Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

_____ _____

Yuanbo Song Date

Investigating the Role of Spatial Structure in Genetic Hitchhiking and Sweep Detection

By

Yuanbo Song
Master of Science

The Department of Physics

Daniel Weissman
Advisor

Daniel Sussman
Committee Member

Laura Finzi
Committee Member

Ilya Nemenman
Committee Member

Accepted:

Kimberly Jacob Arriola
Dean of the James T. Laney School of Graduate Studies

Date

Investigating the Role of Spatial Structure in Genetic Hitchhiking and Sweep Detection

By

Yuanbo Song

Advisor: Daniel Weissman

An abstract of
A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in The Department of Physics
2024

Abstract

Investigating the Role of Spatial Structure in Genetic Hitchhiking and Sweep Detection
By Yuanbo Song

A "selective sweep" occurs when a beneficial allele (a variant form of a gene) rapidly increases in frequency and becomes common in the population. This process causes "genetic hitchhiking", in which some nearby genetic variants also increase in frequency because they are statistically associated with the beneficial allele in the population. Traditionally, studies on selective sweeps have been in the context of well-mixed populations, in which every individual has an equal opportunity of interaction and reproduction. However, many real-world scenarios involve spatially structured populations where individuals only interact locally. This raises the pertinent question: How does spatial structure influence the detection and interpretation of selective sweeps? Our first step is to implement selective sweeps with spatial structure in the `msprime` simulation software package. Then we can use the simulation data to assess the impact of spatial structure on standard methods used to detect selective sweeps.

Investigating the Role of Spatial Structure in Genetic Hitchhiking and Sweep Detection

By

Yuanbo Song

Advisor: Daniel Weissman

A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in The Department of Physics
2024

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Daniel Weissman, for his invaluable guidance and support throughout my research journey. His willingness to provide me with the opportunity to join his group and the freedom to pursue my interests has been instrumental in my growth as a researcher. His mentorship, coupled with the flexibility he offered, allowed me to explore my potential and develop my skills in a nurturing environment.

I am also immensely grateful to my group members, who have been a constant source of support and inspiration. Their attendance at my practice sessions and their constructive feedback have significantly contributed to my personal and professional development. Their camaraderie and collaborative spirit have made my research journey an enriching and enjoyable experience.

I would like to extend my heartfelt thanks to my committee members, Dr. Daniel Sussman, Dr. Laura Finzi, and Dr. Ilya Nemenman, for their time, dedication, and willingness to participate in my qualifier and thesis defenses. Their insightful feedback and honest opinions have helped me identify areas for improvement, which are invaluable for both my academic pursuits and personal growth.

To everyone who has been a part of my academic journey, thank you for your encouragement, insights, and friendship. Your contributions have been invaluable, and I am deeply appreciative of the role each of you has played in my success.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Genetic hitchhiking

In a population, when a beneficial mutation appears, one that gives an individual an advantage in survival or reproduction, it can spread rapidly due to natural selection, a process known as a selective sweep [40]. This beneficial mutation is located on a stretch of DNA that also contains neutral genes. As the frequency of the beneficial mutation increases due to natural selection, the entire stretch of DNA, including these neutral genes, also becomes more common in the population. This is because recombination, the process that shuffles genetic material during meiosis, struggles to separate closely linked genes. Consequently, neutral genes located near the beneficial mutation are "hitchhiked" along with it, increasing in frequency as well. However, other versions of these neutral genes, present in different stretches of DNA that did not have the beneficial mutation, may decrease in frequency or even be eliminated. This phenomenon is known as "genetic hitchhiking" [39, 25].

This reduction in diversity around the beneficial allele persists as a distinguishable genetic signature for an extended period after the sweep has ended. This signature is

the basis for many methods used to infer sweeps [31, 33, 9, 22, 37].

A well-documented example of a selective sweep in humans is the evolution of lactose persistence, as discussed by Tishkoff et al. [44]. Typically, humans lose the ability to digest lactose, the main sugar in milk, as they grow into adulthood. However, in populations where milk consumption became a significant part of the diet, individuals with mutations that allowed them to continue digesting lactose into adulthood had a survival advantage. The gene responsible for lactose persistence became the beneficial allele in this context, and its frequency increased in these populations due to natural selection. As the lactose persistence gene spread, it likely brought along nearby neutral genes on the same stretch of DNA, a phenomenon that exemplifies genetic hitchhiking. The selective pressure for lactose tolerance in these populations was strong enough to sweep the beneficial allele and its neighboring genetic variants through the gene pool, leaving a distinct genetic signature that can be traced back through our evolutionary history.

To offer a visual perspective on this, let us turn our attention to Figure 1. Here, we represent genetic diversity using a metric called heterozygosity. In simple terms, heterozygosity [35] measures the variability of genes at a particular locus in a population. The graph denotes a decline in heterozygosity around a specific point, marked as 0 Morgan (a unit measuring genetic linkage), suggesting that selection is centered around this point.

In genetics, a Morgan (abbreviated 'M') [43] is a unit of measurement that quantifies the recombination fraction between genes on a chromosome, effectively representing genetic map distance. The maximum possible value of genetic linkage is 0.5 Morgan, which occurs when two genes are unlinked. This means that there is a 50% chance of recombination between the two genes during the formation of reproductive cells (meiosis).

Heterozygosity is a metric used to quantify genetic diversity at a specific locus

within a population. It is calculated as the probability that two randomly chosen alleles from the population at that locus are different. Mathematically, heterozygosity (H) can be expressed as:

$$H = 1 - \sum_{i}(p_i^2)$$

where p_i is the frequency of the i-th allele at the locus. High heterozygosity indicates a greater variety of alleles, while low heterozygosity suggests a predominance of one allele.

The figure 1 illustrates a decline in heterozygosity around a specific point, marked as 0 Morgan. This particular data, generated through MSMS simulations and visualized using Matplotlib in Python, captures the essence of the genetic hitchhiking effect in a quantitative manner.

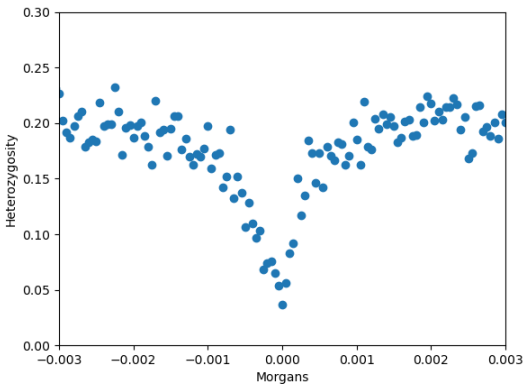

Figure 1: An average heterozygosity graph of the region under selection from the simulation.

1.2 Properties of Selective Sweeps

Selective sweeps leave behind distinct patterns in the genome, marked by a reduction in genetic diversity around the region of the beneficial allele. To understand these

patterns, we need to examine the key characteristics of selective sweeps.

The selective locus specifies the precise genomic location of the beneficial mutation, revealing which genomic regions are under selection. Next, the selective strength represents the intensity of a beneficial mutation. Highlights the mutation that improves the fitness of the population. Selective sweeps can be categorized as either "hard" or "soft" [13]. A hard sweep arises from a single beneficial mutation spreading through the population, typically resulting in a reduction in genetic diversity around the selected region. This is because all individuals with the beneficial allele have inherited it from the same original mutant, along with a similar set of nearby genetic variants.

On the contrary, a soft sweep occurs when multiple separate beneficial mutations, occurring independently, spread through the population. This type of sweep tends to preserve higher genetic diversity around the beneficial allele region, as the beneficial allele originates from multiple individuals with different surrounding genetic material. The term "soft" is used to describe this scenario because the sweep is not as "hard" or strong in its effect on reducing genetic diversity. In a soft sweep, the genetic signature of selection is more subtle, and the genetic diversity around the beneficial mutation is not as sharply reduced.

Distinguishing between hard and soft sweeps is crucial for understanding the selection time scale and genetic variation within populations [12, 28]. The type of sweep can indicate the relative timing of coalescence and the initiation of the selective sweep. In a hard sweep, coalescence, or the merging of lineages in a genealogical tree, occurs after the onset of the selective sweep, making it a relatively recent event. Conversely, during a soft sweep, coalescence takes place before positive selection begins, resulting in a higher level of genetic diversity post-sweep. In summary, hard and soft sweeps impact on genetic diversity and the timing of coalescence differ significantly.

1.3 The need to incorporate spatial structure in models of hitchhiking

Selective sweeps, processes through which beneficial genetic mutations become more common within a population, show distinct dynamics in well-mixed versus spatially structured populations [29, 40, 23, 46, 7, 5, 2].

In the well-mixed model, where each individual equally interacts and produces offspring with any other member, beneficial mutations typically spread according to the logistic growth model. This type of population has traditionally been the primary focus of hitchhiking studies [39, 19].

In contrast, many real-world populations possess spatial structures, meaning individuals mainly engage with their nearby members. In a spatially structured population, beneficial mutations propagate more gradually through traveling waves rather than the standard logistic growth [11]. In a 1D spatially structured population, the growth of these beneficial mutations is linear. In a 2D spatially structured population, the growth adopts a quadratic pattern. As outlined by Barton et al. [4] and reinforced by Min et al. [29], a selective sweep in a spatially structured population leads to a lesser reduction in genetic diversity compared to a identical-sized well-mixed population. The slower pace of selective sweeps in structured environments provides more opportunities for mutations and recombinations to arise, potentially boosting genetic diversity [3].

One way that genetic hitchhiking can be measured is through its effect on the site frequency spectrum (SFS). The SFS provides a summary of how the derived alleles are distributed at varying frequencies within a sample population.

Pictorially, it can be represented as a histogram where the x-axis represents the frequency of an allele in the population, and the y-axis shows how many sites (or loci) have that frequency. Comparison of SFS of well-mixed populations and spatial

populations can indicate the effects of selective sweeps on genetic diversity, offering insight into the impact of selection and the demographic history of populations.

Min et al. [29] discovered that the 1D spatial structure significantly modifies the SFS signatures produced by genetic hitchhiking. Fig.2 displays the SFS produced from hitchhiking in both well-mixed populations and 1D spatial structures.

Specifically, in populations occupying a 1D range, selective sweeps propagate as Fisher waves instead of logistically. In a spatially structured population, the progression of a sweep is more gradual compared to a well-mixed population. This slower spread provides alleles, which begin to hitchhike midway through a sweep, ample opportunity to proliferate and achieve high frequencies. Unlike in well-mixed populations where hitchhiking mainly favors alleles present at the early stages of a sweep, spatially structured populations allow alleles introduced later in the sweep to achieve high frequencies. Furthermore, in these 1D structured populations, recombination plays a much more potent role in restoring genetic diversity when compared to well-mixed populations.

Notably, in spatially structured populations, the Site Frequency Spectrum (SFS) exhibited a long, flat tail, indicating higher genetic diversity after the sweep. This unique SFS could potentially bias estimates of selective sweeps, emphasizing the importance of considering spatial structure when analyzing genetic data.

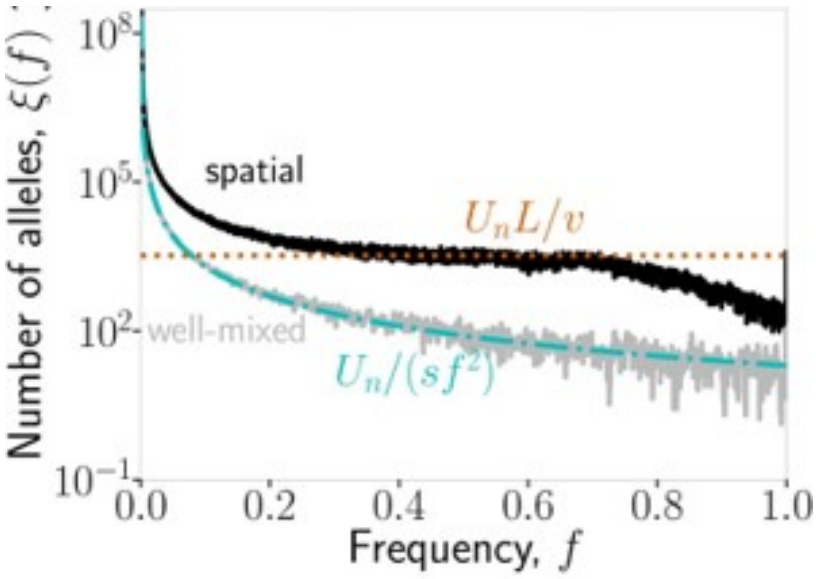

Figure 2: Comparison of Spatial and Well-Mixed Populations on SFS. Both spatial (black curve) and well-mixed (grey curve) populations have the same population size ($N = 10^7$) and selective coefficient ($s = 0.05$). The flat high-frequency tail in the spatial SFS (matching the dotted orange line formula), which stands out higher than that of the well-mixed SFS (matching the dot-dashed cyan curve formula). From [29]

CHAPTER 2

Implementing hitchhiking with spatial structure in *msprime*

Our research goals of studying the impact of spatial structure require simulating numerous replicates of large populations across multiple subpopulations (up to 1000 demes). We need a efficient population genetics simulator that allows us to generate data with realistic population parameters.

2.1 Background

2.1.1 Overview of Simulation Methods

Simulations in population genetics can be broadly categorized into two types based on the directionality of time: forward simulations and backward simulations. The following Figures 3 and 4 illustrate the contrasting directionality of these simulations.

In forward simulations, the process is chronological, beginning with an initial population and proceeding generation by generation into the future. This mirrors the natural progression of evolutionary events as they occur in real populations.

Backward simulations, also known as coalescent simulations, operate in reverse. They start from the present and trace the genealogical lineage of alleles back through time. This reverse engineering of the ancestral tree is efficient because it disregards lineages that do not contribute to the gene pool of the current population. Hence, there is no need to store information about these "dead" lineages, which would be irrelevant for the final genetic data.

Both methods have their distinctive advantages and applications, depending on the research question at hand. Forward simulations are generally more intuitive, as they replicate the actual process of evolution as we understand it. Backward simulations, while less intuitive since they operate in reverse chronological order, can be computationally more efficient. This is because they do not store information about every individual in a population, only those that contribute to the genetic makeup of the sampled individuals at the present time.

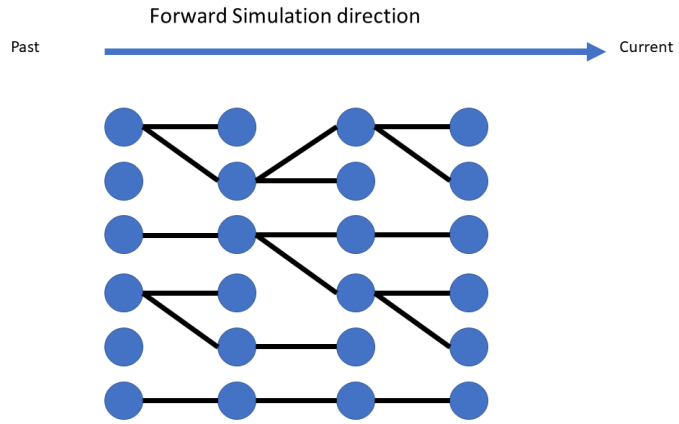

Figure 3: Forward Simulation Direction: This figure demonstrates a forward simulation, where the process unfolds from the past towards the present.

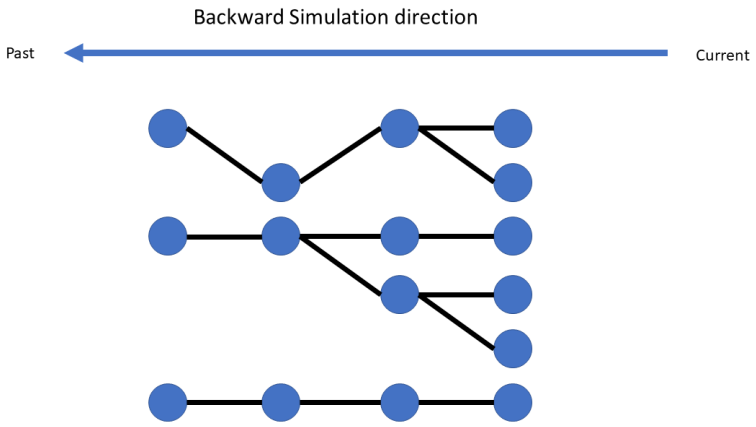

Figure 4: Backward Simulation Direction: In contrast to Figure 3, this diagram represents a backward coalescent simulation, where time flows from the present to the past. The nodes symbolize sampled individuals or genes, and the lines depict ancestral lineages. The simulation only retains lineages that contribute to the current gene pool, disregarding those that do not leave descendants, thereby optimizing data storage.

2.1.2 Simulation Packages

MSMS is a mature and widely used genetic simulation package recognized for its comprehensive functionality. It supports a variety of evolutionary scenarios, including the ability to simulate selective sweeps within spatially structured populations. MSMS adeptly combines forward and backward simulations, allowing for a detailed examination of evolutionary processes [16]. Despite its extensive capabilities, MSMS is limited by its computational efficiency, particularly when simulating large populations across many demes, which makes it less practical for large-scale studies.

In contrast, *msprime* is a more recent addition to the coalescent simulation toolkit. It has rapidly gained popularity due to its efficient and scalable algorithm. With linear time complexity and an innovative approach to data storage, *msprime* performs large-scale simulations with remarkable speed, making it stand out among coalescent simulators [20, 21, 30, 6]. Despite these advantages, *msprime* currently lacks the functionality to simulate selective sweeps in spatially structured populations, offering

this feature only within neutral models. This limitation highlights a development opportunity for *msprime* to accommodate more complex spatial structures in the future.

2.2 Research Question

How can we implement selective sweeps within spatial structures in the *msprime* simulation package?

2.3 Current Progress

2.3.1 Generating beneficial Allele Trajectory

Msprime simulates selective sweeps using a backward-in-time approach. It typically starts with a population that has already fixed a beneficial allele and then traces the ancestry of samples backward until it reaches the start frequency of the sweep. However, *msprime*'s standard selective sweep simulation does not incorporate spatial structure, limiting its ability to model gene flow between different demes or populations. Additionally, the *msprime* trajectory generating function does not guarantee a hard sweep, which is the case we want to study.

To overcome these shortcomings, there is a need for a forward-time simulation approach that can capture the dynamic fluctuations of allele frequencies and incorporate spatial structure to model gene flow. Our methodological foundation is based on insights from Min et al. [29], Barton et al. [3] and MSMS [10]. We aim to integrate the strengths of forward and backward simulations.

The forward simulation is designed to model the dynamics of allele frequencies during selective sweeps in spatially structured populations. It starts with defining a number of demes (spatial units), each containing a fixed number of individuals.

The simulation is parameterized by several key factors: the selection coefficient (s) which quantifies the fitness advantage of the beneficial allele, the migration matrix that determines the rates at which individuals migrate between demes, and the total duration of the simulation (tfinal).

At the beginning of the simulation, a beneficial allele is introduced into one of the demes. In each generation, the simulation performs the following steps:

1. **Migration:** Individuals migrate between demes according to the migration matrix. This step updates the allele counts in each deme by accounting for the influx and outflow of alleles due to migration.

2. **Selection:** The allele frequencies in each deme are updated based on the selection coefficient. This step models the increase in frequency of the beneficial allele due to its fitness advantage.

3. **Drift:** The new generation of individuals in each deme is sampled based on the updated allele frequencies. This step introduces stochasticity into the simulation, representing genetic drift.

The simulation continues for a predefined number of generations or until the beneficial allele reaches a certain frequency threshold in the population. If the beneficial allele becomes extinct, the simulation can restart from the introduction of the allele, allowing for the exploration of different evolutionary trajectories. The output of the simulation is a trajectory that records the allele frequencies in each deme over time.

2.3.2 Integrating Forward and Backward Simulation

The software *msprime* is equipped with a built-in backward coalescent simulation. This program allows only one event to occur during each time interval, which can

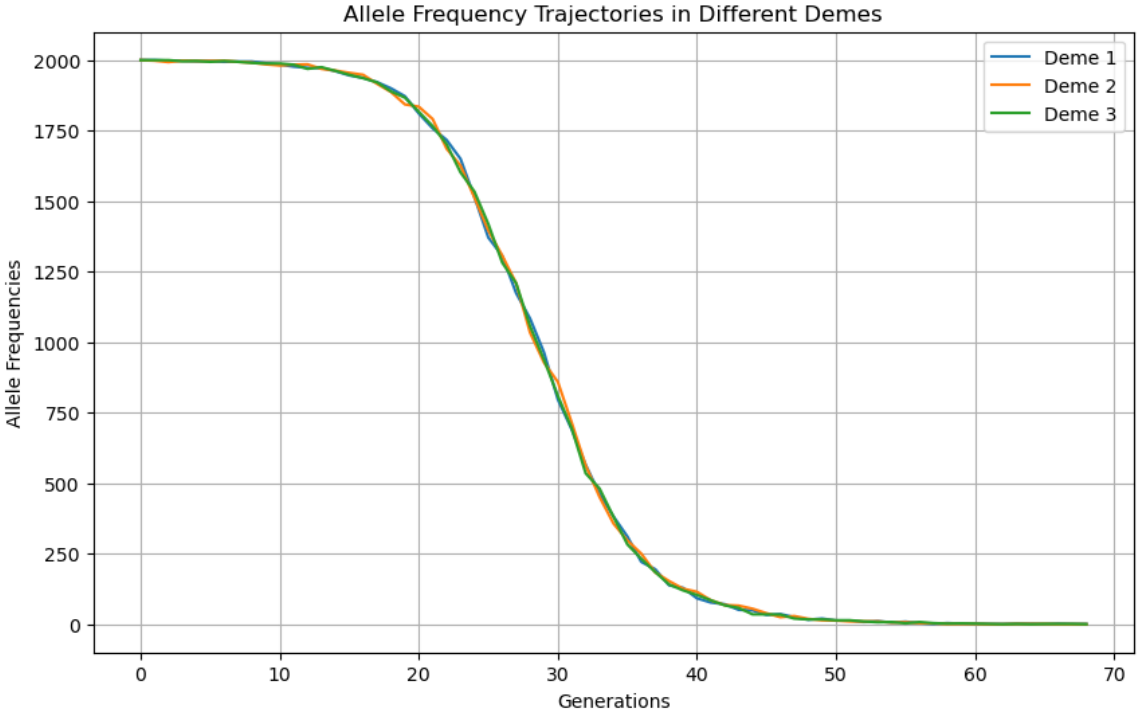

Figure 5: Allele Frequency Trajectories in Different Demes Over Generations. This figure illustrates the results of a forward genetic simulation depicting the allele frequencies in three separate demes (Deme 1, Deme 2, and Deme 3) over 70 generations. Each deme initially contains 2000 individuals (N), with a selection coefficient (s) of 0.25. The simulation's total time limit is 200 generations, although the sweep only takes 70 generations. The number of demes (L) is set to 3, and the migration rate is determined as the inverse of the number of demes (migration_rate = 1/L). The similar trajectories across demes suggest homogenization of allele frequencies due to migration and selection effects.

include processes such as recombination, coalescence, or migration. For this type of simulation, specific information need to be provided, such as the migration matrix, population sizes, sweep trajectory, and time slice. The core of this process is the Hudson simulation function [17], which examines populations and identifies those with ancestors of interest. This function then anticipates the next significant genetic event, be it migration, recombination, or common ancestral. The coalescent history is captured in the tables which are a representation of the genetic data. The detail of the table structure is documented in tskit manual [21].

In our project, we have ensured consistency between the time scales of our forward and coalescent simulations. Both simulations operate in units of generations, aligning the temporal framework of allele frequency tracking and population dynamics. This alignment allows us to focus solely on generation time, eliminating concerns about transitioning between coalescent time and generations.

However, challenges remain. There are a few key decisions we need to make regarding the simulation, such as choosing between discrete and continuous time, as well as discrete allele counts and their continuous frequency counterparts.

The coalescent simulations often conclude with multiple lineages despite the simulation's initiation with a single beneficial allele. This observation could suggest a discrepancy that may stem from the limitations of the deterministic model applied during the initial stages of the allele spread. The Braverman paper [8], which informs this implementation, defines the start of the deterministic frequency trajectory based on the point where the probability of the new mutant's extinction is negligible. However, when the initial frequency is very close to zero, the model may not fully capture the stochastic effects that are pronounced during these early stages of the sweep [18].

On the *msprime* GitHub repository's issue tracker #2242, a developer of *msprime*, Gertjan Bisschop, clarifies that if the provided start and end frequencies are too close to 0 or 1, the deterministic approximation may not accurately reflect the stochastic nature of the genetic sweep. This particularly affects the interpretation of the initial frequency as an absolute lower bound. In practice, while a hard sweep is expected to result in a distribution of neutral lineages with the favored allele concentrated around a single lineage, stochastic elements can cause variations from this expectation.

Given these comments, adjustments may be needed for handling initial conditions and how we simulate the early stages of allele spread, where random effects are strong. Due to these inconsistencies, we cannot directly input the allele frequencies from the forward simulation into the coalescent simulation.

We first write code for a well-mixed population, and then test the simulation to ensure integration is performed correctly. *Msprime*'s setup file contains built-in test protocols, which we will make use of it to evaluate the validity of our coalescent simulation. Once the outputs align, our subsequent step will be to implement migration between subpopulations.

CHAPTER 3

Assessing the Impact of Spatial Structure on Sweep Inference Methods

As we work towards implementing spatial structures in *msprime* simulations, the next logical step is to examine the performance of the existing selective sweep inference methods. The state of the art in selective sweep inference has evolved significantly, with various methods being developed to detect different types of selective sweeps. These methods are used in a wide range of scenarios, from identifying regions of the genome under selection in human populations to understanding adaptive evolution in model organisms. For example, SweepFinder, SweeD and OmegaPlus have been employed to detect selective sweeps in Drosophila and human populations [31, 33, 1, 36, 12, 14]. These tools are critical for researchers aiming to understand the evolutionary history of species.

For example, researchers have used these methods to study the domestication of crops by identifying selective sweeps associated with traits like seed size and plant architecture [15]. In the study by Sattath et al. [36], they shows two distinct types of selective sweeps: a common type with weaker effects and a rarer, stronger type from

Drosophila simulans. In human populations, these methods have been employed to detect regions under selection due to factors like diet, climate, and disease resistance [34]. In conservation biology, identifying selective sweeps can help understand how species adapt to changing environments or human-induced pressures [45].

These methods have traditionally been tested in simple, unstructured populations. However, the presence of spatial structure introduces additional complexities that may affect the performance of these inference methods. Methods like SweepFinder2 and diploS/HIC have been developed to incorporate the inference of selective sweeps in spatially structured populations and other compounding factors [9, 22].

In this chapter, we will outline our approach to evaluate the capability of existing selective sweep inference packages in detecting selective sweeps within spatially structured populations. We plan to explore various configurations, starting with basic linear (1D) setups and progressing to more complex two-dimensional (2D) landscapes.

Min et al. [29] have raised concerns that spatial structure might introduce biases in inferring past evolutionary events. Our goal is to assess the robustness of these methods in the context of spatial complexities. We aim to quantify the impact of spatial structure on the detection power of these methods and explore any potential biases they may introduce in estimates of selective strength and sweep hardness/softness. Understanding and accounting for these effects is crucial for making accurate inferences in evolutionary studies.

At this stage, our analysis is based on preliminary tests and theoretical considerations. As we continue to develop and refine the spatial structure implementation in *msprime*, we will update our evaluation to include more comprehensive empirical results.

Our primary focus is on Sweepfinder2 and diploS/HIC because of their unique methodologies and the distinct properties of selective sweeps that they infer. We have chosen Sweepfinder2 because it uses a likelihood ratio test, giving us a solid statistical

base and helping us measure how well traditional models perform. This enables the determination of both the location and strength of selective events from genetic data.

Kern and Schrider [22], employed a machine learning-based method, termed diploS/HIC, for robust identification of soft and hard sweeps. We plan to access the performance of diploS/HIC for spatial structured population.

The combination of Sweepfinder2 and diploS/HIC allows for a comprehensive analysis of selective sweep.

3.1 Parameter choice

In our simulations, we specifically model a weakly spatially structured population. For such a structure under neutral evolution, the spatial pattern within the population is not immediately apparent. This means the spatial structure has a negligible effect on neutral alleles, influencing only the selective alleles.

We want to determine the spatial structure effect through the hitchhiking pattern. To achieve that, we need to understand relationships between various time scales to find right parameters for the simulation. To break down the relationships, we first need to understand the notations and parameters:

- N: Total population size.
- D: Effective migration or diffusion rate. It informs us about how lineages disperse across the spatial domain.
- L: Length of the spatial range. It defines the distance over which the spatial effects and sweeps.
- s: Selection coefficient of the sweeping allele. A larger s indicates stronger selection in favor of the allele.

- r: The recombination rate between the selected locus and other loci in the genome.

Our parameters of interest include:

- T_{past}: The duration elapsed after the occurrence of a selective sweep.
- T_{coal}: Neutral coalescence time.
- T_{mix}: Dispersal time for a lineage across the range, given by $\frac{L^2}{D}$. This expression is derived from the theory of diffusion [11].
- T_{sweep}: The time taken for a sweep to become fixed.

First, we must ensure that spatial structures exert a negligible impact on neutral alleles. For this to be the case, the rate of spatial mixing must exceed the coalescence rate, leading to the relationship $T_{\text{mix}} \ll T_{\text{coal}} \approx N$. In such conditions, the neutral allele frequency spectrum closely aligns with a well-mixed model [27, 24]. Specifically, the density of mutant alleles at frequency f matches $p(f) \approx \frac{2NU}{f}$ [47].

Simultaneously, it is essential for the spatial structure to influence selective alleles. For sweeps possessing a sufficiently robust selective strength, the duration T_{sweep} is significantly shorter than T_{coal} and potentially T_{mix}. Such sweeps, heavily influenced by spatial structure, progress like a Fisher wave at a speed roughly equal to $\sqrt{Ds}$, making $T_{\text{sweep}} \approx \frac{L}{\sqrt{Ds}}$ [11].

Another key parameter to consider is the recentness of the sweeps for them to be detectable. Only those sweeps recent enough leave traces unerased by genetic drift. This means $T_{\text{past}} \ll T_{\text{coal}}$.

In summary, the hierarchy of time scales for many natural populations is $T_{\text{sweep}} \ll T_{\text{mix}} < T_{\text{past}} \ll T_{\text{coal}}$.

3.2 SweepFinder2

SweepFinder2 [9] is a computational tool that employs the site frequency spectrum (SFS) to identify genomic regions potentially under the influence of recent positive selection. SweepFinder2 provides high accuracy in detecting genomic regions under diverse selective sweep scenarios. Its effectiveness is underscored by its widespread use in various studies [48, 26, 32, 41].

SweepFinder2's output serves two purposes, offering the likelihood ratio (LR) and the α metric. LR is a measure of which model better fits the observed data. This ratio contrasts the fit of the hitchhiking model against the neutral model for the given dataset. A higher LR value translates to a higher confidence in the presence of a selective force at any given point on a genome.

Simultaneously, the alpha metric in SweepFinder2 is related to the selective sweep's strength and is computed as:

$$\alpha = \frac{r \ln(2N)}{s}$$

. The alpha value is inversely proportional to the selective strength. A lower alpha value suggests a stronger selective sweep, as it indicates a higher selection coefficient for a given recombination rate and population size.

Once we have fed the simulation data into SF2, we can obtain a LR and alpha prediction for each genomic window we've specified. We will then visualize these results. Note that the alpha value around selective loci is meaningful; elsewhere, without selective sweeps, the selective strength does not make sense. Since the SFS is produced by our simulation, we already know the selective locus and strength in advance. We can compare the inferred results with the actual values to determine how the spatial structure biases the selective locus and strength.

3.3 diploS/HIC

The diploS/HIC [22] is a method designed for 'Soft/Hard Inference through Classification'.

The diploS/HIC framework is a sweep type classification tool that uses deep learning, specifically a deep convolutional neural network (CNN) architecture, to classify genomic windows based on their evolutionary trajectory.

Instead of feeding raw results from coalescent simulations, like those from msms, directly into the CNN, diploS/HIC converts this raw data into 12 summary statistics. Each of these statistics is calculated for 11 sub-windows in the genomic data, allowing the CNN to focus on the most meaningful patterns without being overwhelmed by raw data noise. The final output is a prediction of the probability of each selective type for every window.

3.4 Preliminary Results

Fig.6 presents scatter plots of both LR and alpha values predicted by SweepFinder2 across two distinct spatial structure settings: the well-mixed and the 10x10 grid scenarios.

Spatial structures should distort the traditional signals of selective sweeps. Some methods might be less adept at deciphering complex spatial scenarios compared to simpler, well-mixed populations.

We anticipate that the increase in the high-frequency tail of the SFS, caused by the spatial structure, might lead inference methods to deduce a lower selective strength compared to that of a well-mixed population. The inference of the selective locus might also be impacted, as a smaller region of the SFS will be affected by the selective sweep.

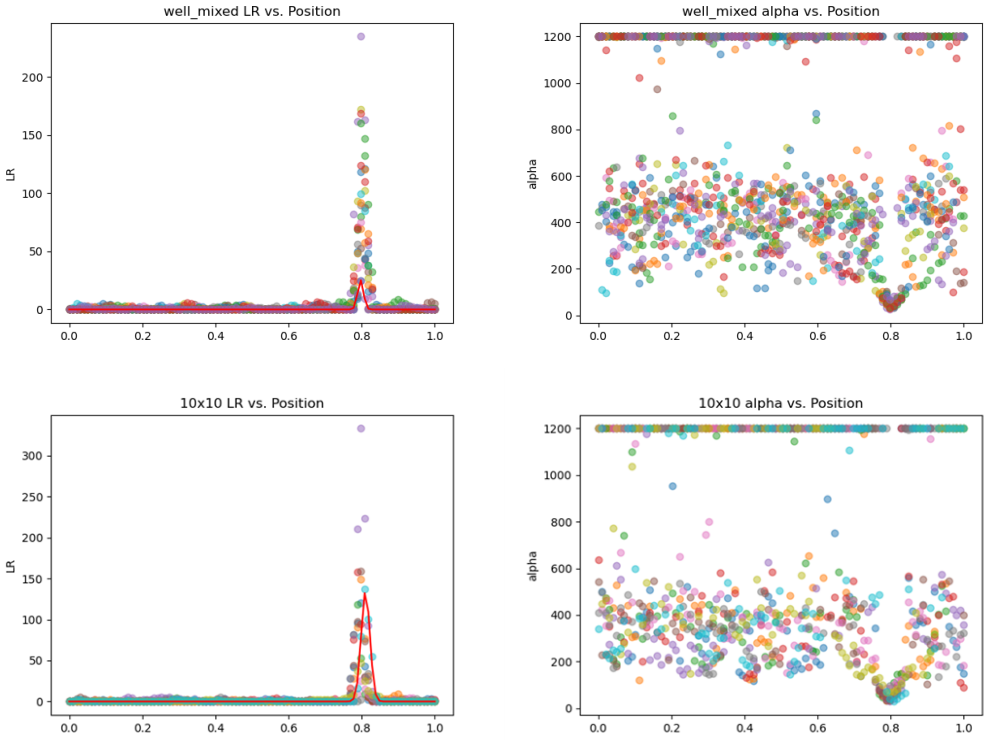

Figure 6: Scatter plots showcasing distinct population structures: well-mixed and spatially structured on a 10x10 grid. The x-axis represents the relative position across the genome, spanning from 0 to 1, while the y-axis depicts the likelihood ratio (LR), with a higher value indicating a better fit for the selective model. The distribution of the alpha value across both scenarios - well-mixed and the 10x10 grid - is also presented. Note that the alpha value around selective loci is meaningful; elsewhere, without selective sweeps, the selective strength does not make sense. The selective locus was set at 0.8.

Quantitatively, the width of the hitchhiking region on the genome serves as an indicator to infer the selective coefficient. This width is approximately represented by $s/\ln(Ns)$. In real-world observations, Tavares et al. [42] identified strong selective sweeps in nature. However, in the case of spatially structured *Antirrhinum majus* population, there was only reductions in diversity observed within narrow genomic windows than what predicted by well-mixed assumption.

We should be able to quantify this potential misclassification in inference methods. We can analyze the results by comparing the predicted sweep type, selective strength,

and selective locus to those from the simulations. However, the details of the impact remain unclear based on the preliminary data we have. As you can see from Figure 6, SweepFinder2 could detect selection with a power of around 0.8. We suspect that we don't have a large enough population, so we moved to implement spatial structure in *msprime* to generate data that meet our weekly structure constraint.

Additionally, Min et al. [29] discussed how the "soft shoulder" phenomenon might lead to misclassification of sweep types in spatially structured populations. The term "soft shoulder" refers to the potential presence of high-frequency recombinant haplotypes around the selected locus, as highlighted by Schrider et al. [38]. This phenomenon mimics multiple initial background haplotypes, which make hard sweep appear softer. In 1D spatially structured populations, these soft shoulders are often more frequent and closer to the swept locus, potentially causing hard sweeps to be mistakenly identified as soft.

Bibliography

[1] Alachiotis, N. and Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, 1:79.

[2] Barton, N. H. (2000). Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 355(1403):1553–1562.

[3] Barton, N. H., Etheridge, A. M., Kelleher, J., and Véber, A. (2013a). Genetic hitchhiking in spatially extended populations. *Theoretical Population Biology*, 87:75–89.

[4] Barton, N. H., Etheridge, A. M., Kelleher, J., and Véber, A. (2013b). Inference in two dimensions: Allele frequencies versus lengths of shared sequence blocks. *Theoretical Population Biology*, 87:105–119.

[5] Battey, C. J., Ralph, P. L., and Kern, A. D. (2020). Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data. *Genetics*, 215(1):193–214.

[6] Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., Quinto-Cortés, C. D., Rodrigues, M. F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A. W., Wong, Y., Gravel, S., Kern,

A. D., Koskela, J., Ralph, P. L., and Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229.

[7] Bradburd, G. S. and Ralph, P. L. (2019). Spatial Population Genetics: It's About Time. *Annual Review of Ecology, Evolution, and Systematics*, 50(1):427–449.

[8] Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., and Stephan, W. (1995). The Hitchhiking Effect on the Site Frequency Spectrum of DNA Polymorphisms. *Genetics*, 140(2):783–796.

[9] DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., and Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics (Oxford, England)*, 32(12):1895–1897.

[10] Ewing, G. and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

[11] Fisher, R. A. (1937). The Wave of Advance of Advantageous Genes. *Annals of Eugenics*, 7(4):355–369. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1937.tb02153.x.

[12] Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLOS Genetics*, 11(2):e1005004. Publisher: Public Library of Science.

[13] Hermisson, J. and Pennings, P. S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352.

[14] Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., and Przeworski, M. (2011). Classic selec-

tive sweeps were rare in recent human evolution. *Science (New York, N.Y.)*, 331(6019):920–924.

[15] Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., Li, M., Fan, D., Guo, Y., Wang, A., Wang, L., Deng, L., Li, W., Lu, Y., Weng, Q., Liu, K., Huang, T., Zhou, T., Jing, Y., Li, W., Lin, Z., Buckler, E. S., Qian, Q., Zhang, Q.-F., Li, J., and Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, 42(11):961–967.

[16] Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

[17] Hudson, R. R. and Kaplan, N. L. (1988). The Coalescent Process in Models with Selection and Recombination. *Genetics*, 120(3):831–840.

[18] Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The Coalescent Process in Models with Selection. *Genetics*, 120(3):819–829.

[19] Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989). The "hitchhiking effect" revisited. *Genetics*, 123(4):887–899.

[20] Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5):e1004842. Publisher: Public Library of Science.

[21] Kelleher, J., Thornton, K. R., Ashander, J., and Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, 14(11):e1006581. Publisher: Public Library of Science.

[22] Kern, A. D. and Schrider, D. R. (2018). diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3 Genes|Genomes|Genetics*, 8(6):1959–1970.

[23] Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777.

[24] Kimura, M. and Maruyama, T. (1971). Pattern of neutral polymorphism in a geographically structured population. *Genetics Research*, 18(2):125–131. Publisher: Cambridge University Press.

[25] Kojima, K.-I. and Schaffer, H. E. (1967). SURVIVAL PROCESS OF LINKED MUTANT GENES. *Evolution; International Journal of Organic Evolution*, 21(3):518–531.

[26] Li, J., Zhang, L., Zhou, H., Stoneking, M., and Tang, K. (2011). Global patterns of genetic diversity and signals of natural selection for human ADME genes. *Human Molecular Genetics*, 20(3):528–540.

[27] Maruyama, T. (1972). Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics*, 70(4):639–651.

[28] Messer, P. W. and Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, 28(11):659–669.

[29] Min, J., Gupta, M., Desai, M. M., and Weissman, D. B. (2022). Spatial structure alters the site frequency spectrum produced by hitchhiking. *Genetics*, 222(3):iyac139.

[30] Nelson, D., Kelleher, J., Ragsdale, A. P., Moreau, C., McVean, G., and Gravel, S. (2020). Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLOS Genetics*, 16(5):e1008619. Publisher: Public Library of Science.

[31] Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11):1566–1575.

[32] Pavlidis, P., Jensen, J. D., and Stephan, W. (2010). Searching for Footprints of Positive Selection in Whole-Genome SNP Data From Nonequilibrium Populations. *Genetics*, 185(3):907–922.

[33] Pavlidis, P., Živković, D., Stamatakis, A., and Alachiotis, N. (2013). SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology and Evolution*, 30(9):2224–2234.

[34] Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley,

D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918.

[35] Samuels, D. C., Wang, J., Ye, F., He, J., Levinson, R. T., Sheng, Q., Zhao, S., Capra, J. A., Shyr, Y., Zheng, W., and Guo, Y. (2016). Heterozygosity Ratio, a Robust Global Genomic Measure of Autozygosity and Its Association with Height

and Disease Risk. *Genetics*, 204(3):893–904.

[36] Sattath, S., Elyashiv, E., Kolodny, O., Rinott, Y., and Sella, G. (2011). Pervasive Adaptive Protein Evolution Apparent in Diversity Patterns around Amino Acid Substitutions in Drosophila simulans. *PLoS Genetics*, 7(2):e1001302.

[37] Schrider, D. R. and Kern, A. D. (2016). S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLOS Genetics*, 12(3):e1005928. Publisher: Public Library of Science.

[38] Schrider, D. R., Mendes, F. K., Hahn, M. W., and Kern, A. D. (2015). Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps. *Genetics*, 200(1):267–284.

[39] Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35. Publisher: Cambridge University Press.

[40] Stephan, W. (2019). Selective Sweeps. *Genetics*, 211(1):5–13.

[41] Svetec, N., Pavlidis, P., and 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(7):1549–1556.

[42] Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., Copsey, L., Elleouet, J., Burrus, M., Andalo, C., Li, M., Li, Q., Xue, Y., Rebocho, A. B., Barton, N. H., and Coen, E. (2018). Selection and gene flow shape genomic islands that control floral guides. *Proceedings of the National Academy of Sciences*, 115(43):11006–11011. Publisher: Proceedings of the National Academy of Sciences.

[43] Teas, H. J., editor (1969). *Genetics and Developmental Biology: The Thomas Hunt Morgan Centennial Symposium*. University Press of Kentucky.

[44] Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39(1):31–40.

[45] Vilà, C., Sundqvist, A.-K., Flagstad, O., Seddon, J., Björnerfeldt, S., Kojola, I., Casulli, A., Sand, H., Wabakken, P., and Ellegren, H. (2003). Rescue of a severely bottlenecked wolf (Canis lupus) population by a single immigrant. *Proceedings of the Royal Society B: Biological Sciences*, 270(1510):91–97.

[46] Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annual Review of Genetics*, 47:97–120.

[47] Wakeley, J., King, L., Low, B. S., and Ramachandran, S. (2012). Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman's Coalescent. *Genetics*, 190(4):1433–1445.

[48] Wilson Sayres, M. A., Lohmueller, K. E., and Nielsen, R. (2014). Natural selection reduced diversity on human y chromosomes. *PLoS genetics*, 10(1):e1004064.