Detecting positive selection using pair of consecutive SNPs

Eirini Christodoulaki

Department of Biology

University of Crete

This dissertation is submitted for the degree of

Master of Science

UoC / IMBB-FORTH November 2015

Acknowledgements

And I would like to acknowledge my supervisors Dr. Pavlidis Pavlos (Institute of Molecular Biology and Biotechnology/Foundation of Research and Technology, Hellas) and Dr. Iliopoulos Ioannis (University of Crete, Faculty of Medicine) who supported this effort with their inexhaustible knowledge and of course, unwavering patience and understanding...

Abstract

An interesting research topic in population genetics is the detection of loci that are targets of positive selection using Single Nucleotide Polymorphism (SNP) data. Many tests have been developed for the identification of genomic signatures of positive selection based on the skew of the distribution of SNP frequencies (Site Frequency Spectrum or SFS) that is caused by selection. There are several methods to detect strong positive selection or in other words, a selective sweep event. Current methods for detecting selective sweeps based on SFS do not take into consideration the linkage disequilibrium in combination with the skew in the allele frequencies in SFS. In this study we describe a model that combines both genomic signatures of positive selection, that is, LD and the skewed patterns of allele frequencies. By extensive simulations we show that is possible to detect a selective sweep by using the joint distribution of the allele frequencies of pairs of polymorphic sites (joint-SFS). Furthermore, we show that using pairs of SNPs, instead of independent sites, results in higher efficiency to detect selective sweeps. Furthermore, we try to find a mathematical formula that calculates the probability of observing a pair of Single Nucleotide Polymorphisms (SNPs) at a certain distance from a positively selected site, using a diffusion approximation.

Table of contents

Chapter 1

Introduction

The major evolutionary force that results in a better adaptation of organisms into a changing environment is positive selection. When a species or a population adapts to its environment, several genes might be under strong selection. Detection of selective sweeps helps population geneticists to identify candidate genes that are affected by strong recent positive selection. To this end, the quantification of variation is important and can be measured by genome scans in samples of individuals. If, for instance, a region in the genome shows low variation, statistical tests help us to decide if a gene under selection is located in the nearby region [15]. But what happens when a beneficial mutation appears in a population? What is a selective sweep and which are the genomic signatures that are observed during a selective sweep? And in the end, how can we detect such patterns of recent, strong, positive selection (selective sweeps) using genomic SNP data?

1.1 Selective sweep and its genomic signatures

When a new beneficial mutation arises in a population and increases its frequency throughout the population due to natural selection, the standing genetic variation in neighboring regions is affected. As a result, three genomic signatures can be observed along the genome, i)reduced level of variability, ii)increased level of linkage disequilibrium, and iii)skewed pattern of allele frequencies [2].

As this beneficial mutation spreads throughout the population and goes to fixation, the adjacent genomic regions that are physically linked to the selected site are dragged also to fixation along with the beneficial mutation. Otherwise, they are discarded together with the less fit alleles during the process that is called genetic hitchhiking [16]. As a result, the level of variability in the region near to the positively selected site is reduced. However, in some individuals, the selected locus can be separated from adjacent loci by recombination.

In this case, there is a decreased strength of hitchhiking as the distance from the selected locus is increased and consequently the variation in this region is increased again. Therefore, recombination is a force that tends to diminish the sweep effect as the genetic distance from the selected site increases[4].

Additionally, as the beneficial allele goes to fixation, the level of linkage disequilibrium, that is produced between linked neutral polymorphisms, increases, due to hitchhiking effect. Specifically, as the beneficial mutation increases in frequency it drags with it the neutral loci that are linked with it and as a result, there is an increase in linkage disequilibrium. Also, during the selective sweep, recombination does not have much time to break the linkage that is produced between neutral polymorphisms and as a consequence, there are alleles with high frequencies and long-range associations between them. [17].

There have been several studies based on patterns of LD. These studies started on 2004 with Kim and Nielsen, who used numerical simulations to study LD. Later, they were extended by Stephan et al.[25] who obtain analytical expressions for measures of LD after a selective sweep. Also, in another study [17] Pffafelhuber and Stephan used the star-like approximation for the genealogy at the selected site in order to describe the patterns of LD . This approximation of the joint genealogy at the two neutral loci, described splits in the wild-type background and showed that it can predict the increase of LD close to a selected site and the elimination of LD between both sides of the selected site.

As a consequence of both genomic signatures of selection that was analyzed above, there is a distortion of the Site Frequency Spectrum in the case of selective sweep [2]. In our study, we are interested in the detection of recent selective sweep events by studying the signature of selection on the distribution of allele frequencies. For this purpose, we use the joint-Site (or 2D) frequency spectrum of two loci. Both, SFS and joint SFS are analyzed extensively below.

1.2 Site frequency Spectrum

Positive selection can be identified between species using divergence data and within a species using polymorphism data. Divergence data are used to identify older selective events, whereas polymorphism data are used to identify recent selective events [11]. In this study, we are interested in recent selective events in a genome, so we use only polymorphism data (it is an intra-species study). In order to detect recent strong positive selection we have to identify regions along the genome that show evidence of a selective sweep. We use the skewed patterns of allele frequencies that are caused by strong positive selection, or simply the Site Frequency Spectrum (SFS).

The Site Frequency Spectrum is a histogram of i entries whose i^{th} entry represents the number of polymorphic sites at which the mutant allele is present in i copies within the sample [27]. There are two types of Site Frequency Spectrum, the unfolded SFS and the folded SFS. Let's say that we have a sample of n human chromosomal sequences from a population of N diploid individuals, with m segregating sites (positions in a sequence alignment that show differences (polymorphisms) between related genes). There are two different measures of site frequencies, ξ_i and n_i. Then, i is the number of polymorphic sites at frequency i/n in the sample or in other words, the number of segregating sites that the mutant base is present on i sequences in the sample and the other $n-i$ sequences have the ancestral base. In the case that the ancestral base is known, so it is possible to distinguish the ancestral from the mutant base, the unfolded site frequency ξ_i is estimated. Conversely, when the ancestral base is not known and it is impossible to distinguish it from the mutant base, the folded site frequency (n_i) is estimated. Therefore, n_i is the number of sites at which the less frequent base is present on i sequences out of n. The expected values of these quantities can reveal how population-level processes shape genetic variation [27].

Measure of variation using polymorphism data is very important. If a population behaves nicely, which means constant population size and no selection, then the observed SFS is connected to the population-scaled mutation rate θ (=4Nμ).Then E(ξ_i)= θ/i, for all i [19]. But, it is not the same in the case when the population evolves under positive selection. For example, for a panmictic population that evolves neutrally, it is predicted that the SFS is decreasing monotonically. That is, the frequency of i-derived variants is decreasing as i increases. In contrast, in the case where the population evolves under positive selection, the high-frequency derived alleles increase. This means that the scaled SFS under selective sweep differs from the scaled SFS under neutrality [1].

There are many different methods available to detect selective sweeps from genomic SNP data based on sweep patterns shown in the SFS. These tests help us to quantify the skew in the SFS of a sample of a population in relation to that expected under neutrality [19]. These tests are based on 4 different estimations of the scaled mutation rate $\theta = 4N\mu$, where N is the effective population size and μ the mutation rate per bp and per generation. The four θ estimators are presented below:

- Waterson's θ [28] that is based on the number of segregating sites

$$\theta_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}, \tag{1.1}$$

where n is the sample size

• The average number of pairwise differences between the n sequences of the sample is also an unbiased estimator of θ

$$\theta_\pi = \frac{\sum_{i<j} \pi(i,j)}{\binom{n}{2}} \tag{1.2}$$

• Fu and Li's statistic that is based on the number of singletons M [9],

$$q_M = \frac{n}{n-1} M \tag{1.3}$$

• Fay and Wu's θ_H which is based on the frequency distribution of derived alleles and by using data from an outgroup species it can distinguish the ancestral state of a polymorphism and then detects selective sweeps by an excess of high-frequency-derived polymorphisms [7].

$$\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} k_i i^2 \tag{1.4}$$

These estimators provide four measures of nucleotide diversity. Their combinations can be used to detect selection. The most famous statistic to detect recent selection is Tajima's D which compares the θ estimates from Tajima and Waterson. Specifically it is the difference of those estimates normalized by the variance of the difference.

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{Var(\theta_\pi - \theta_W)}} \tag{1.5}$$

The problem probably is to get a good idea about the variance of the difference of the estimators because you can find that Tajima's D is from a normal distribution but the mean is not zero and the variance is not one (basically, it is very difficult to construct a neutrality test based on each possible combination of θ). Looking at the distribution that results from this test, D depends on the value of θ so the distribution of D can be done by all θ values. Tajima proposed an approximate procedure to do this but you have to assume constant population size. The problem is that demography is not included in this test. As a result, positive Tajima's D values may mean either balancing selection or evidence for mildly bottlenecked demographic events, or population substructure. In all cases, the key-signature is the excess of intermediate-frequency allelic variants. Respectively, the negative values of Tajima's D show that there are new mutations that result in an excess of low-frequency variants compared to the neutral expectations. Thus, this is evidence for directional selection or population expansions or strong bottleneck events. Notably, in bottlenecks you can take both positive or

negative Tajima's D values depending on the time frame, the frequency of the haplotypes that survived after the bottleneck.

Other Tajima's-like-tests are Fu & Li (1993) ([9]), F^* and D^*, which are based on singletons,

$$F^* = \theta_T - \theta_M \tag{1.6}$$

$$D^* = \theta_W - \theta_M \tag{1.7}$$

and Fay and Wu's (2000) H [7],

$$H = \theta_\pi - \theta_H \tag{1.8}$$

which is influenced mostly by the high-frequency *derived* alleles.

In the last fifteen years, more advanced tests were developed to detect recent and strong positive selection. Probably, the most known test has been developed by Kim and Stephan [12] that uses the unfolded SFS . This test combines the expected reduction in variation with the expected skew in the SFS under the presence of recombination. The test, known as Kim and Stephan test, constructs a maximum likelihood approach to detect sweeps and give an estimation of the location of the site of the advantageous allele and the strength of selection, given the recombination rate.

Nielsen et al. [14] extended the Kim and Stephan test to detect deviations from neutrality due to SNP ascertainment bias and demographic history, under the assumption that a selective sweep has been completed.

In our study we are based on Nielsen et al. [14] test to detect selective sweeps, not using single sites but pairs of consecutive SNPs, by calculating the joint Site Frequency Spectrum. Usually the term 'joint SFS' or '2D SFS' is used to denote the SFS calculated simultaneously in a structured population and has been firstly mentioned by Wakeley and Hay [26] who implemented it among SNP sites between two different populations, to infer divergence times and gene flow. Here, however, the term 'joint SFS' or '2D SFS' means something different: the joint SFS between consecutive SNP sites along the genome. This was derived from the idea that under positive selection two consecutive nucleotide positions, displaying SNPs, may be linked due to the selective sweep effect and the reduced effect of recombination nearby the selected site. In this way, the loss of information between linked sites when calculating the SFS, to detect positive selection, can be compensated [3].

But, why do we use pairs of SNPs? Most theoretical studies of selective sweeps have focused on a model with one selected and one partially linked neutral locus [18]. But, maybe,

a selective sweep generates distinct patterns on multi-locus allele frequencies.So, we use a model of one selected and two consecutive neutral loci. In this way, we take advantage of both genomic signatures of positive selection, the increase in LD between neutral sites and the skew in the allele frequencies. Furthermore, using pairs is more accurate because selective sweeps do not only affect sequence diversity at a single neutral locus, but also the joint allele distribution of several partially linked neutral loci [17].

1.3 Difussion Theory

Here, we want to extend the site-frequency spectrum theory to consider the case where the sites are linked and they are not independent. We will use diffusion approximation to estimate the joint distribution of allele frequencies at linked sites.

The quantity we are interested in is the gene frequency or, differently, the proportion of a gene in the population that we study. In mathematical theory of population genetics, as it was founded by Fisher, Haldane and Wright, one of the problems is to investigate the change of gene frequencies when they are affected by mutation, natural selection, migration and random sampling of gametes (genetic drift). This is a stochastic process due to random sampling of gametes, which is a chance event. A mathematical approach that was developed to study similar problems, and it is proved that is the most powerful, is called Diffusion Theory. In this approach, the change in gene frequencies is a continuous stochastic process since the gene frequencies in large populations change almost continuously with time.

Generally, in such stochastic processes in population genetics, we have two major problems, (i) the gene frequency distribution at equilibrium and (ii) the probability of gene fixation. These problems are very important when we are trying to understand the genetic structure of the population of our interest. Gene frequency distributions have been studied extensively by Wright and have been extended by the work of Kimura [13] on the distribution of gene frequencies under irreversible mutation. This concept of gene frequency distribution is applicable not only to individual gene but even to nucleotide pairs. The second problem, the probability of fixation of a mutant gene in a population, is very important in evolutionary genetics studies and it has been studied firstly by Fisher and Haldane and extended by Kimura who used Kolmogorov backward equation to solve the problem. Motoo Kimura [13] was the first that gave a solution to the diffusion equation. Also, Robertson developed a theory in relation with this problem but for small population numbers. Using diffusion theory, we can study the fate of a gene in the population which can be either fixed or lost from the population

within a finite time interval. The process of change in gene frequencies can be treated using diffusion equations and particularly the *Kolmogorov backward equation*,

$$L = \frac{1}{2} \sum_{i,j=1}^{K} \alpha_{ij}(\chi) \frac{\partial^2}{\partial \chi_i \, \partial \chi_j} + \sum_{i=1}^{K} b_i(\chi) \frac{\partial}{\partial \chi_i}. \tag{1.9}$$

The Kolmogorov backward equation is one of the simplest and most useful equations in diffusion theory, where the $\alpha_{ij}(\chi)$ in the first term represents the drift and $b_i(\chi)$ in the second term represents the infinitesimal covariance matrix (general formula for multi-variate diffusion processes). Generally, this equation describes the change of allele frequencies due to drift (first term of the equation after the equality) and selection (the second term of the equation), but without mutation.

For example, let's say that we have an allele of interest that starts with a specific frequency and then we have different time intervals where the time is scaled by the effective population size, N. By using the diffusion method we can get an expectation for our allele frequency with genetic drift given a certain starting frequency and the sample size.

Furthermore, we can use the backward equation to get properties relevant to population genetics studies. For example, we can calculate the mean time of alleles starting at given frequencies until they become fixed in the population. Also, we can calculate the mean time of loss. Furthermore, in our case, we can also ask how long do we expect two neutral alleles to be segregating in the population of a given size given a certain starting frequency.

For example, when the allele frequency is $1/2N$ (i.e. a new mutation that occurs in the population) we can ask: 'What is the fate of this new mutation?', ' What is the probability that this new mutation will be fixed in the population?'. 'How does this probability change with the population size?'. And in the case that it gets lost 'what is the probability of loss?' A further question is, how long does it take for neutral mutation on average until this mutation been fixed in the population? For this last question, time is estimated to be $4N$ generations which means that it takes very long time for a neutral allele starting as a new variant to survive and become fixed in the population.

Chapter 2

Materials and Methods

2.1 The Model

Most theoretical studies of selective sweeps have focused on a model with one selected and one partially linked neutral locus[18]. However, genetic data are available for many partially linked loci. This raises the question that maybe, selective sweeps also generate distinct patterns on multi-locus allele frequencies [18]. Furthermore, as it is known, selective sweeps do not only affect sequence diversity at a single neutral locus, but also, the joint allele distribution of several partially linked neutral loci [17]. For these reasons, in our study, we consider a three locus model of one selected site and two partially linked neutral loci. For each locus, we assume that there are only two allelic types (i.e. a biallelic segregation model) and with simulations we show that using a multi-locus model results in high power in the detection of selective sweeps.

Assume a beneficial allele S which arises in a population of N haploid individuals at time $t = 0$, it has a selective advantage of s with respect to the wild type allele, and it increases in frequency until it fixes in the population. The effect of selection on removing bad alleles or with other words, the scaled selection coefficient is $\alpha = sN$. In our model, we assume that only when the beneficial allele fixes in the population, selection can be detected. So, we let T be the time of fixation . Furthermore, assume that reproduction in the population follows a Wright-Fisher model. As time is measured in units of N generations, the frequency path of the beneficial allele in the population can be described by the differential equation:

$$dX = \alpha X(1-X)coth(\alpha X)dt + (X(1-X))^2 dW \tag{2.1}$$

where $W =$ standard Brownian motion and $X_0 = 0$.

If we take genetic drift into account, the process stops when fixation time of the beneficial allele is about $2log(\alpha)/\alpha$, when the frequency of the beneficial allele is $X_T = 1 - \varepsilon$.

In our model two neutral variants are partially linked to the selected locus at $t = 0$ and increase their frequencies together with the beneficial allele [24]. We refer to the neutral locus as the left, L, and right, R, neutral locus. As a result we have two possible geometries for study, either (a) the neutral loci are on the same side of the selected site or (b) the selected locus is in the middle of both neutral loci. Furthermore, for the selected S-locus we have the wild type b allele and the beneficial B allele. For the other two neutral loci left and right (or L-loci and R-loci), we have the L, l and R, r alleles, respectively.

In our study, we chose to focus on the (a) geometry (Figure 2.1) as it was referred above and all of our results and the statistical method for sweep detection, that we suggest, are based on this model. During reproduction, recombination events might occur and can break up the association of these three loci. So, we have to consider the scaled recombination rates (scaled with the distance on the genome) which are different for the two geometries presented above. For the geometry of interest we denote the recombination rates between the selected and neutral loci by ρ_{SL}, and ρ_{LR}, respectively. Note that the recombination rate between the selected and the R neutral locus is $\rho_{SR} = \rho_{SL} + \rho_{LR}$. Generally, recombination rate is scaled to the distance along the genome.

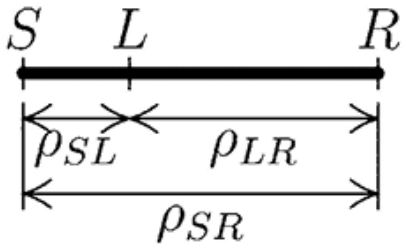

Fig. 2.1 The studied geometry where S is the selected site and L and R the two neutral, consecutive, polymorphic sites that are located on the same side of the selected site. Also, ρ_{SL}, ρ_{LR}, ρ_{SR} are the scaled recombination rates between the sites.

2.2 Computational and theoretical estimation of 2D-SFS

In the following section we describe how a 2D-SFS (or Joint Site Frequency Spectrum) can be estimated both computationally by simulations and theoretically by diffusion theory.

2.2.1 Computational estimation of a 2D-SFS

We performed two types of simulations to generate genealogies under neutrality and positive selection assuming that for each locus there are only two allelic types. For the neutral simulations of populations with constant size, we used the standard coalescent simulation program *ms* [21]. Two models reflecting the population's history were chosen, simulating genealogies with both constant size and experiencing a bottleneck phase. In order to specify that the demographic parameters change over time for the bottleneck effect, the -eN switch of *ms* was used. Particularly, we set that the population size shrunk 100 times at the time point $t_1 = 0.002$ and expanded to the original size at the time point $t_2 = 0.04$. The time when the demographic change occurs is measured from the present backwards in units of $4N_0$ generations.

Two sample sizes were set, with 20 and 500 individuals, respectively, and the sampling process was repeated for 1000 times. The θ value was set to 200 (where $\theta = 4N_0\mu$). For all possible combinations of the above parameters, we allowed for recombination ρ to occur at rates 1 or 10 (where $\rho = 4N_0r$). We also considered that the population was not completely isolated and allowed gene flow with rate $4N_0m = 0.2$, where m measures the fraction of each sub-population that is made up of new migrants in each generation. To simulate genealogies under positive selection and the trajectories for the selected positions we used the mssel algorithm created (and kindly provided) by R.R. Hudson. We simulated trajectories for 60 bottleneck scenarios, all for 10,000 generations, setting the following parameters: bottleneck severity, eliminating 40, 200, 400, 4000 and 8000 generations during the effect; bottleneck length, allowing for the effect to last for 80, 200 and 400 generations; bottleneck times, setting when the effect starts in units of $4N$ generations counting from the present backwards: 0.005, 0.01, 0.03, 0.05, 0.055, 0.1, 0.5. The position of the selective sweep was set to be in at position 0.5.

Based on our sets of simulations we took all pairs of SNPs (partially linked neutral loci) which appear consecutively in the alignment positions and we calculated the distance of each pair from the selective sweep position. We used the position of the SNP that was closer to the sweep position (by construction this is the location 0.5) as the value to calculate this distance. For n SNPs, there are $n(n-1)/2$ pairs of SNPs. We grouped some of these pairs in classes, since they were too many to handle. This classification is meaningful when trying to detect positive selection, since the middle class SNPs are rarely indicators of selective processes, whereas singletons or nearly fixed mutation SNPs are landmarks of positive selection. Distances of each SNP pair from the sweep position were plotted against the frequency of each SNP pair (Figure 3.1). We used linear interpolation in order to infer continuous values for the distances from the selective sweep position. Given these values,

we calculate the probabilities (likelihoods) to find a certain pair of SNPs in a given distance from the selective sweep.

We calculated the natural logarithm of these probabilities for all SNP-pairs and all positions in the alignment for both scenarios of neutrality and selection. Then we inferred the ratios of the log-likelihood of the selective sweep scenario over the neutrality scenario. Since we assume that the sites in the alignment evolve independently, we are able to add the log-likelihood ratios. We then used the Composite Likelihood Ratio Test (CLR, [12]) implemented in a PERL script to construct a Maximum Likelihood approach for testing selective sweep (null hypothesis) over neutrality (alternative hypothesis) in real data.

Since we used the conditional probability to observe a pair of SNPs in a certain distance from the selective sweep, we need to take into account the parameter α. This is $\alpha = (rlogN_e)/s$, where r is the recombination rate, N_e the effective population size and s the selection coefficient. The CLR test is a function of α so we need to estimate α and find its value that maximizes the CLR. For real data it is needed to calculate the p-value for the maximum CLR value by simulating data under the alternative hypothesis (neutrality scenario).

2.2.2 Theoretical estimation of the Joint Site Frequency Spectrum of two polymorphic sites

In order to study theoretically, the Joint Site Frequency Spectrum (2D-SFS), we need to derive the mathematical formula that describes the probability of a pair of consecutive SNPs to be observed at a certain distance from a positively selected site. This probability encloses two terms, the first term describes the probability of the pair given that there is recombination between the polymorphic sites and the second term is the probability of the pair without recombination between them. For simplicity we will not consider the case of recombination in this study. Before we start working with our 3 loci model, we will define Green's function that we will implement in the next section.

2.2.3 Green's Function

The Wright-Fisher model describes the process of genetic drift as random sampling with replacement. Another way, more efficient to mathematical analysis, to model the process of genetic drift is the diffusion approximation [6, 8, 13, 22]. When the population size is larege

then the Wright-Fisher model is approximated by multidimensional diffusion process with infinitesimal generator [5, 6].

$$L = \frac{1}{2} \sum_{i,j=1}^{K} \alpha_{ij}(\chi) \frac{\partial^2}{\partial \chi_i \, \partial \chi_j} + \sum_{i=1}^{K} b_i(\chi) \frac{\partial}{\partial \chi_i} \tag{2.2}$$

where $a_{ij}(\chi) = \chi(\delta_{ij} - \chi_j)$ is the infinitesimal covariance matrix and the $b(\chi)$ is the infinitesimal drift vector. Diffusion model describes the evolution of allele frequencies under random drift (second term of the above equation), selection (first term of the above equation) and no mutations.

Here, we need the process of the K allele diffusion (as we have a studied geometry of 3 sites, one selected and 2 neutral loci) which can be described by Green's function $G(\chi, \chi')$ [10, 20] which (in continuous space) is the solution of the differential equation:

$$LG(\chi, \chi') = -\delta(\chi - \chi') \tag{2.3}$$

with boundary condition $G(\chi, \chi') = 0$ where $\Delta := [\chi = (\chi_1, \ldots, \chi_K) : \exists\, k$ such that $\chi_k = 0]$ is the boundary condition, χ' is an interior point of Δ and $\delta(\chi - \chi')$ is the delta function of Dirac. Generally, Green's function $G(\chi, \chi')$ is a Markov chain process and shows the expected number of visits to frequency χ' starting from a frequency χ.

Suppose,now, that $G(u, v)$ is the solution of the equation:

$$LG(u; v) = -\delta(u - v) \tag{2.4}$$

with boundary condition to be G(z;v) = 0 for any point z at the boundary and v an interior point of Δ. Then

$$G(u; v) = \int_0^\infty p(X(t) = v | X(0) = u) \partial t, \tag{2.5}$$

where $p(X(t) = v|X(0) = u)$ is the transitional probability density or simpler the probability for an allele to reach a frequency v at time t when the starting frequency at time 0 is u.

Let's define the Green's function $G(\chi, \chi')$ for the interval $[u, v]$ by the property that:

$$g(\chi) = \int G(\chi, \chi') f(\chi'), \mathrm{d}\chi' \tag{2.6}$$

satisfies $Lg = -f$ for $u < \chi < v$ with $g(u) = g(v) = 0$

Then, if $\tau_{u,v} = T_u \wedge T_v$ is the exit time from (u,v) then:

$$E_\chi(\tau_{u,v}) = \int_u^v G(\chi,\chi')\frac{\partial}{\partial\chi'} \tag{2.7}$$

Green's function is used when we want to know where the process spends its time before reaching a boundary point at a time τ and then we can also estimate E_τ.

2.2.4 Joint Site Frequency Spectrum of two polymorphic sites

Let's consider a sample of N haploid individuals and assume an infinite-sites-model without recombination. The infinite-sites-model assumes that two mutations can not occur at the same site of the chromosome both. Let's also consider two mutations in two different polymorphic sites that occurred in the past, the first one a at time t_1 and the second b at time t_2 (where $t_1 < t_2$) and in the present population they have frequency $p+1$ and p_2, respectively (where $0 < p_1, p_2 < 1$). We have four different genotypes A_1 which carries only the mutation a, A_2 which carries only the mutation b, A_3 which carries both a and b and the wild type A_4 (no mutations) with frequencies X_1, X_2, X_3, X_4, respectively.

Our goal is to calculate the joint distribution of p_1 and p_2. We split this in two cases depending the allele type where the second mutation occurred. At that time (when the second mutation occurred) in the population there are the genotypes A_1 or A_4 which means that after the second mutation we take the genotypes A_2 or A_3, respectively (Figure 2.2).

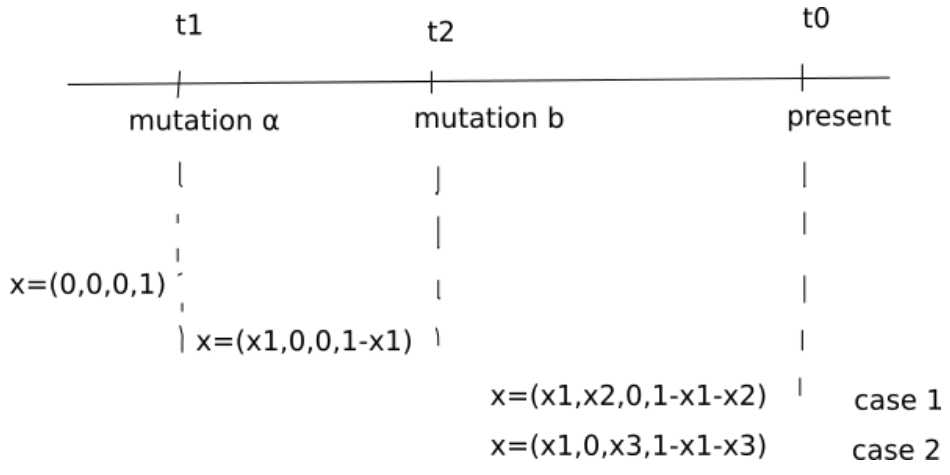

Fig. 2.2 This is the diagram of the 2 mutation events and the allelic types that they create. Mutation α occurs at time t1 and mutation b occurs at time t2. If the mutation b occurred within the wild type allele then it creates the allele type A2 whereas if it occurs within the A1 allele then creates the allele A3. The two cases corresponds to two different boundaries that are described in the text.

Firstly, let's take the case when the second mutation occurred within the wild type allele. If we assume that both sites are polymorphic then the time point t_0 (the time that we take the sample) in the population there are both A_1 and A_2 genotypes but there is not restriction for A_4 as it can be either present or extinct. This means that we need to consider for the case where the A_4 is extinct in the present population and so $X_1 + X_2 = 1$ (the frequencies of the other alleles are complementary) and the case where the A_4 is present and then $X_1 + X_2 < 1$.

We have 3 time points, t_2 where the second mutation occurred, t_1 where the first mutation occurred and t_0, the time that we take the sample (present). Our goal, here, is to calculate the probability to have A_1 allele in frequency X_1 and A_2 allele in frequency X_2 in the population at the time that we take the sample (t_0). For this, we need two transition probabilities. Specifically, when the first mutation arises in the population at time point t_1 it has frequency $\delta = 1/N$ and so, the transition probability where the frequency X_1 of A_1 allele changes from δ at t_1 to u at t_2 is denoted as $P(X_1(t_2) = u|X_1(t_1) = \delta)$. Also, the transition probability of X_1 and X_2 from u and δ at t_2 to x_1 and x_2 at t_0 is $P(X_1(t_0) = x_1 X_2(t_0) = x_2|X_1(t_2) = u, X_2(t_2) = \delta)$. Then the probability that we want to derive is

$$f1(x1,x2) = \int_{-\infty}^{0} dt_2 \int_{-\infty}^{t_2} dt_1 \int_{0}^{1} P(X_1(t_2) = u|X_1(t_1) = \delta) \tag{2.8}$$

$$P(X_1(t0) = x_1 X_2(t_0) = x_2|X_1(t_2) = u, X_2(t_2) = \delta)u(1-u)du \tag{2.9}$$

Using the Green's function for the above equation, we can re-write it as follows:

$$f_1(x_1,x_2) = \int_{0}^{1} (1-u)G(\delta;u)T_1(u,\delta;x1,x2)du \tag{2.10}$$

where G(δ;u)=2δ/u and $T_1(u,\delta$; x1,x2) is the time spent in X1=x1 and X2=x2 when the starting frequencies are u and δ respectively, and it contains the case of x1+x2<1 inside the boundary and the case of x1+x2=1 at the boundary. Specifically, in our case that both mutant alleles are neutral, the $T_1 functionisexpressedas$:

$$T_1(u,\delta;x1,x2) = T_a(\delta,u;x2,x1) + T_b(x_2;1-u-\delta,\delta)\delta(x1+x2-1) \tag{2.11}$$

where the first term after the equality is the time spent inside the boundary and the second term is the time spent at the boundary (δ() function is used to constrain x_1 and x_2 to be sum 1 along the boundary).

Let's consider, now, the second case, when the second mutation occurred within the A_1 allele.In this case,assuming again that both sites are polymorphic, the alleles that we

can see in the population after the second mutation are A_1, A_3 or A_4. At the time point t_0 the frequency of the first mutation in the population is $X_1(t_0)+X_3(t_0)$ and the frequency of the second mutation is $X_2(t_0)$.In this case, A_3 and A_4 must be present in the population in order to be fulfilled the assumption of polymorphic sites which means that $X_3(t_0) > 0$ and $X_4(t_0) > 0$. Also, A_4 must be present as in different case the the first mutation site would be not polymorphic. But as the first mutation site is polymorphic due to the presence of A_3 and A_4, there is not restriction for A_1 as it can be extinct ($A_1 = 0$).

Our goal here is to calculate the probability of $X_3(t_0) = x_3$ and $X_4(t_0) = x_4$ which similar to the first case is proportional to:

$$f_2(x_1,x_3) = \int_0^1 uG(\delta;u)T_2(u,\delta;x_1,x_3)du \tag{2.12}$$

where $G(\delta;u) = 2\delta/u$ and $T_2(u,\delta;x_1,x3)$ is the time spent in $X_1 = x_1$ and $X_3 = x_3$ when the starting frequencies are u and δ respectively, and it contains the case of when $x_1 + x_3 < 1$ inside the boundary and the case of $x_1 = 0$ when the A_1 is extinct. Specifically, in our case that both mutant alleles are neutral, the T_2 function is expressed similar to the case 1, as:

$$T_2(u,\delta;x1,x2) = T_c(\delta,u-\delta;x_3,x_1) + T_d(x_3;u-\delta,\delta)\delta(x1) \tag{2.13}$$

where the first term after the equality is the time spent inside the boundary and the second term is the time spent at the boundary $x_1 = 0$.

At the end, combining the two previous cases, the probability of the frequencies of the two polymorphic sites being p1 and p2 equals

$$g(p1,p2) = f_1(p_1,p_2) + f_2(p_1-p_2,p_2)I(p_1 > p_2) \tag{2.14}$$

up to a difference in normalization constant. Analytical formulas $T(y;x_1,x_2)$ (mean time spent at the boundary) were derived by [29] using *Gauss hypergeometric function* and *Jacobi polynomials*.

Chapter 3

Results

3.1 Simulations Results

As it was described in Chapter 2, based on our sets of simulations we took all pairs of consecutive SNPs that appear in the alignment and we calculated the distance of each pair from the selective sweep site. We used the position of the SNP that was closer to 0.5 as the value to calculate this distance. For n sequences the SNP pairs that exist are $(n-1)^2$, since we assumed that all sites display a SNP of some class.After, we grouped some of these pairs in classes, since they were too many to handle. This classification makes more sense when trying to detect positive selection, since the middle class SNPs are rarely indicators of selective processes, whereas singletons or nearly fixed mutation SNPs are landmarks of positive selection. Distances of each SNP pair from the sweep position were plotted against the frequency of each SNP pair. Then, we plotted the conditional probability of all the pairs against their position on the alignment (scaled position). The formula that we used for the calculation of the probability was derived by Nielsen et.al 2005 (formula (6)) [14] that describes the probability of a single SNP to be observed at a certain distance from a selective sweep position.

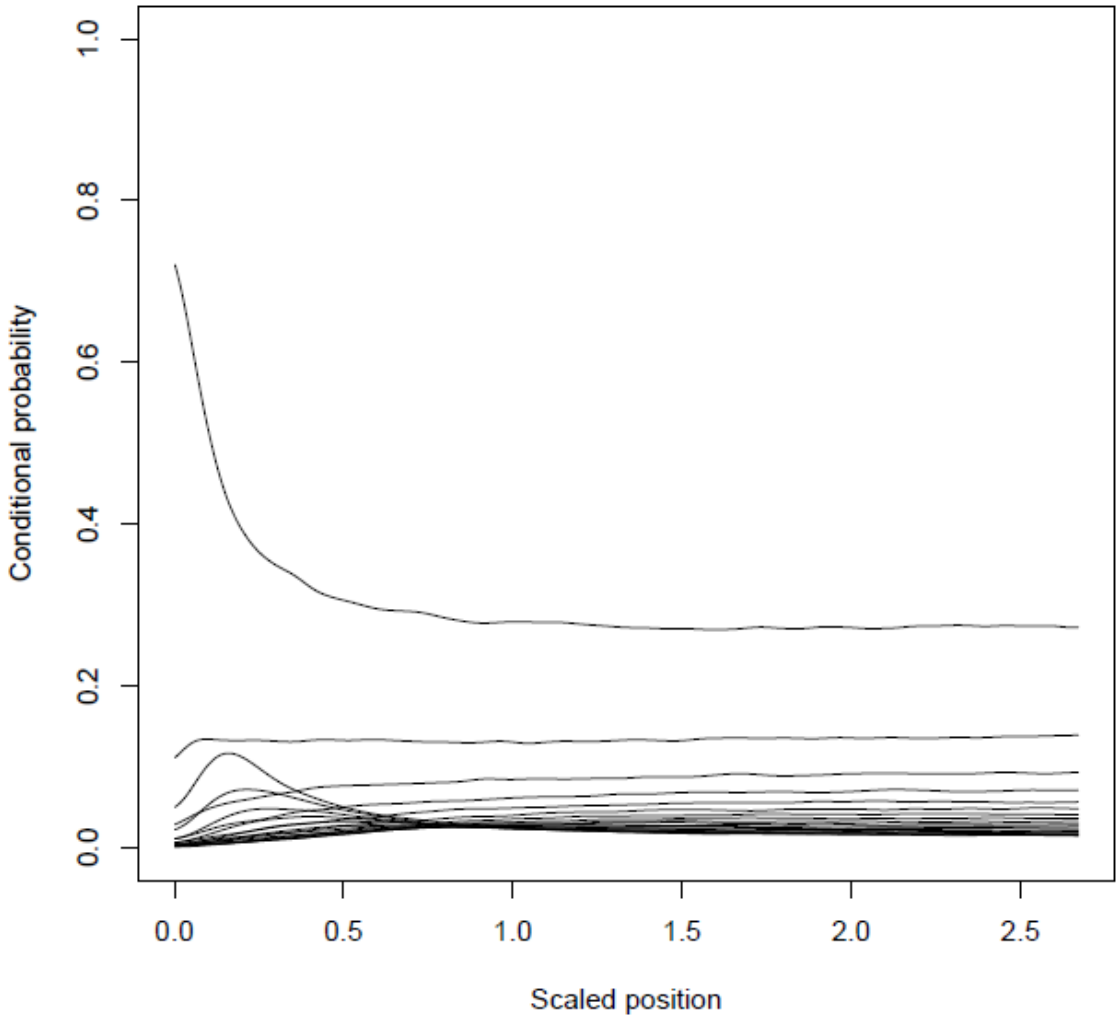

Fig. 3.1 Computational estimation of the distribution of a 2D-SFS

3.2 Distribution of 2D-SFS with Diffusion Theory

We used *Mathematica* 8.0.0 software to derive the distribution of the formula (2.14) that was presented above and describes the probability of the frequencies of two segregating sites to be p_1, p_2 respectively. Unfortunately, because of the complexity of the mathematics, it takes so long for it to run and 4 months now, has not yet, given a result.

Chapter 4

Discussion

In our study we are trying to extend the work of Nielsen (2005) [14] and construct a test to detect positive selection based on the distribution of the Joint Site Frequency Spectrum.

For this purpose, we considered a three-loci model of one selected and two partially linked neutral loci. For each locus we assumed that there are only two allelic types. Calculating the joint SFS proved to be statistically more informative to the Site Frequency Spectrum, since the detection of a selective sweep is usually obscured by the demographic effect on the SFS and the existing methods do not incorporate demographic models [23].

So, one of the advantages of this study is that, unlike other methods for detection of selection in the past, we consider for demography. This is very important as demographic scenarios affect the SFS in a similar pattern like positive selection and in most cases it is very hard to distinguish between the two signals. In both cases the variation is decreased, but for different reasons. In the case of positive selection, this is due to hitchhiking effect whereas in the case of bottleneck the reduction in the diversity is because of the reduction of the effective population size. For this reason, We inferred simulated models for demography, particularly for 60 cases of bottleneck effects, to understand where the signal of the SFS is affected by demography and where it is affected by selection.

Previous statistical tests like Tajima's D ignore the demographic history of a population and as a result both positive and negative Tajima's D values do not indicate selection necessarily. This is because demography produces the same skew in the distribution of allele frequencies as positive selection does. And as a result taking positive or negative values of D, it means nothing. You have to combine this estimator with coalescent simulations that consider the demographic history of the studied population and then you are able to reject neutrality hypothesis.

Furthermore, we are trying to detect selective sweeps, not using single sites but pairs of consecutive SNPs, which means that we are studying the joint Site Frequency Spectrum,

a fact that provides many advantages. This was derived from the idea that under positive selection two consecutive polymorphic sites, may be linked due to the selective sweep effect. In this way, the loss of information between linked sites when calculating the Site Frequency Spectrum, to detect positive selection, can be compensated.This study brings together the two aforementioned issues in an attempt to improve the power of detecting selective sweeps along the genome.

The major advantage of this method is that it combines both genomic signatures of positive selection, the increase in Linkage Disequilibrium between neutral sites and the skew in the site frequency spectrum . Also, as it has been shown by previous studies, selective sweeps affect not only single sites but they generate distinct patterns on multi-locus allele frequencies. Furthermore, the simulations have shown that using pair of SNPs result in high power in sweep detection.

References

[1] Andreas Wollstein, W. S. (2015). Inferring positive selection in humans from genomic data. *Investigative Genetics*.

[2] BARTON, N. H. (1998). The effect of hitch-hiking on neutral genealogies. *Genetical Research*, 72:123–133.

[3] Boitard S, Schlötterer C, F. A. (2009). Detecting selective sweeps: A new approach based on hidden markov models. *Genetics*, 181(4):1567–78–494.

[4] Cutter, Asher D., P. B. A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet*, 14.

[5] Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer, Berlin.

[6] Ewens, W. (1979). *Mathematical Population Genetics*. Springer, New York.

[7] Fay, JC.; Wu, C. (2000). Hitchhiking under positive darwinian selection. *Genetics*.

[8] Fisher, R. A. (1931). Xvii.—the distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh*, 50:204–219.

[9] Fu, YX.; Li, W. (1993). Statistical tests of neutrality of mutations. *Genetics*.

[10] Karlin, S., T. H. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

[11] Kelley JL, C. H. (2008). Positive selection in the human genome: From genome scans to biological significance. *Annual review of genomics and human genetics*.

[12] Kim Y, S. W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*.

[13] Kimura (1964). *Diffusion models in population genetics*. J.Appl. Probab.

[14] 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):1566–75.

[15] Nielsen R, Hellmann I, H. M. B. C. C. A. (2007). Recent and ongoing selection in the human genome. *Nature reviews Genetics*, 8(4):857–868.

[16] Nurminsky, D. (2005). *Selective Sweep*. Springer, New York.

[17] Pfaffelhuber, P., Lehnert, a., and Stephan, W. (2008). Linkage disequilibrium under genetic hitchhiking in finite populations. *Genetics*, 179(1):527–37.

[18] Pfaffelhuber, P. and Studeny, A. (2007). Approximating genealogies for partially linked neutral loci under a selective sweep. *Journal of Mathematical Biology*, 55(3):299–330.

[19] R, R. (2013). Learning natural selection from the site frequency spectrum. *Genetics*.

[20] Roach, G. (1982). *Green's Functions*. Cambridge University Press, Cambridge.

[21] RR, H. (2002). Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18:337–8.

[22] S., W. (1942). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 24:253–259.

[23] Simonsen KL, Churchill GA, A. C. (1995). Properties of statistical tests of neutrality for dna polymorphism data. *Genetics*, 141:413–429.

[24] Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23:23–35.

[25] Stephan, W., Song, Y. S., and Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*.

[26] Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res.*

[27] Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts Company Publishers, Greenwood Village, Colorado.

[28] Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*.

[29] Xie, X. (2011). The site-frequency spectrum of linked sites. *Bulletin of Mathematical Biology*, 73(3):459–494.

Appendix A

Brief definitions

Selective sweep

Selective sweep is spread of an advantageous allele throughout the population by strong positive natural selection and associated loss of variation near it (by hitchhiking)

Hitchhiking

Hitchhiking is spread of other nearby **neutral** alleles along with the advantageous one because of linkage (lack of recombination)

Recombination

Recombination the force that breaks down allelic combinations.

Neutral variation

Neutral variation does not affect the fitness of the organisms. It is only affected by random drift.

Genetic drift

Genetic drift is the change in allele frequencies in a population due to random sampling of gametes.

Positive selection

Positive selection is selection acting upon new advantageous mutations.

Site Frequency Spectrum

Site Frequency spectrum is a histogram whose i^{th} entry is the number of polymorphic sites at which the mutant allele is present in i copies within the sample.

Genetic linkage

Genetic linkage is the tendency of alleles that are located close together on a chromosome to be inherited together during meiosis.

Linkage disequilibrium

Linkage disequilibrium is the non-random association of alleles at different loci.