap between the two states. The ability of our method to predict the state of a sample can be shown in the ROC curves in Figure 3.3.

The area under the curve of each sample can be shown in the table 3.1. Unfortunately, since the calculation for the 2D model is quite time consuming, the number of samples for these estimation is small (150 samples) which can maybe lead to the under performance of the 2D approach. In this first approach, we do not have any maximization of the likelihood, thus making the distinguish between neutral and selection data more clear. The fact that all our test are above 60% is an indicator that our method can be used for detection of selective sweeping.

3.3 Evaluating Performance of SweeD-Sim compared with SweeD

In this section we demonstrate the results of SweeD-Sim and SweeD in data sets with moderate bottleneck (Model 20), moderate to severe bottleneck(Models 24, 36)

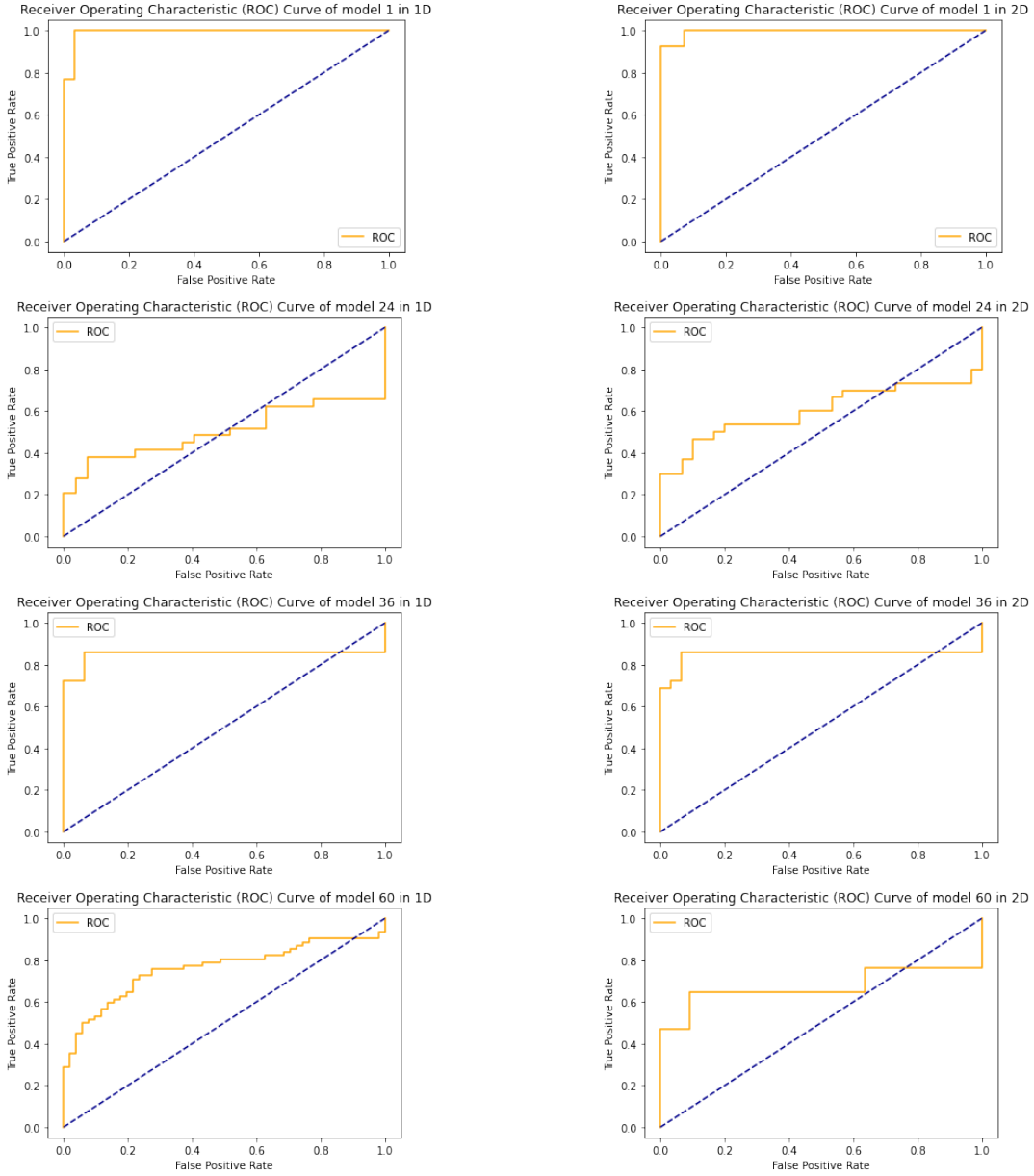


FIGURE (3.3) ROC curves in different bottleneck scenarios in 1D and 2D in 150 samples

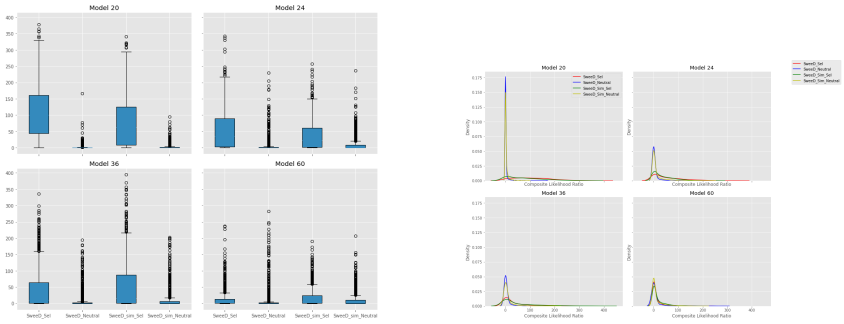
and severe bottleneck (Model 60). Both tools maximize the logarithm of CLR in accordance with the distance from the candidate selection site. The results of both SweeD and SweeD-Sim are examined in two ways. At first, since in our data sets we know the exact location of selection, we compare the CLR of the region closer to the selection. For the neutral data, we take the result of the same region, although since their is no selection, we select this specific region for purposes of comparison. The distribution of the results can be show in Figure 3.4 and their ROC curves and AUC percentages in Figure 3.5 and Table 3.2 respectively.

For the second part we extracted the maximum log likelihood regardless of the position. The distribution of the results can be show in Figure 3.6 and their ROC curves and AUC percentages in Figure 3.7 and Table 3.3 respectively.

Although the results of SweeD-Sim's AUC values are smaller than those of SweeD,

Area Under the Curve (%)		
Model	1D	2D
Model 1	99	99
Model 24	62	60
Model 36	85	85
Model 60	76	67

TABLE (3.1) Percentage of area under the curve for different bottleneck models in 1D and 2D. In all cases the 1D approach outperforms slightly the 2D approach



(A) Boxplot of CLR results of different scenarios in the site of selection

(B) Density Plot of CLR results of different scenarios in the site of selection

FIGURE (3.4) The Boxplot and Density plot of CLR results of different scenarios in the site of selection strongly indicate that both tools can insulate selection from neutrality with a little overlap

Area Under the Curve (%)		
Model	SweeD	SweeD-Sim
Model 20	97	86
Model 24	80	71
Model 36	70	67
Model 60	60	63

TABLE (3.2) Percentage of area under the curve for different bottleneck models using SweeD-Sim and SweeD

their differences are not that consequential to declare our tool as not applicable. The case of the the greater maximum CLR in model 60 in SweeD-Sim, where we have a severe case of bottleneck, is also a strong indicator that our approach can exhibit better results in severe demographic scenarios, where preexisting tools fail to distinguish neutrality from selection.

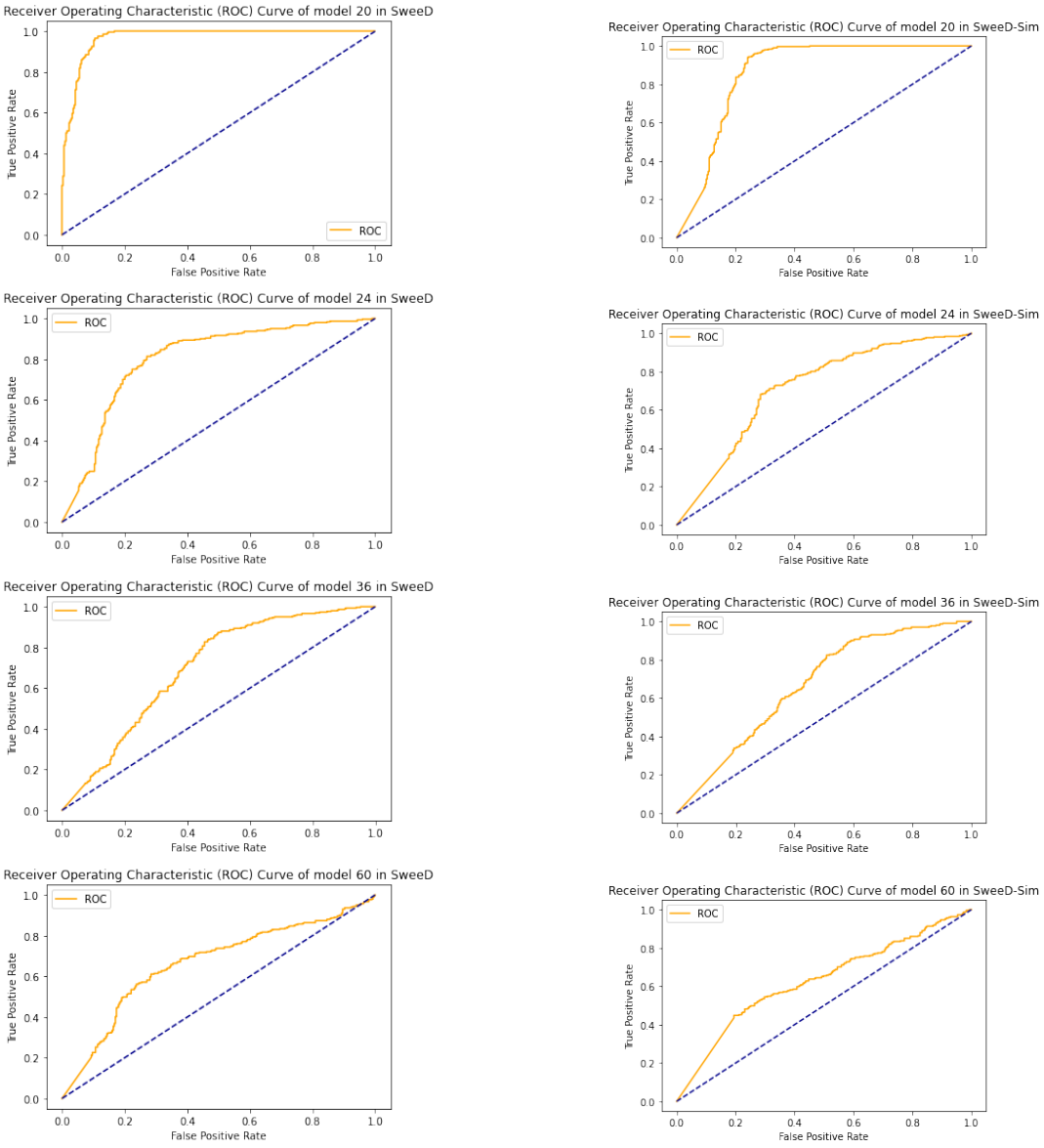
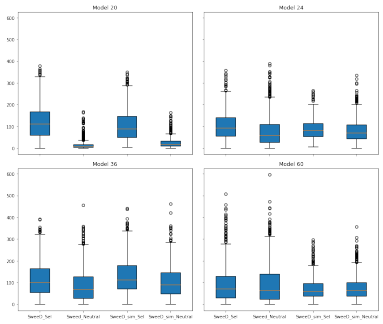


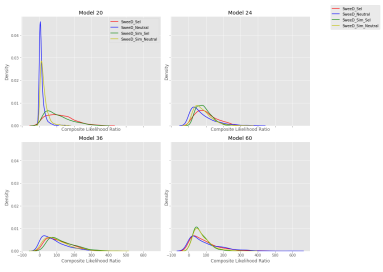
FIGURE (3.5) ROC curves in different bottleneck scenarios with SweeD (left side) and SweeD-Sim (right side). In model 20, the curve tends to move away from the $y = x$ line, indicating a strong performance from both tools, while in models 36 and 60 the line tends to reach it indicating a poor performance

Area Under the Curve (%)		
Model	SweeD	SweeD-Sim
Model 20	95	92
Model 24	63	56
Model 36	64	63
Model 60	47	52

TABLE (3.3) Percentage of area under the curve for different bottleneck models using SweeD-Sim and SweeD



(A) Boxplot of maximum CLR results of different scenarios in the entire loci



(B) Density Plot of maximum CLR results of different scenarios in the entire loci

FIGURE (3.6) Maximum CLR test results of different scenarios. While in model 20 the distinguish is strong, the other models have overlapping CLR values.

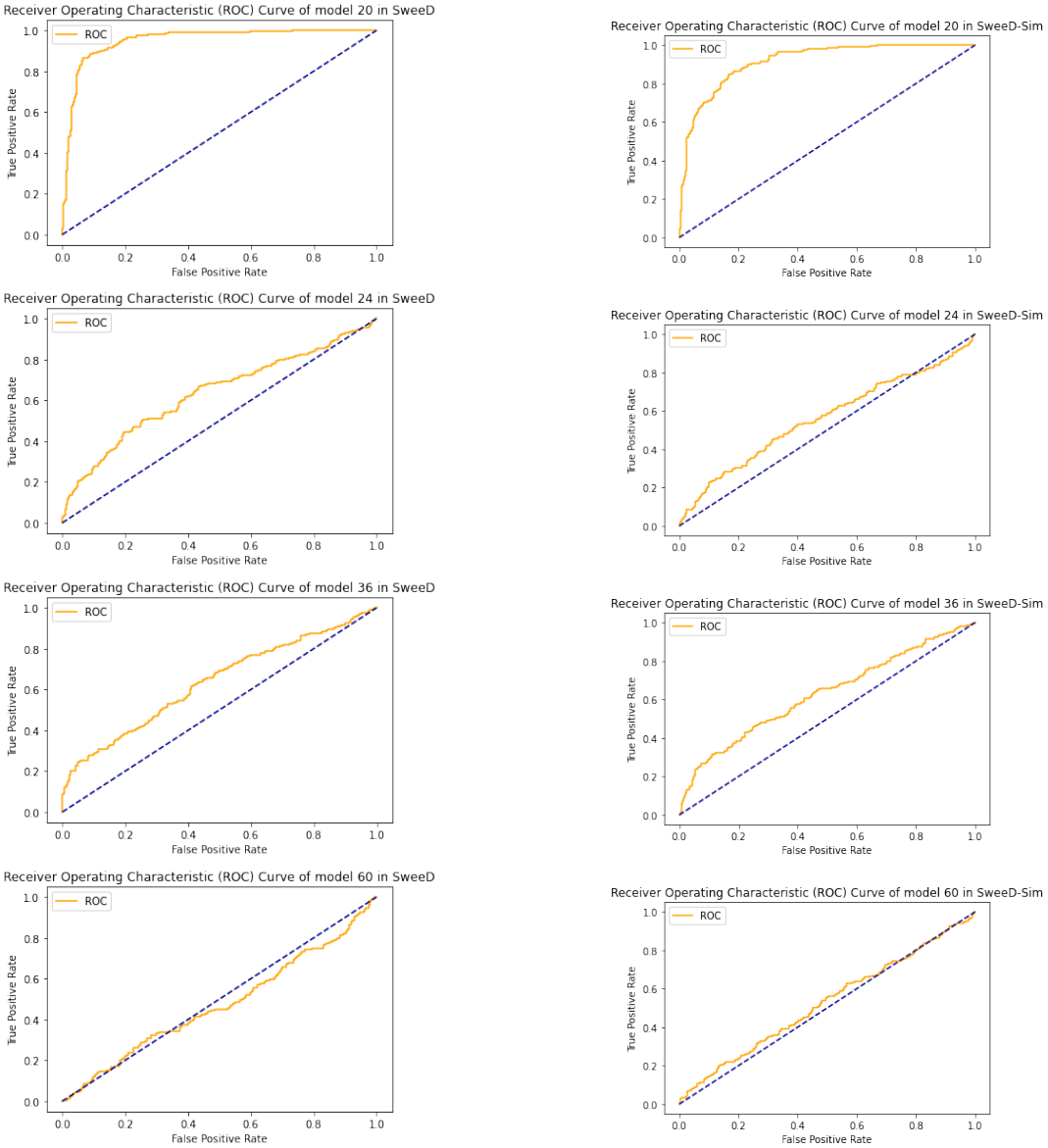


FIGURE (3.7) ROC curves of the maximum CLR alues in different bottleneck scenarios with SweeD (left side) and SweeD-Sim (right side). In model 20, the curve tends to move away from the $y = x$ line, indicating a strong performance from both tools, while in the other models the line tends to reach it, indicating a poor performance

Chapter 4

Discussion

Selective sweeps and their well recognized signatures have been heavily utilized for detecting sites under selection. A variety of tools have been implemented which can detective sweeps with high confidence if the demographic history of a population is not very complicated. However, in the case of populations with complex demographies that share almost identical signatures, the distinguish between sites under selection and sites under neutrality remains a challenging task. In this manuscript we tried to approximate this issue with the usage of simulation and take advantage of the non parametric kernel density estimation in order to approach each population separately, based entirely on its own history, in order to determine if its patterns are a result of demography, selection, or both.

Regarding the results, our method can successfully recognise patterns of selective sweeps thus making it quite useful in detecting selection. In cases of mild bottlenecks SweeD-Sim has a high accuracy, similar to SweeD, a wildly used tool for selective sweep detection. In moderate bottleneck scenarios, SweeD has a higher accuracy than SweeD-Sim but their differences in their values are not disparate enough to discard it. More over, SweeD-Sim seems to have better results in the case of severe bottleneck. Taking into account that this is a primal study which utilizes simulations instead of mathematical equations, the results are indeed quite promising and the method can be further explored. First the size of the simulation is still under consideration. In this study we used 1000 simulations but this number is still under question, whether is it large or small enough to provide the selection model. In addition, we can take a step further to investigate the strength of this approach to other demographic scenarios that its patterns differ from the expected signatures that other tools can detect with high confidence. In the 2D approach our results show that it needs further investigation in terms of efficient implementation so we can analyze a larger number of samples and obtain more significant results.

In conclusion, this work provides a new approach to approximate difficult evolutionary processes not mathematically, but with the use of simulations and kernel density estimators, creating a useful tool for selective sweep detection. We hope that a further development of this implementation in terms of efficiency will make it even more accurate in populations with complex patterns, thus providing a tool that can overcome problems that all mathematical-based tools have faced over the years.

Bibliography

[1] J. M. Smith and J. Haigh, "The hitch-hiking effect of a favourable gene," *Genetics Research*, vol. 23, no. 1, pp. 23–35, 1974.

[2] B. Charlesworth, M. Morgan, and D. Charlesworth, "The effect of deleterious mutations on neutral molecular variation.," *Genetics*, vol. 134, no. 4, pp. 1289–1303, 1993.

[3] W. Stephan, "Selective sweeps," *Genetics*, vol. 211, no. 1, pp. 5–13, 2019.

[4] M. Aguade, N. Miyashita, and C. H. Langley, "Reduced variation in the yellow-achaete-scute region in natural populations of drosophila melanogaster.," *Genetics*, vol. 122, no. 3, pp. 607–615, 1989.

[5] W. Stephan and C. H. Langley, "Molecular genetic variation in the centromeric region of the x chromosome in three drosophila ananassae populations. i. contrasts between the vermilion and forked loci.," *Genetics*, vol. 121, no. 1, pp. 89–99, 1989.

[6] A. J. Berry, J. Ajioka, and M. Kreitman, "Lack of polymorphism on the drosophila fourth chromosome resulting from selection.," *Genetics*, vol. 129, no. 4, pp. 1111–1117, 1991.

[7] D. J. Begun and C. F. Aquadro, "Levels of naturally occurring dna polymorphism correlate with recombination rates in d. melanogaster," *Nature*, vol. 356, no. 6369, pp. 519–520, 1992.

[8] N. L. Kaplan, R. R. Hudson, and C. H. Langley, "The" hitchhiking effect" revisited.," *Genetics*, vol. 123, no. 4, pp. 887–899, 1989.

[9] T. Wiehe and W. Stephan, "Analysis of a genetic hitchhiking model, and its application to dna polymorphism data from drosophila melanogaster.," *Molecular Biology and Evolution*, vol. 10, no. 4, pp. 842–854, 1993.

[10] N. H. Barton, "The effect of hitch-hiking on neutral genealogies," *Genetics Research*, vol. 72, no. 2, pp. 123–133, 1998.

[11] J. H. Gillespie, "Genetic drift in an infinite population: the pseudohitchhiking model," *Genetics*, vol. 155, no. 2, pp. 909–919, 2000.

[12] J. M. Braverman, R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan, "The hitchhiking effect on the site frequency spectrum of dna polymorphisms.," *Genetics*, vol. 140, no. 2, pp. 783–796, 1995.

[13] J. C. Fay and C.-I. Wu, "Hitchhiking under positive darwinian selection," *Genetics*, vol. 155, no. 3, pp. 1405–1413, 2000.

[14] P. Pavlidis and N. Alachiotis, "A survey of methods and tools to detect recent and strong positive selection," *Journal of Biological Research-Thessaloniki*, vol. 24, no. 1, pp. 1–17, 2017.

[15] Y. Kim and W. Stephan, "Detecting a local signature of genetic hitchhiking along a recombining chromosome," *Genetics*, vol. 160, no. 2, pp. 765–777, 2002.

[16] M. Kimura, "Theoretical foundation of population genetics at the molecular level," *Theoretical population biology*, vol. 2, no. 2, pp. 174–208, 1971.

[17] J. D. Jensen, Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante, "Distinguishing between selective sweeps and demography using dna polymorphism data," *Genetics*, vol. 170, no. 3, pp. 1401–1410, 2005.

[18] R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante, "Genomic scans for selective sweeps using snp data," *Genome research*, vol. 15, no. 11, pp. 1566–1575, 2005.

[19] P. Pavlidis, D. Živković, A. Stamatakis, and N. Alachiotis, "Sweed: likelihood-based detection of selective sweeps in thousands of genomes," *Molecular biology and evolution*, vol. 30, no. 9, pp. 2224–2234, 2013.

[20] M. Przeworski, "The signature of positive selection at randomly chosen loci," *Genetics*, vol. 160, no. 3, pp. 1179–1189, 2002.

[21] R. R. Hudson, "Testing the constant-rate neutral allele model with protein sequence data," *Evolution*, pp. 203–217, 1983.

[22] J. F. C. Kingman, "The coalescent," *Stochastic processes and their applications*, vol. 13, no. 3, pp. 235–248, 1982.

[23] R. R. Hudson and N. L. Kaplan, "The coalescent process in models with selection and recombination.," *Genetics*, vol. 120, no. 3, pp. 831–840, 1988.

[24] C. Becquet and P. Andolfatto, "Signatures of a population bottleneck can be localised along a recombining chromosome," tech. rep., Tech. rep. http://przeworski. uchicago. edu/cbecquet/MasterThesis. pdf, 2003.

[25] R. R. Hudson, "Generating samples under a wright–fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, pp. 337–338, 2002.