Inference from Two Non-Equilibrium Models in Population Genetics

by

Benjamin Marco Peter

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rasmus Nielsen, Chair
Professor Montgomery Slatkin
Professor Yun S. Song

Fall 2014

Abstract

Inference from Two Non-Equilibrium Models in Population Genetics

by

Benjamin Marco Peter

Doctor of Philosophy in Integrative Biology

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Rasmus Nielsen, Chair

Improvements in sequencing technologies and the resulting increased availability of genetic data call for new and more sophisticated analysis methods. Particularly in ecological genetics and evolutionary biology, questions that can be addressed are often limited by the availability of analysis tools and statistical inference procedures. Non-equilibrium models in particular have been relatively poorly studied, mainly because analytical approaches are challenging and many useful and well-known results make equilibrium assumptions. However, using heuristic methods and strongly simplified models, we can make progress and arrive at procedures that help us gaining new insights from population genetic data. After an introductionary first chapter, in the second chapter, I develop an Approximate Bayesian Computation procedure to distinguish selection from standing variation from selection on a *de novo* mutation. This method is applied to human genetic data where we identify two genes, ASPM and PSCA, that are most likely affected by selection on standing variation. In the third chapter, I develop an inference procedure to infer the origin of a range expansion, introducing the directionality index statistic ψ. Applying this method to human data, we find a most likely origin of humanity in southern Africa, and evidence of the main expansion routes into Asia, finding evidence for a Southern route. In the fourth chapter, I extend the work on range expansions by developing an analytical model based on branching processes, which gives a biological interpretation to ψ, and allows us to measure the decay of genetic diversity with distance. An application to *Arabidopsis thaliana* reveals that we are able to infer both recent expansions in the Americas, as well as expansions from the last glacial maximum in Europe. Between these three chapters, I use different approximation procedures to introduce inference procedures for models where direct likelihood calculations are not available.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Inference methods have become one of the main focus of population genetic theory in the last few decades, ever since genetic data became widely available (Nielsen et al., 2005). As sequencing technologies further improve, become cheaper and ubiquitously available, this trend is likely to further increase. In evolutionary biology, many old questions can be addressed in light of this new data. In this thesis, I present inference methods addressing two major problems in evolutionary biology: Distinguishing between different modes of positive selection, and inference under a range expansion model.

In the next chapter, I will present a method to distinguish two of the major types of positive selection: selection on a *de novo* mutation and selection on standing variation. The two scenarios differ in the order in which mutation and the appearence of new selective pressures occur: Under the selection on standing variation model, we assume that a potentially beneficial mutation may arise and segregate in the population before the population ever experienced the conditions under which these mutations are beneficial; the population may pre-adapt to novel environments. On the other hand, selection on a *de novo* mutation is the scenario where new mutations evolve only after an environment has changed – in that regime, the population is necessarily slower to adapt, as it has to wait for beneficial mutations to occur after the environment changes. Both modes of adaptations are likely to be important under different conditions and in Chapter 2 I will present one of the first empirical tests to distinguish these two modes of selection using genetic data. I approach this by combining several common statistics used to detect selection in an Approximate Bayesian Computation (ABC) framework, which allows me not only to discriminate between the two models of selection, but it also provides estimates of the age of the selected allele and the selection coefficients acting on them. I use simulations to assess the power and accuracy of the method, and apply it to seven of the strongest sweeps currently known in humans. I identify two genes, ASPM and PSCA, that are most likely affected by selection on standing variation and I find three genes, ADH1B, LCT and EDAR in which the adaptive alleles seem to have swept from a new mutation. I also confirm evidence of selection for one further gene, TRPV6. In one gene, G6PD, neither neutral models, nor models of selective sweeps, fit the data, presumably because this locus has been subject to balancing selection.

The inference methods for range expansion models are presented in Chapters 3 and 4. Range expansions are thought to be very common in nature, as one of the main processes through which species change their habitat boundary, and are thought to occur in all taxa from bacteria through plants and animals, on scales from petri dishes to the entire globe, and on time scales from weeks to millennia (Hallatschek et al., 2007; Hewitt, 1999; Taberlet et al., 1998). Nevertheless, inference under explicit range expansion models has been a little studied topic so far, primarily because until this decade it was prohibitively expensive to generate relevant data for almost all species (but see e.g. François et al., 2008; Ramachandran et al., 2005). However, as we are entering the era where assessing genome-wide genetic data for many individuals becomes economically feasible, we can address range expansions in light of this new data. In Chapter 3, I first develop a statistical test to reject equilibrium isolation-by-distance. Equilibrium models have been commonly assumed for inference procedures (Excoffier, 2004; Hey and Nielsen, 2004; Nielsen and Wakeley, 2001), however, it is not clear how often the equilibrium assumption actually holds. We develop a test to reject equilibrium migration models, while being largely agnostic about the underlying population structure. While the range expansion model is only one of many possible alternative models, in some cases historical or geological records may suggest it to be appropriate. For these cases, I propose a method to infer the location of the origin of the expansion.

Under the range expansion model, inference is based on the novel statistic ψ (the directionality index) that detects asymmetries in the two-dimensional allele frequency spectrum of pairs of population. These asymmetries are caused by the series of founder events that happen during an expansion and they arise because low frequency alleles tend to be lost during founder events, thus creating clines in the frequencies of surviving low-frequency alleles. Using simulations, I show that ψ is more powerful for detecting range expansions than both F_{ST} and clines in heterozygosity. I also show how the approach can be adapted to more complicated scenarios such as expansions with multiple origins or to infer barriers to migration, and I illustrate the utility of ψ by applying it to a data set from modern humans.

A further parameter of the range expansion model that is of great interest is the strength of the founder effect. In contrast to Chapter 3, where I largely theory and argue empirically, in Chapter 4 I develop a population genetic model of a range expansion. Through a series of approximations, I arrive at a branching process interpretation of a range expansion, in which founder effects can be interpreted as an increase in genetic drift caused by an increased offspring variance at the wave front. This allows us to measure the strength of the founder effect, dependant on an effective founder size. I demonstrate that the predictions from the branching process model fit very well with Wright-Fisher forward simulations and backwards simulations under a modified Kingman coalescent, and further show that estimates of the effective founder size are robust to possibly confounding factors such as migration between subpopulations. I apply my method to a data set of *Arabidopsis thaliana*, where we find that the founder effect is about three times stronger in the Americas than in Europe, which may be attributed to the more recent, faster expansion.

The common element of these three chapters are their focus on inference from models where calculations of full likelihoods is not feasible. Through approximations, we arrive at methods whose properties can be assessed using simulations, and, subsequently, can be applied to explore the evolutionary history of humans and *Arabidopsis thaliana*.

Chapter 2

Distinguishing between Selection from Standing Variation and Selection on a *De Novo* Mutation

2.1 Introduction

Most organisms harbor large amounts of, mostly neutral or nearly neutral, standing genetic variation (Hernandez et al., 2011; Hurst, 2009; Kimura, 1985; Ohta, 1992) As environments change, alleles that previously segregated neutrally, or were only weakly affected by selection, may become targets of strong selection. Examples of a change in environment that could induce such a change include invasion of a new habitat or niche through dispersal, climate changes, and introduction of novel disease agents. This type of selection, in which selection acts on already segregating alleles, is called selection from standing variation (SSV). We contrast this model with the more commonly assumed model of selection on a *de novo* mutation (SDN). In the SDN model the selection pressure already exists when a new mutation is introduced into the population. In addition, there are several more complicated scenarios of selection. The case where an allele under selection has multiple independent origins has received particular attention (Hermisson and Pennings, 2005; Pennings and Hermisson, 2006a,b), and is often also referred to as selection from standing variation. In this paper, we focus on the case where all copies of an allele are identical by descent, and do not consider multi-origin alleles. Of great interest is the question of which mode of selection has been more frequent in the evolution of a species (Hermisson and Pennings, 2005; Innan and Kim, 2004). In particular, if we observe a selected variant, which mode of selection is more likely to have occurred? Theoretical results by Hermisson and Pennings (2005) find that SDN should be common if selection is strong and mutation rates are low, in all other cases we expect SSV to be more prevalent.

2.1.1 Statistics Affected by Selection

Detection of selected regions has been a major goal in population genetics in recent years (Akey et al., 2004; Sabeti et al., 2002, 2006; Williamson et al., 2005). Rather than working with the full data, all of these studies simplified their data by using various statistics designed to detect the signal of selection (see e.g. Nielsen et al., 2005; Sabeti et al., 2006). These statistics may be classified in different categories, based on the information they exploit. First, functional differences between different codon positions, and the substitution rates of synonymous and non-synonymous sites were used by Hudson et al. (1987) and McDonald and Kreitman (1991). Another approach relies on finding related populations, where selection acts on only one of them. This leads to locus-specific high population differentiation, which may be detected by statistics such as F_{ST} (Lewontin and Krakauer, 1973) or XP-EHH (Sabeti et al., 2007). A third category of statistics is based on the length of haplotypes associated with a given allele. Haplotypes associated with the selected allele will on average be younger than haplotypes carrying the derived allele, and there will therefore be fewer recombination events that break up the haplotypes. Statistics such as EHH (Sabeti et al., 2002) and iHS (Voight et al., 2006) were developed to detect this pattern. Finally, the site frequency spectrum (SFS) can also be used to detect departures from neutrality and hence selection. SFS based statistics usually compare various estimators of the population mutation rate θ. The first and perhaps most well-known statistic in this category is Tajima's D (Tajima, 1989), but the statistic can be generalized (Achaz, 2009; Fu, 1997), and other statistics such as Fay and Wu's H (Fay and Wu, 2000) belong to the same family.

2.1.2 Distinguishing SSV and SDN

In this study, we are interested in distinguishing the SDN and SSV models of evolution for a single putatively adaptive mutation. Barrett and Schluter (2008) identify three possible ways of identifying SSV: i) the selected allele may occur in an ancestral population, ii) an allele is shown to be older than the environment it is adaptive in and iii) the signature of selection at linked loci, the selected sweep, is different between SSV and SDN. Our approach is based on differences in the genetic signature of selection, but when possible, we will compare to inferences based on i) and ii). To understand the difference between the SSV model and the SDN model, it is important to realize that all the information regarding selection, and mode of selection, is captured by the allele frequency trajectory through time. In other words, the full allele frequency path through time would be a sufficient statistic for the selection coefficient, if it was known. As selection acts only to change the allele frequency in the selected site, and does not act directly on adjacent sites, the effects of the selection on linkage disequilibrium, haplotype patterns, allele frequencies in linked sites, etc., are only through the effects caused by the change in allele frequency of the selected allele (hitch-hiking effects). This observation is the foundation for standard population genetic theory on selective sweeps (e.g. Kaplan et al., 1989; Spencer and Coop,

2004) and forms the basis for several simulation methods, in which the path of the selected mutation is first simulated and then neutral simulations are performed conditional on the allele frequency path (Spencer and Coop, 2004). Such simulation methods would be invalid if the allele frequency path did not contain all information regarding the selection coefficient acting on the selected mutation. Similarly, if the path of an allele is the same under the SSV and the SDN model, no additional genomic data could help us distinguish between the two models. Armed with this insight, we can further explore the differences between the two models. Figures 2.1a,b depict the trajectory, the number of copies of the selected allele through time for an SSV and SDN model. Looking backward in time, the adaptive alleles are selected at first in both models, and during this stage the two models do not differ at all. In the SSV model, however, the mutation stops being advantageous at some point in the past. Backwards from this time point, the mutation in the SSV model acts as a neutral allele, whereas the mutation in the SDN model is under selection. As selection is the same in the phase when both alleles are selected, the difference between the models is during the phase in which selection is acting on the mutation in the SDN model but not in the SSV model. How big is this difference? It depends on two parameters: the selective advantage of the mutation under the SDN model, and the frequency of the mutation at the time when selection first start acting in the SSV model. A good measure of the difference might be the allele age distribution at this point, which is plotted in Figure 2.1c and Figure 2.1d for a mutation at a frequency of 1% and 5%, respectively. Unfortunately, it turns out that the difference it is rather small: While the allele age of a mutation at a low frequency does depend on the selection coefficient, the difference is very small if selection is weak. Clearly, it will be much easier to distinguish between the two models if selection is strong and if the frequency of the mutation is initially high in the SSV model. However, we cannot observe the trajectory directly, but only the diversity at linked site. It has been shown that the genetic signature of sweeps from standing variation differs in three important aspects from the signature of sweeps from new mutations (Przeworski et al., 2005): at the same selection coefficient, the signal of selection from standing variation is 1) weaker and 2) affecting a narrower region. As a third difference, we expect an increased variance in both allele age and trajectory. Under the SSV model, the selected allele may be present on several haplotypes when selection starts, and these haplotypes will be affected equally strongly by selection. Thus, there will be more variation compared to SDN, and the, loss-of-diversity signal of selection will be weaker. The fact that the signal of selection affects a narrower region is due to the fact that the selected allele is older in the SSV model, and hence recombination had more time to break it up (Figure 2.1a, 2.1b). The increase in variance is evident from the large variance in the neutral phase of the allele trajectory in Figure 2.1b, and the wider distribution of the allele age of neutral alleles in Figure 2.1c and 2.1d. In Figures 2.1e and 2.1f we give the expected distribution of Fay and Wu's H (Fay and Wu, 2000) and EHH (Sabeti et al., 2002), two statistics used to detect selection, and where we show that the signal is indeed expected to be weaker and affecting a narrower region under the SSV model. The objective of this paper is to develop and explore a statistical method for distinguishing between SSV and SDN models, and for providing associated estimates of relevant parameters. However, the method we develop is not intended as a new method for performing scans for selection in genome-wide data or for quantification of genome-wide levels of selection. For computational reasons, other methods might be more suitable for such genome-wide analyses. We focus on illustrating the method on a few loci previously hypothesized to be under selection in humans, but the method could as well be applied to other human loci or data from other species.

2.1.3 Approximate Bayesian Computation

To exploit the characteristics of selective sweeps discussed in the previous section, we combine different statistics and calculate them for different genomic regions. Using combinations of statistics to improve inference is not a new concept, and has been applied previously (e.g. Grossman et al., 2010). Here, we choose an Approximate Bayesian Computation (ABC) framework for combining statistics (Beaumont et al., 2002; Tavaré et al., 1997). ABC has the advantage that it extends naturally to allow both model choice and parameter estimates under a given model. ABC was developed to estimate parameters of complex models in manageable computer time, and has been widely used in population genetics, most frequently to infer parameters for complex models of demographic history (Beaumont et al., 2002; Csilléry et al., 2010; Fagundes et al., 2007; Peter et al., 2010; Wegmann et al., 2010). Several implementations of the ABC algorithm have recently been published (Cornuet et al., 2008; Jobin and Mountain, 2008; Wegmann et al., 2010), and in the past few years, various variations of the algorithm have been developed (Blum and François, 2009; Marjoram et al., 2003; Sisson et al., 2007). ABC is a rejection sampling algorithm used to calculate the posterior distribution of a parameter under a given model, used frequently when the likelihood cannot be calculated analytically. In ABC inference, a large number of data sets are simulated using parameters randomly drawn from a prior distribution. If a simulation does not match the observed data, it is rejected, otherwise it is retained. However, if the data is complex, the probability of a match is prohibitively low, and two important approximation steps are used: First, the data is transformed into a set of summary statistics. If these statistics are sufficient (i.e. retain all the information present in the data), this step is exact. However, in many cases, including this study, no sufficient statistics are known, and this step results in a first approximation step. In many cases, however, this transformation will still result in very low acceptance probabilities. Therefore, the condition of an exact match is relaxed. Specifically, the summary statistics based on the simulations (S) are compared to the summary statistics of observed data (S^*). Using some distance measure δ, simulations are retained if $|\delta(S, S^*)| < \epsilon$ for an arbitrarily small distance ϵ. Frequently, some post-sampling adjustment is used in an attempt to correct for the error introduced in the second approximation step, and posterior distributions are estimated from the parameters of the retained simulations. In this study, we propose to use ABC to distinguish between a selective sweep from a new mutation and a selective sweep from standing variation. We use simulations to determine which parts of the parameter space the method has power to make this distinction, and aim to estimate parameters under both models. We then apply our method to seven genes that were previously reported to be under selection.

2.2 Results

2.2.1 Accuracy of Parameter Estimates

We first wanted to assess how accurately we can estimate the selection coefficient and the age of the selected mutation from the SSV and SDN models. For this purpose, we performed ABC inference on simulated data sets with known parameter values. Results for a case of moderately strong selection ($\alpha = 400$) are given in Figure 2.2, with α being the population scaled selection coefficient $\alpha = 4Ns$. As can be seen from the figure, the mode is an accurate estimator of the true value for both models. However, in the SSV case the posterior distribution is much broader than under the SDN model, and the 95% confidence interval extends to the edges of the prior, indicating low accuracy in the estimate. For the initial frequency parameter, f_1, the posterior differs only marginally from the prior, and therefore this parameter cannot be reliably estimated.

2.2.2 Accuracy of Model Choice

We aim to identify parameter regions where we can distinguish between the SSV and the SDN model. As a control, we also consider a model of neutral evolution (NT), where an allele increases to high frequency solely due to genetic drift. In particular, we are interested in three parameters that are expected to have a strong influence on model choice accuracy: the selection parameter α, the frequency of the mutation when it became selective advantageous, f_1, and the current frequency of the selected allele f_{cur}. In Figures 2.3 and 2.6, we explore the accuracy of our model choice procedure in three series as a function of α, f_1, and f_{cur}. We find (Figure 2.3a) that in cases where $\alpha < 100$; the method cannot reliably distinguish between selection and a neutral model. This is not surprising, as for such values of α, standard neutrality tests have little or no power to detect selection (Simonsen et al., 1995; Teshima et al., 2006). For selection coefficients of $\alpha = 100$ and $\alpha = 200$ the neutral model has a very low posterior probability and would be rejected, but we still do not have sufficient power to distinguish the signals from SSV from SDN. Only under strong selection ($\alpha = 1,000$) do we have reasonable power to distinguish between SSV and SDN. Thus, we find that there is a parameter range of α between 100 and 500, in which selection can be reliably detected, but the two models of selection are statistically indistinguishable. In the second series (Figure 2.3b), we vary the initial allele frequency (f_1). We find that simulations under the SSV model, with $f_1 = 1\%$, are identified as SDN models, but that the accuracy in model choice increases with f_1. For larger values of f_1, we can detect selection when selection is strong (α =1,000). For high initial allele frequency ($f_1 = 20\%$) we correctly infer the true mode of selection even when α is 200. This suggests that the ability to distinguish the two models increases with f_1. Furthermore, we also find a negative relationship between the estimated value of f_1 for a data set and the posterior probability of the SDN model (Figures 2.7, 2.8 and 2.9): As we would expect, the larger the estimate of f_1, the lower is the posterior probability of the SDN model, and we find a strong negative correlation ($R^2 = 0.51$) between these two quantities based on 1,000 simulations. In the third series (Figure 2.3c), we investigate the effect of the current allele frequency f_{cur} on the model comparison. For simulations under the SSV model, we find that the accuracy strongly decreases with f_{cur}. For $f_{cur} = 0.2$, we classify slightly less than half of the data sets correctly. This is in contrast to simulations under the SDN model, where the power to correctly classify simulated data sets gradually increases with f_{cur}. Thus, in studies aimed at detecting selection on standing variation, the false positive rate should depend only slightly on f_{cur}, but the false negative rate is expected to increases drastically when f_{cur} is low. Figure 2.4 illustrates how the selection parameter (α) and the initial allele frequency (f_1) affect the accuracy of model choice between the SSV, SDN and NT models for three values of f_{cur} ($f_{cur} = 0.95$, $f_{cur} = 0.8$ and $f_{cur} = 0.5$). As in Figure 2.3, the number of correctly assigned data sets increases with α, f_1 and f_{cur}. Under the SSV model, the gradient with which the power declines is strongest when f_{cur} is large (95%, Figure 2.4a), and becomes less pronounced for smaller f_{cur} (see Figure 2.4c and 2.4e). For $f_{cur} = 95\%$ (Figure 2.4a), there is a region with $f_1 > 0.05$ and $\alpha > 1,000$ where there clearly is very high power to infer the correct model. On the other hand, for $\alpha < 200$ or $f_1 < 0.03$, we make incorrect inferences more than half of the time, indicating that in these regions of the parameter space, the signal of the sweep is too weak to discriminate between the SSV and SDN models. While that global pattern is the same for $f_{cur} = 0.8$ and $f_{cur} = 0.5$ (Figure 2.4c,e), the distinction between regions where we can and cannot assign simulated data sets correctly is less pronounced. Quite surprisingly, however, we find that for $f_{cur} = 0.8$, the number of correctly assigned data sets increases when selection is low. The same trend holds for $f_{cur} = 0.5$ (Figure 2.4e), however here the influence of selection is even weaker, and inference becomes quite ambiguous, with posterior probabilities ranging from 60% to 80% in the entire parameter space. In contrast, the pattern is much simpler for simulations under the SDN model (Figures 2.4b, 2.4d and 2.4f), where the probability to correctly identify the model increases with decreasing f_{cur}. When f_{cur} is set to 0.95, we need a selection coefficient of $\alpha = 1,500$ to make confident inferences. For f_{cur} of 0.8 and 0.5, this value decreases to 900 and 300, respectively. In summary, a high current allele frequency increases the power to distinguish between SSV from SDN (Figures 2.3c, 2.4). The frequency with which the SDN model is correctly inferred increases slightly with decreasing f_{cur}, presumably because the selected phase makes up a larger proportion of the trajectory.

2.2.3 Applications

We illustrate our model choice procedure by analyzing seven genes that have previously been identified as candidates for being under selection. These genes are ADH1B, ASPM, EDAR, G6PD, LCT, PSCA and TRPV6. The genes were selected using the following set of criteria: i) there is evidence for selection from a previous study, ii) a putative causal mutation has been identified and iii) the putative causal site has reached a high frequency in at least one population, but has not yet reached fixation. In addition, we also analyzed four regions that were noncoding and presumably neutral. We retrieved polymorphism data from the 1000 Genomes Project low coverage data (1000 Genomes Project Consortium, 2010) using tabix (Li, 2011). Ancestral genotypes were inferred by comparison to the homologous chimpanzee allele. If a signal of selection was present in more than one population, we used data for the population where the selected site was most frequent, to facilitate inference. Model choice and parameter estimation were performed using the procedures described in the methods section. In contrast to the inference on simulated data sets, here we explicitly model varying recombination rates and the complex demographic history of the human population. Results for the sample genes are given in Figures 2.5 and 2.9, as well as Table 2.1. For six of the seven genes analyzed, the neutral scenario was strongly rejected with a posterior probability of less than 1%, and we can confirm the prior evidence that these genes are under selection. Three of those genes, ADH1B, EDAR and LCT, were found to be under selection from a new mutation and one gene, TRPV6 could not be assigned with any significant probability to either model. Two genes, ASPM and PSCA, were found to be under selection from standing variation. Finally, none of the three models provided a good fit to observed data in the G6PD gene, suggesting that neither of the models is appropriate for this gene. In the following paragraphs, we will discuss each gene in some detail, and give estimates for selection coefficient and time when appropriate. All estimates are given with a point estimate for the mode, and the lower and upper bound of a 95% Highest Posterior Density interval in brackets. Estimates in years were made assuming a generation time of 25 years.

2.3 Discussion

2.3.1 Applications

2.3.1.1 ADH1B

The ADH1B gene encodes one of three subunits of the Alcohol dehydrogenase (ADH1) protein, a major enzyme in the alcohol degradation pathway that catalyzes the oxidization of alcohols into aldehydes. ADH1B is part of a 60kb gene cluster on chromosome 4, encoding for all three ADH1 subunits. Selection on the major ADH gene complex has received major attention as it is suggested to be one of the major genetic causes of alcoholism risk (Li et al., 2007), and a possible cause of the “alcohol flush” phenotype prevalent in many Asian populations, where individuals turn red due to increased acetaldehyde levels in the blood after alcohol consumption (Peng et al., 2010). As a result, the genes are well studied and several non-synonymous polymorphisms are known to have various effects on enzyme activity (Eng et al., 2007; Osier et al., 2002). One particular allele, Arg47His, has been proposed to be under selection based on several lines of evidence: First, the derived Histidine allele results in an increased enzymatic activity. Second, age estimates of the derived allele based on its frequency correlate with the onset of rice domestication (Li et al., 2007; Peng et al., 2010) and the availability of fermented beverages (McGovern et al., 2004). In our analysis, we analyzed the CHB population where the allele is found at a frequency of 0.71 in the 1000 genomes data. For this data set, we could clearly reject the neutral model, with a posterior probability of 10^{-8}. The SDN and SSV models have posterior probabilities of 78.3% and 21.7%, respectively, indicating slightly stronger evidence in favor of the SDN model. Under this model, we estimate a selection coefficient of $s = 0.036$ (0.009 - 0.19), and an age of the mutation of 11,100 (1,900 - 42,900) years It is remarkable that this age corresponds very well with the arrival of rice agriculture and the availability of fermented beverages in China around 10,000 year ago (Peng et al., 2010). Our finding of evidence for a *de novo* sweep is conflicting with the fact that the derived 47His allele also occurs at a high frequency in Western Asian populations, but only at low frequencies in Central Asian and Indian populations (Li et al., 2007), a pattern of genetic variation that has previously been suggested to be a result of selection on standing variation (Li et al., 2007).

2.3.1.2 ASPM

The ASPM (abnormal spindle-like microcephaly associated) gene has been identified as a major determinant of brain size (Bond et al., 2002). Much attention has been focused on the difference between humans and chimpanzee in that gene, and several studies (Kouprina et al., 2004; Zhang, 2003) have quantified these differences and found an unusual high amount of fixed substitutions between these two species, indicating positive selection on the branch between humans and chimps. In addition, recent ongoing selection was proposed based on the finding that a single haplotype was unusually frequent in several populations (Mekel-Bobrov et al., 2005). However, the interpretation of their results stirred considerable debate (Currat et al., 2006; Mekel-Bobrov and Lahn, 2007; Timpson et al., 2007; Yu et al., 2007), with researchers pointing out that the haplotype distribution found by Mekel-Bobrov et al. (2005) is not that unusual (Yu et al., 2007) and that neutral demographic scenarios are able to produce haplotype distributions similar to the one observed in ASPM (Currat et al., 2006). We used the non-synonymous SNP A44871G (rs41310927) for our study, which was identified in (Mekel-Bobrov et al., 2005) as a putative causal variant in our analysis. We found evidence for selection on standing variation ($\mathbb{P}(SSV) = 0.87$), with little support for the neutral and SDN model with posterior probabilities of 0.13 and 2×10^{-7}, respectively. We estimate a rather weak selection coefficient of 0.029 (0.003-0.170), and estimate that selection started to act

17,412 (771-56,443) years ago, and an age of the mutation of 97 (17-289) ky. This is considerably older than the estimate of 5,800 years for the most recent common ancestor of the selected allele by Mekel-Bobrov et al. (2005), a difference that might be due to the fact that we assume a different demographic history.

2.3.1.3 EDAR

The EDAR gene region has been suggested to be under selection in East Asians based on multiple genome scans (Akey et al., 2004; Voight et al., 2006; Williamson et al., 2005) and has been studied in more detail by Bryk et al. (2008). EDAR encodes a cell-surface receptor that activates a transcription factor (Bryk et al., 2008; Fujimoto et al., 2008), and, among other phenotypes, has been associated with the development of distinct hair and teeth morphologies (Fujimoto et al., 2008; Kimura et al., 2009). A non-synonymous SNP (rs3827760, V370A) has been associated with these phenotypes, and has been confirmed in an in vitro study to enhance the activity of the EDAR gene (Bryk et al., 2008). The rs3827760 SNP lies in a DEATH-domain that is highly conserved within mammals (Sabeti et al., 2007), and is found at a very high frequency in East Asian and American individuals, but is absent from all European and African populations (Bryk et al., 2008). In the 1000 genomes data, EDAR shows the strongest signal of selection for EHH, Tajima's D and Fay & Wu's H among all genes we analyzed. This is reflected in our model choice analysis, where we find a 88.5% probability that the V370A polymorphism originated from a new mutation. The probability for the SSV model was 13.3%, and the neutral model did not receive any measurable support. We estimated a very high selection coefficient of $s = 0.15$ (0.04, 0.31), and an origin of the mutant allele 3,000 (1,400, 6,900) years ago. This estimate is most likely too young, as the allele is also present in Native American population and so is strongly expected to have been present before the colonization of America. A possible explanation for this is that selection does not act codominantly on EDAR. Comparing our codominant model with a model where the dominance parameter h was allowed to vary between 0 and 1 resulted in a strong favor for the more complex model (Bayes Factor = 36). Under this model we estimate a selection coefficient of $s = 0.14$ (0.07-0.31), but a much older age of the allele of 11,400 (4,300-43,700) years. This is at the lower end of estimates for the time of colonization of the Americas (Fagundes et al., 2007; Waters and Stafford, 2007), indicating that the derived allele might have moved into the American populations at a low frequency. This hypothesis is consistent with the very high divergence of the EDAR region between the Mexican and Chinese populations, where we find an F_{ST} of 0.36 (excluding the conserved DEATH-domain), which is much higher than the genome-average F_{ST} of 0.069 between these two populations (Altshuler et al., 2010). This may indicate that the 370A allele has risen in frequency largely independently between these two populations.

This is in contrast to the analysis of Bryk et al. (2008), who estimated that the derived 370A allele has been fixed 10,740 years ago. However, both in the 1,000 genomes data and the data of Bryk et al. (2008), the site is still segregating within the CHB population.

While we cannot exclude the scenario of fixation and recent reintroduction of the ancestral allele, the high divergence between Native Americans and East Asians seems to favor a more recent sweep.

2.3.1.4 G6PD

The G6PD gene is located on the X chromosome, and is one of the best studied cases of selection in humans (Sabeti et al., 2002; Saunders et al., 2002; Tishkoff et al., 2001; Verrelli et al., 2002). The G6PD gene encodes the Glucose-6-phosphate dehydrogenase protein, the first enzyme in the pentose phosphate pathway. The G6PD gene has long been associated with reduced-efficiency erythrocytes (Beutler, 1994; Carson et al., 1956), and several hundred variants causing various levels of reduction in catalytic activity have been discovered (Nkhoma et al., 2009), leading to a significantly reduced fitness in affected individuals. As a benefit, however, G6PD deficiency provides resistance to malaria (Ruwende et al., 1995) and therefore even strongly deleterious alleles rise to considerable frequencies in populations where malaria infections are epidemic. Due to these antagonistic selective pressures, G6PD in populations affected by malaria is one of the best examples of balancing selection described in the human genome. We use the A/A- polymorphism (rs1050828), identified by (Tishkoff et al., 2001) as the putative site under selection. When applying our method, however, none of the models provided a good fit to the data, indicating that the models we used are too simplistic for the complicated history of G6PD (see Figure 2.5). The combination of summary statistics with a low EHH, very low IHS and high, non-significant values for Tajima's D and Fay and Wu's H cannot be captured by either of our models. This is not surprising given that the selection on the G6PD locus cannot be described as a selective sweep, but is the effect of balancing selection. It is encouraging the method in this case indirectly, through a poor model fit, helps determine that the simple selective sweep models considered here are not appropriate for this locus.

2.3.1.5 LCT

In most mammals, the ability to digest lactose, a common disaccharide in milk, decreases when they stop being milk-fed. In contrast, in many humans the main enzyme used to digest lactose into monosaccharides, continues to be expressed even in adults, a phenotype known as lactase-persistence (Bersaglieri et al., 2004; Enattah et al., 2002; Hollox et al., 2001). Several presumably independent alleles have been identified that confer the same phenotype (Tishkoff et al., 2007) in different populations. The first and possibly best-characterized allele is the C/T-13910 polymorphism (rs4988235) that is particularly prevalent in Northern European populations and has been shown in Finnish populations to be 100% associated with the lactase phenotype (Kuokkanen et al., 2003). Further evidence that the T-13910 allele is causal for the persistence phenotype is given by in vitro analyses (Olds and Sibley, 2003; Troelsen et al., 2003) that found increased enhancer activity.

We analyzed the FIN population from the June 2011 data release of the 1,000 Genomes Project, using the C/T-13910 polymorphism as the selected site. We found a 98.7% posterior probability for the SDN model and only a 1.2% posterior probability for the SSV model, indicating that this particular LCT allele most likely was under selection shortly after it arose. We estimated a rather low selection coefficient of 0.025 (0.003-0.19), and an origin of the mutation 11,200 (1,500-64,900) years ago. Our estimate is much older than the estimates by Bersaglieri et al. (2004), who estimated a selection coefficient between 0.09 and 0.19, and an age of the mutation of 1,625-3,188 years using a deterministic approximation based on the observed frequency of the allele. The fact that they used a deterministic approximation may explain the fact that we have wider confidence intervals. Our estimate is more consistent with the estimate of Tishkoff et al. (2007) who used the width of the sweep region to date the selected allele to an age of 7,998 years and obtained an estimated selection coefficient of 0.069. Our estimates are also in good concordance with the estimate of Itan et al. (2009). In their study, they modeled the spread of lactase persistence through Europe using a spatially explicit ABC model, which takes advantage of the arrival of dairy farming in various locations. They estimated a selection coefficient of 0.095 in dairy farmers and a slightly older age for the selected allele (7,441 years). While all studies suggest a more recent origin of the selected allele, we note that the confidence intervals on both the selection coefficient and age of the sweep overlap between all four studies. A complimentary approach to dating the age of an allele, and estimating selection coefficients from modern DNA data, is the usage of ancient DNA (Burger et al., 2007; Malmström et al., 2010; Plantinga et al., 2012). Indeed, the derived allele of the LCT the C/T-13910 polymorphism as was found in a single copy in a 5,000 year old sample from Sweden (Malmström et al., 2010), and at a higher frequency of 27% in the Basque country in a sample of approximately the same age (Plantinga et al., 2012). In contrast, the derived allele was absent from an Eastern European sample roughly 7,000 years old (Burger et al., 2007). These findings are in good agreement with our estimates and other estimates on genetic data. Based on this ancient DNA evidence, it has been speculated that the LCT allele may have swept from standing variation (Plantinga et al., 2012), mainly due to the fact that the derived allele is found at a rather high frequency only two millennia after the introduction of agriculture in that population. However, if the allele was mostly neutral before the arrival we would expect it to be rather old, and in particular we might also expect to see the derived T allele in African populations, which is not the case. Calculating the expected age of an allele at a frequency of 27% (Kimura and Ohta, 1973; Maruyama, 1974) results in expected ages of 480ky and 6,500 (2,500 -36,000) years for neutrality and selection, respectively, using our estimated selection coefficient and an effective population size of 10,000. While these estimates based solely on allele frequencies should be interpreted with great caution, they nevertheless show that our estimate of a *de novo* selected mutation is consistent both with the observed allele frequencies around 5,000 years ago and an assumed origin of dairy farming 11,000-12,000 years ago.

2.3.1.6 PSCA

The prostate stem cell antigen gene (PSCA) on chromosome 8 has been proposed to be under selection by (Bhatia et al., 2011) based on an analysis of population differentiation in a global array of human populations. A non-synonymous SNP in PSCA (rs2294008) is known to be involved in various forms of cancer (Wu et al., 2009), and we therefore used it as the causal site in our analysis. Interestingly, the derived allele is present in all human populations although the frequency varies considerably between different populations (Bhatia et al., 2011). The highest derived allele frequencies of more than 75% are reported in West African and East Asian populations, whereas some sub-Saharan African and most Native American populations have allele frequencies below 20%. This worldwide distribution of the allele was interpreted as evidence of selection from standing variation (Bhatia et al., 2011). Our analysis confirmed this hypothesis based on analyses of data only from the Yoruban population, with the SSV model receiving a posterior probability of 86.0%, compared to a posterior probability of 23.9% for the SDN model, and 1.2% for a neutral model Under the SSV model, we estimate a selection coefficient of 0.035 (0.004 - 0.15), with selection having started 8,000 (1,000-54,900) years ago, and the allele being 191 (50 698) thousand years old. The fact that the mutation is distributed globally supports our inference of a sweep based on standing variation.

2.3.1.7 TRPV6

TRPV6 is in the heart of a 115kb region on chromosome 7 that has been reported to be under selection (Akey et al., 2004; Stajich and Hahn, 2005) and has been closely investigated by Akey et al. (2006). TRPV6 codes for a protein subunit that encodes cation pores, particularly for calcium ions (Akey et al., 2006; Birnbaumer et al., 2003). TRPV6 was found to be in a region of accelerated evolution on the human lineage, as indicated by an elevated ratio of non-synonymous to synonymous fixed differences (Akey et al., 2006). In particular, three non-synonymous mutations segregating in humans were found, with a striking diversity pattern; the derived allele was at an intermediate frequency in all African population, but at frequencies of 90% and more in the rest of the world. In addition, both Tajima's D and Fay and Wu's H statistics were significantly negative for non-Africans and European-Americans. For this reason, we restricted our analysis to the CEU population, and used the first of the non-synonymous SNP (rs4987682) as the focal site for our analysis, as it was the only one that was in the N-terminal region of the TRPV6 protein, the suggested target of selection (Akey et al., 2006). While the neutral model could be rejected with a posterior probability close to zero (8×10^{-7}), the separation of SSV and SDN model remained inconclusive, with posterior probabilities of 0.55 and 0.45 respectively. The estimate of the selection coefficient was very similar for both models $s_{SSV} = 0.032$ (0.005-0.25), $s_{SDN} = 0.023$ (0.007-0.08), but the confidence interval is much smaller under the SDN model, as expected. Furthermore, the estimated age of the allele differed between models: Under the SSV model, the mutation is inferred to be 211 (29-697) ky old, but became selected only 7,600 (900-43,300) years ago. Under the SDN model, the mutation arose and became selected 23,400 (6,400-70,400) years ago. These findings are in good concordance with the patterns of diversity found previously, (Akey et al., 2004, 2006) and in particular the evidence that the signature of selection is shared between all non-African populations and thus selection started likely less than 100,000 years ago. Also, the estimate under the SSV model that selection started less than 10,000 years ago is concordant with the role of TRPV6 in absorbing calcium (Akey et al., 2006).

2.3.1.8 Neutral Regions

In addition to these genes, we also analyzed four putatively neutral regions that were 5Mb away from our candidate genes. This distance should be big enough that the neutral region are not impacted by the selective sweep, but are likely influenced by the same mutational processes as the selected regions. For all these regions, the neutral model had the highest posterior probability, with posterior probabilities of 0.758, 0.932, 0.994 and 0.999 for the four regions. This indicates we are indeed able to discern selected from neutral regions.

2.3.2 Conclusions on Data Applications

The distribution of summary statistics in Figure 2.10 illustrates the impact of choice of summary statistic for model inference (Marin et al., 2011). Very high values of EHH are clearly indicative of the SDN model, at both a 10kb and 20kb distance. Both the SSV and SDN models are associated with low IHS values, whereas the neutral regions have IHS values closer to zero. Tajima's D and Fay and Wu's H are both very informative for model comparison, with SDN genes having very low D values, SSV genes having D values close to zero and neutral regions having positive D values. The main exception is the LCT gene, however, which we inferred to be selected from a *de novo* mutation, but which has a high D. The signal for SDN apparently comes more from the high EHH and low IHS values.

In general, our results are highly concordant with previous studies of these genes. Our estimates tend to be slightly older for several genes, which might be due to us assuming a recent bottleneck and thus a period of relaxed selection, whereas the majority of other authors assumed a constant population size for their timing inferences (Peng et al., 2010), which results in a stronger signal of selection (Figure 2.6). For one gene, G6PD, we could not make any inferences, because we could not reproduce the observed pattern of diversity using simulations of positive directional selection. G6PD shows an extremely narrow region of reduced diversity, surrounded by a region of high diversity. This may be due to balancing selection between malaria resistance and reduced efficiency oxygen transport introducing a signal that cannot be reproduced by our simple model of directional selection. Alternatively, the X-linked mode of inheritance of this locus is not concordant with the assumptions of our model. This also highlights one of the dangers of ABC: It is crucial that the models investigated are able to reproduce the data observed; otherwise false inferences may be drawn. This danger inherent to any ABC approach is also highlighted by the fact that misidentification of the selected site will bias model choice results towards SSV (Figure 2.11). This can be explained by the fact that even if the neutral site is closely linked to the selected site, it is likely to "escape" the sweep by recombining away from the selected haplotype, thus giving the signal of selection from standing variation. Similarly, analyzing data simulated under a population bottleneck under a constant size model will bias the results towards stronger selection and SDN (Figure 2.12), presumably due to the younger age of mutations being taken as evidence of strong selection.

2.3.3 Model Choice Accuracy

We have shown that it is much more difficult to estimate the model parameters α, and t_1 from the SSV model than from the SDN model. This is unsurprising, as the SSV model has been shown to have a higher variance in allele age, which results in a higher expected variance for most summary statistics (Innan and Kim, 2004). We further show that there is not enough information to estimate the initial frequency of the sweep f_1. This is unsurprising, as the exact position of the switching point has likely only a minor effect on the data, especially as the effect of selection on the trajectory is weak when the allele frequency of the beneficial allele is low (Kaplan et al., 1989). We further notice that the accuracy of our model choice procedure decreases when the signal of selection is weak. Consistent with previous findings, selection is very hard to detect if α is below about 100 (Hermisson and Pennings, 2005; Teshima et al., 2006; Williamson et al., 2005). This is also the point where our method gains power to distinguish between SVN from SDN. The initial frequency required to detect standing variation is moderate at around 3% for weak selection and 2% for stronger selection. However, selection has to be rather strong, at around $\alpha = 1{,}000$ and initial frequencies have to be above 5% to allow accurate inference. Presumably, this is because below this threshold, the stochasticity of the trajectory is very large even under selection, and the difference between the two scenarios is small (see also Figure 2.1c). These findings are not particularly surprising, as selection scans based on summary statistics have been shown in general to have low power under these conditions (Teshima et al., 2006). These findings certainly limit the scope of our approach. Could we do better with a different strategy? As discussed in the introduction, the ABC approach simplifies data in two ways. First, instead of using the full data, we use an array of summary statistics. Second, we substitute an exact match between observation and simulations with an approximate match, depending on "close" simulations. Regarding the use of summary statistics, we note that summary statistics have been widely used to detect selection from genetic data (Akey et al., 2004; Nielsen et al., 2005; Sabeti et al., 2002;

Voight et al., 2006; Williamson et al., 2005), and currently provide the only way to detect selection from DNA sequence data. No full likelihood based method is available to detect selection from DNA sequence data that could be adapted to distinguish between the two sweep models entertained here. The second simplification step is based on the number of simulations performed and the tolerance interval and is imposed by computational constraints. We examine the effect of different numbers of simulations and tolerance cutoffs on our results by calculating relative error rates of the posterior mean and the false negative rate of the model choice. We show in Table 2.2 that increasing the number of simulation by a large amount or changing the rejection parameter does not significantly improve our results, indicating that we do not lose a lot of information at this stage. This shows that the ABC approach reliably estimates the posterior based on the summary statistics, and as such use all the information available in these statistics. Statistics such as EHH, iHS, Tajima's D, etc., do not contain information that will allow us to provide more reliable estimates. It the light of this, it may appear disappointing that our method does not provide more accurate parameter estimates and more power to distinguish between models. However, it is important to realize that as previously argued, all information regarding selection is in the frequency path of the selected allele (Hudson and Kaplan, 1988; Kaplan et al., 1989). For relatively small selection coefficients and/or small initial frequencies of the selected allele, the paths are very similar for the SSV and SDN models. Even if a full likelihood method could be developed, it is unlikely that it had much more power to distinguish between models. A further simplification in our method is the restriction to a single population. Population differentiation measures, such as F_{ST}, are one of the most successful ways to detect sweeps from standing variation (Barrett and Schluter, 2008; Bhatia et al., 2011), and the inclusion of more realistic models of demography may improve our accuracy. Such models, however, require an additional estimation of multi-population demographic history, which greatly increases the complexity of the model. While we applied our method only to human candidate loci, it should be possible to easily translate it to other species. In particular, as our simulation results suggest that we have more power to distinguish SDN and SSV if selection is strong, species with large population sizes, such as e.g. *Drosophila* or many microorganisms may be very promising targets for a similar study. Another possible target might be species with very strong artificial selection, such as domesticated animals or plants, where we may gain valuable insights on the domestication history of these species. Of course, our approach could also be combined with ancient DNA (e.g. Plantinga et al., 2012), which could provide much narrower confidence intervals on time estimates and also help improve estimates of selection coefficients.

The two selection models we consider here, the SSV and the SDN models, are nested models, Setting $f_1 = 1/2N$ in the SSV model recovers the SDN model. To facilitate Bayesian model choice we assign positive probability to $f_1 = 1/2N$, and base our inferences on a choice between $f_1 = 1/2N$ and $f_1 \sim U(0, 0.2)$ (See Methods). ABC based model choice has recently been criticized and been shown to be biased in some cases where the statistics used are not sufficient (Didelot et al., 2010; Robert et al., 2011). While some of the specific issues raised by (Birnbaumer et al., 2003) are not applicable in our setting because we consider nested models, we do not base our inference on sufficient statistics and the statistical properties of our model choice procedure are, therefore, largely unknown. To address this issue, and in general to validate our approach, we use a method introduced in (Cook et al., 2006). We show in Figure 2.13 that our estimated probabilities only show bias for very small values of the Bayes factor, where there appears to be a bias towards inference of the SDN model for simulations generated under the SSV model with very low values of f_1.

2.4 Methods

2.4.1 Models

In order to keep our problem simple, we condition on two important parameters: We assume that the exact site under selection is known from extraneous information, and we furthermore assume that the allele frequency f_{cur} of that site at the time of sampling, $t_cur = 0$ is known. The interpretation of the parameters is depicted graphically in Figure 2.1a. Unless noted otherwise, we assume a panmictic diploid population of size N with an additive selection model where the ancestral homozygous, heterozygous and derived homozygous genotypes have fitness 1, $1 + s/2$ and $1 + s$, respectively. However, the methodology applied here can easily be adapted to more complex scenarios, e.g. models involving multiple populations, more sophisticated demographic models, and other models of selection. For most simulated data sets, we will report the population scaled mutation rate $\alpha = 4Ns$, as the shape of the allele frequency trajectory depends only on that compound parameter (Ewens, 2004). However, for most of the genes we analyze previous estimates were made directly on s rather than the compound parameter. To facilitate comparisons, we report s for the genes we analyzed.

2.4.1.1 Sweep from a *De Novo* Mutation Model

The sweep from a *de novo* mutation (SDN) models a single selective sweep and has two parameters: the mutation rate μ, and the selection coefficient s. For all simulations, we follow Itan et al. (2009) and record the time t_0 when the mutation arose, as t_0 depends stochastically on s. The prior distributions we use for this model were $\mu \sim U(0.5 \times 10^{-8}, 6 \times 10^{-8})$ and $\log(s) \sim U(-3, -0.5)$, where U is a uniform distribution.

2.4.1.2 Sweep from Standing Variation Model

The sweep from standing variation (SSV) model is identical to the *de novo* mutation model, with the exception that we define a frequency f_1 at which the mutation becomes selected. Unless noted otherwise, the priors for μ and s are the same as in the SDN model, and the prior for f_1 is $f_1 \sim U(0, 0.2)$. In addition to t_0, which is defined analogous to the

SDN model, we are also interested in t_1, the time when the mutation becomes selectively advantageous (i.e. the time when the mutation reaches frequency f_1).

2.4.1.3 Neutral Model

We also consider a neutral model (NT), without any selection. The only free parameter in this model is the mutation rate μ, with the same prior distribution as described under SDN model. As under the selection model, however, we still condition on one site having reached a final allele frequency of f_{cur}, so this model does not correspond to the classical neutral coalescent.

2.4.2 Approximate Bayesian Computation

We use a standard ABC approach (Beaumont et al., 2002; Tavaré et al., 1997), using a post-sampling adjustment in the form of a GLM (Leuenberger and Wegmann, 2010). We used the ABCToolbox package (Wegmann et al., 2010), for specifying priors, rejection sampling and post-sampling adjustment. Unless specified otherwise, we perform 10^5 simulations per model, and retained the 100 (0.1%) simulations with associated Euclidean distance between observed and simulated summary statistics closest to zero. To assess how the number of simulations and acceptance rates influence our results, we analyze 10,000 random data sets with up to 10^7 simulations and varying acceptance rates. We show that these parameters have very little impact on the relative error for both the model choice and parameter estimates in Table 2.2.

2.4.2.1 Details of Statistics Used

We use a diverse array of summary statistics, with the goal of maximizing the information captured, while not incluing any statistics that just add noise. The statistics we used may be broadly classified into statistics based on haplotype patterns, and statistics based on the site frequency spectrum. The haplotype based statistics we used were iHS (Voight et al., 2006) and EHH citepsabeti2002. We recorded EHH in a 10kb, 20kb and 50kb window, centered on the selected site. For the SFS based statistics, we used Tajima's D (Tajima, 1989), Fay & Wu's H (Fay and Wu, 2000), the average number of pairwise differences π, and the number of segregating sites S as statistics. All these statistics are calculated for three regions: A central region of 20kb around the selected site, an intermediate region consisting of all sites 20-50kb away from the selected site, and a faraway region consisting of all sites further than 50kb away from the selected site. Following Wegmann et al. (2009), we linearize all statistics using a Box-Cox-transformation (Box and Cox, 1964). To choose a set of informative summary statistics, we used a Partial Least Squares Discriminant Analysis (PLS-DA) (Lê Cao et al., 2009; Tenenhaus, 1998). PLS-DA is a variant of Partial Least Squares regression, that, similarly to principal component analysis, extracts orthogonal components from a high-dimensional data set (in this case the summary statistics). In contrast to PCA, in PLS-DA these components are chosen such that the covariance between summary statistics and models is maximized (see e.g. Boulesteix and Strimmer, 2007). We did our computations using the “psda” function of the mixOmics package for R, and kept the five first PLS-DA axes (Lê Cao et al., 2009). ABC has two crucial parameters independent of the model it is applied to: The number of simulations n_S and the acceptance rate ϵ. To assess the effect of these parameters on inference, we calculate the accuracy of our model choice estimates for various values of n_S and ϵ (Table 2.2).

2.4.3 Simulations

All our data sets used for both the ABC inference and the assessment of our procedure are simulated using a modified version of the coalescent simulator mbs (Teshima and Innan, 2009). Mbs allows simulation of genetic data sets with a single selected site using the structured coalescent (Hudson and Kaplan, 1988). Mbs first simulates the allele frequency trajectory of the site under selection, and then generates a data set conditional on that trajectory. We simulate allele frequency trajectories using Euler’s method on the unscaled backwards diffusion equation with selection (eq 7.1. in Kimura, 1964). This equation makes it very easy to incorporate population size changes by just changing the variance term. To simulate sweeps from standing variation, we set the selection coefficient (s) to zero the first time the trajectory reaches f_1. To analyze simulated data sets, we generally simulate a 100kb region with a recombination rate of 1.5cM/Mb. For the human genes, we simulate the gene and a 50kb flanking region on both sides, resulting in regions that are usually between 100kb and 150kb wide. Recombination rates and hotspots are modeled by using the HapMap recombination map (Myers et al., 2005) in the application to selected genes. For all simulated data sets we assume a constant-sized population. For the analysis of human genes, we use the population history estimated by (Li and Durbin, 2011). Specific regions and details of the used regions are given in Table 2.3. To ensure that our method does not suffer from a high false positive rate, we also analyze regions 5Mb downstream from the candidate genes, as they are presumably neutral. For three of genes (ASPM, G6PD, and PSCA), no data was available for these downstream regions, so we analyzed the remaining loci. Candidate loci for selection were chosen using the following criteria: i) they were required to have a derived allele frequency between 0.7 and 0.9 and ii) to be as closely to 5Mb away from the actual candidate locus in the upstream gene as possible. We estimate parameters from our models using the standard ABC procedure described above. The parameters we estimate are the mutation rate μ, the age of the sweep t_1 the selection coefficient s and, only under the SSV model, the initial frequency f_1 for the SSV model. In particular we want to determine if our posteriors are unbiased, and if we were able to get reasonable confidence in our estimates. To do this, we simulate data sets with fixed parameters and plotted the average posterior distribution for all parameters in Figure 2.2.

2.4.4 Model Choice

For model choice, our main goal is to calculate the relative probabilities of the models given the data, i.e. $\mathbb{P}(SSV|data), \mathbb{P}(SDN|data)$ and $\mathbb{P}(NT|data)$, which we calculate using the marginal densities as proposed by Leuenberger and Wegmann (2010). To identify parameter regions where there is power to distinguish between the models, we simulate 1,000 data sets each under 30 different scenarios in three series, corresponding to three parameters of interest: The strength of selection α, the frequency when the mutation became selective advantageous f_1 and f_{cur}, the frequency at which the mutation is observed. To test the algorithm for approximating Bayes factors, we also use a simulation approach. The estimator of the posterior probability from k simulations, $\hat{p}_k(m|x)$, should have the property $\lim_{k\rightarrow\infty} \hat{p}_k(m|x) = \mathbb{P}(m|x)$ where m is a model indicator function for a specific model. Also, for a particular draw from the posterior, $m^{(0)}$, we expect $\mathbb{P}(m^{(0)} = m|x) = \mathbb{P}(m|x)$, if the simulation algorithm works properly. In other words, $\hat{p}_k(m|x)$, should asymptotically equal $\mathbb{P}(m^{(0)} = m|x)$, i.e. if $\hat{p}_k(m|x) = c$, we expect a proportion c of simulations to have been obtained from model m. Equivalently, for an estimated log Bayes factor B, $\log B = c$, we expect a proportion $10c/(1 + 10c)$ of draws to be from model m. This prediction is tested in Figure 2.13, based on 10,000 random data sets from both the SDN and SSV model.

2.5 Tables

Table 2.1. Genes analyzed in this study. Chr: chromosome, pop: population we analyzed using the population code from the 1000 genomes project; For each gene, we give the favored model(s) and in brackets the posterior probability for that model.

					Estimates			
Gene	chr	function	pop	Model	S	t_1 (years)	t_0 (years)	References
ADH1B	4	Alcohol metabolism	CHB	SDN(0.78)	0.036 (0.009-0.192)	-	11,100 (1,900-42,900)	Osier et al. (2002)
ASPM	1	microcephalism	GBR	SSV(0.87)	0.029(0.003-0.17)	17,400 (800-56,400)	79 (17-288) ky	Mekel-Bobrov et al. (2005)
EDAR	2	NF-B Activation	CHB	SDN(0.88)	0.14 (0.07- 0.31)	-	11,400 (4,300 -43,700)	Bryk et al. (2008)
G6PD	X	malaria resistance	YRI	-	-	-	-	Tishkoff et al. (2001)
LCT	2	lactase persistence	FIN	SDN(0.99)	0.025 (0.004-0.20)	-	11,200 (1500-64,900)	Hollox et al. (2001)
PSCA	8	cancers	YRI	SSV(0.86)	0.035 (0.004 - 0.015)	8,000 (1,000 54,900)	191 (50 698) ky	Bhatia et al. (2011)
TRPV6	7	Calcium absorption	CEU	SSV (0.55)	0.032 (0.005-0.25)	7.600 (900-43,300)	211 (29-697)ky	Akey et al. (2006)
				SDN (0.45)	0.023 (0.007-0.08)	-	23,400 (6,400-70,400)	

Table 2.2. Details of genes and neutral regions analyzed in this study. Chr: chromosome, pop: population we analyzed. All positions given are on the hg19 build of the human genome.

						Selected site		
Gene	chr	function	freq	pop	Region sim	dbSNP id	Position	size
ADH1B	4	Alcohol metabolism	0.71	CHB	100177528-100292572	rs1229984	100239319	115044
ASPM	1	microcephalism	0.48	GBR	197007250-197165824	rs41310927	197070697	158574
EDAR	2	NF-B Activation	0.95	CHB	109460931-109605828	rs3827760	109513601	144897
G6PD	X	malaria resistance	0.21	YRI	153709606-153825233	rs1050828	153764217	115627
LCT	2	lactase persistence	0.56	FIN	136535946-136657220	rs4988235	136608646	121274
PSCA	8	cancers	0.77	YRI	143691875-143834142	rs2294008	143761931	142267
TRPV6	7	Calcium absorption	0.91	CEU	142518960-142633477	rs4987682	142569596	114517
ADH1B.OFF	4	-	0.71	CHB	105177528-105292572	rs7672705	105247146	115044
EDAR.OFF	2	-	0.95	CHB	114460931-114605828	rs34264290	114528334	144897
LCT.OFF	2	-	0.56	FIN	141535946-141657220	rs59101965	141583482	121274
TRPV6.OFF	7	-		CEU	147518960-147633477	.	147583629	114517

Table 2.3. Relative Error for different numbers of simulations and acceptance rates. In this table, we give the relative error of the mean and the false negative rate of the model choice for 1000 data sets randomly simulated under the SSV model with varying number of simulation n_S and proportion of accepted simulations δ. FN=False negative rate in model choice.

n_S	δ	RE(f_1)	RE(log(s))	RE(μ)	RE(log(t_1))	RE(log(t_0))	FN
10^5	10^{-3}	0.0392	0.348	2.92e-9	0.366	0.321	0.258
10^5	10^{-2}	0.0387	0.348	2.91e-9	0.365	0.322	0.240
10^6	10^{-4}	0.0389	0.343	2.95e-9	0.361	0.321	0.250
10^6	10^{-3}	0.0388	0.345	2.93e-9	0.364	0.320	0.236
10^7	10^{-5}	0.0386	0.343	2.86e-9	0.361	0.323	0.258
10^7	10^{-4}	0.0387	0.344	2.97e-9	0.362	0.321	0.246

2.6 Figures

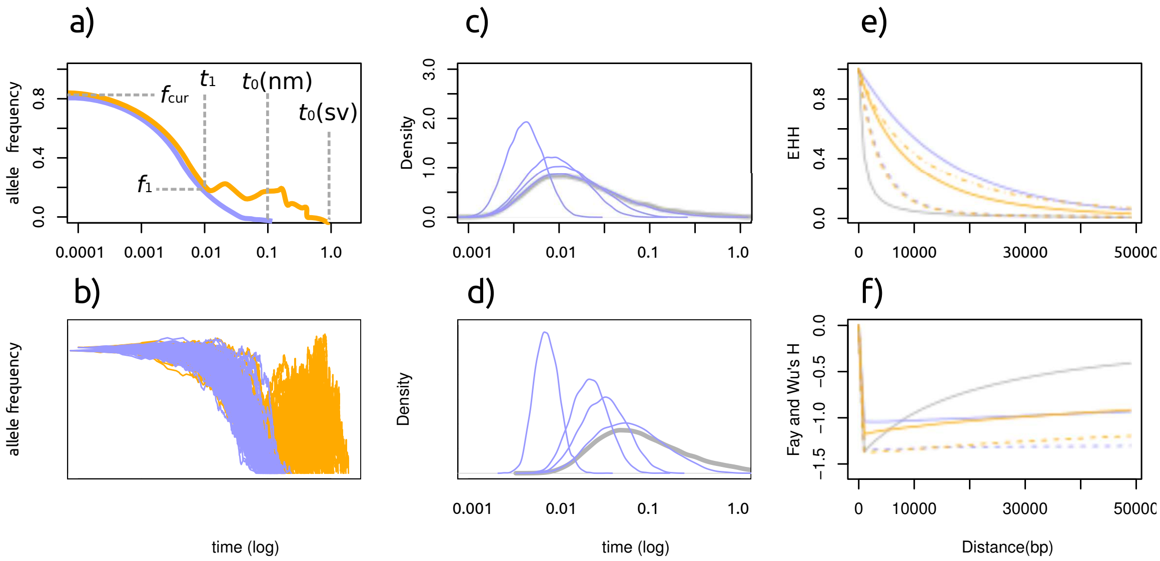

Figure 2.1. Characteristics of a selective sweep from standing variation. orange: sweep from standing variation blue: sweep from a new mutation, grey: neutral model. Panel a: A cartoon of the allele frequency trajectory with relevant parameters: f_1: allele frequency at the time selection started, f_{cur}: allele frequency at the time mutation is observed. t_1: time at which selection started. t_0: time when mutation arose. Panel b: 100 stochastic realizations of the allele frequency trajectory. Panels c,d: Age distribution of an allele at 1% frequency and 5% frequency in a population (log scale). Grey line denotes neutrality, blue lines represent selection with α=20,100,200 and 1000 (right to left). Panels e,f: Distribution of the EHH (e) and H (f) statistic under neutrality (grey), a *de novo* mutation (blue) and standing variation (orange). Full and dashed lines represent selective pressures of α = 1,000 and 200, respectively. The dash-dot line represents α = 4000. Note that the slopes of the curves are different for the two scenarios, and the low H value around 0 under neutrality is due to the conditioning on a high frequency derived allele. Times are given in coalescent units and are plotted on a logarithmic scale.

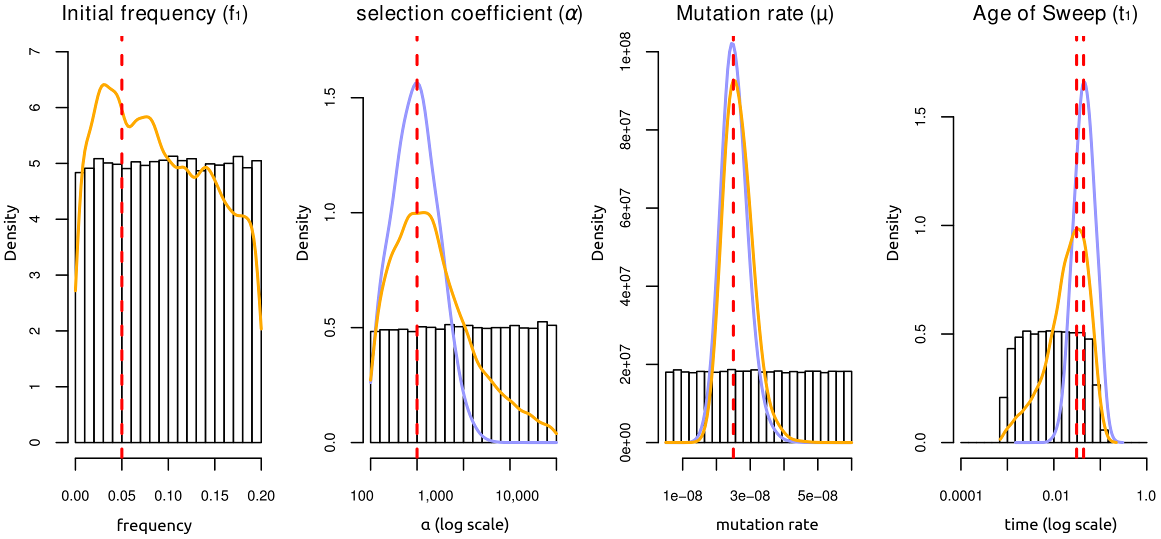

Figure 2.2. Parameter estimation accuracy under SSV and SDN model. Prior distributions are given as histograms; the orange and blue lines depict the average posterior distribution from 100 replicates of the parameters under the SSV and SDN model, respectively. The vertical dashed red line gives the parameters used for the simulation: $\alpha = 400$, $\mu = 2.5 \times 10^{-8}$, $f_1 = 0.05$, $\log(t_1) = -1.51$ (SSV)/-1.36(SDN). Estimates for the SSV model are less accurate for all parameters except μ, and 95% confidence intervals of estimates under the SSV model span the entire prior range for f_1, α and t_1. The age of the sweep is given in coalescence units.

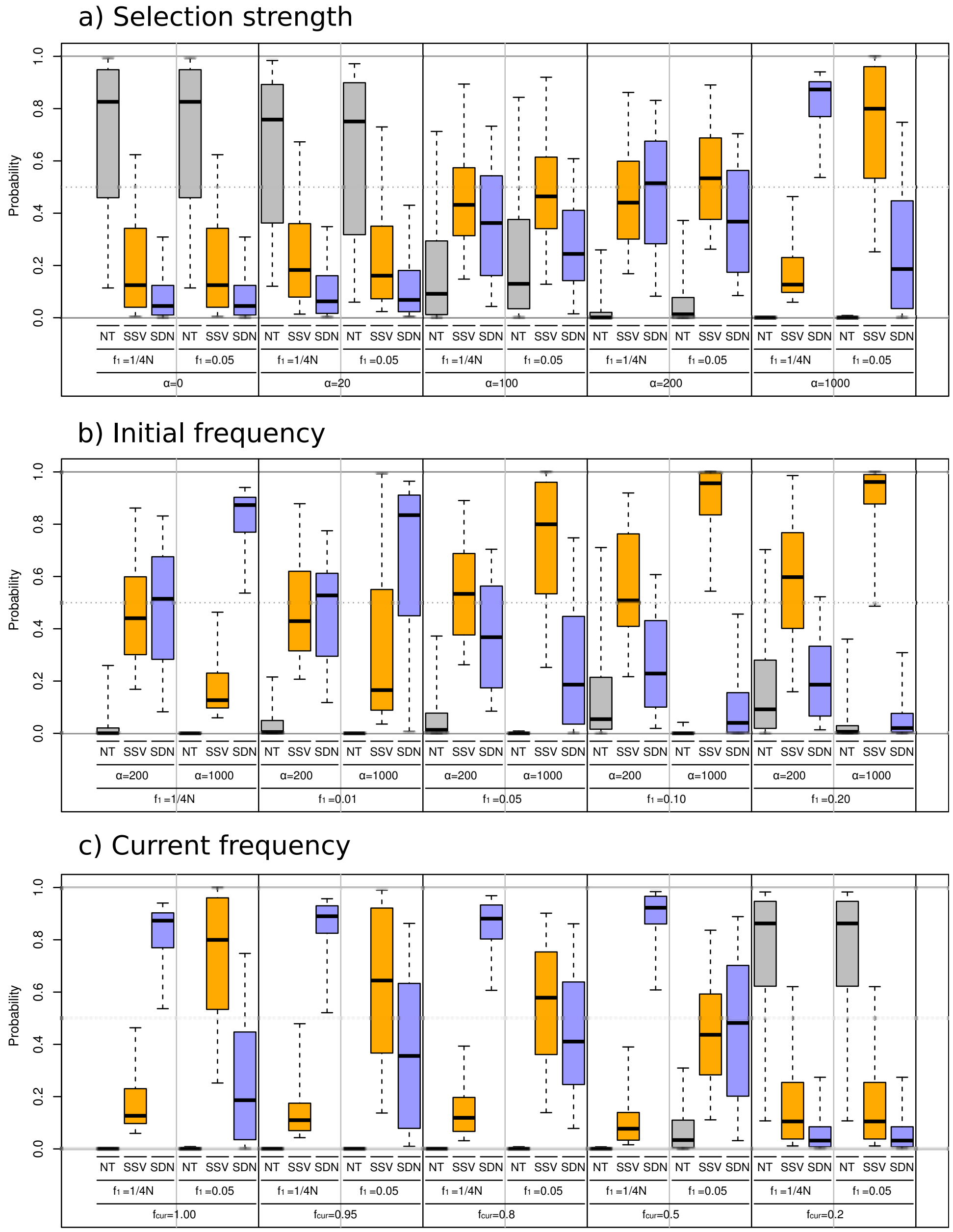

Figure 2.3. Simulation results for ABC model choice procedure. We simulated data using the fixed parameter values given in the lower part of the figure. The boxplots show the lower and upper quartiles, the median and the limits of a 95% interval of the posterior probability for the NT (grey), SSV (orange) and SDN (blue) models, respectively. Panel a: We compare the effect of the increasing selection coefficient α. Panel b: The effect of increasing initial frequency f_1. Panel c: The effect of the current frequency f_{cur}, In panels a,b f_{cur} was set to 0.95, and in panel c, α=1,000.

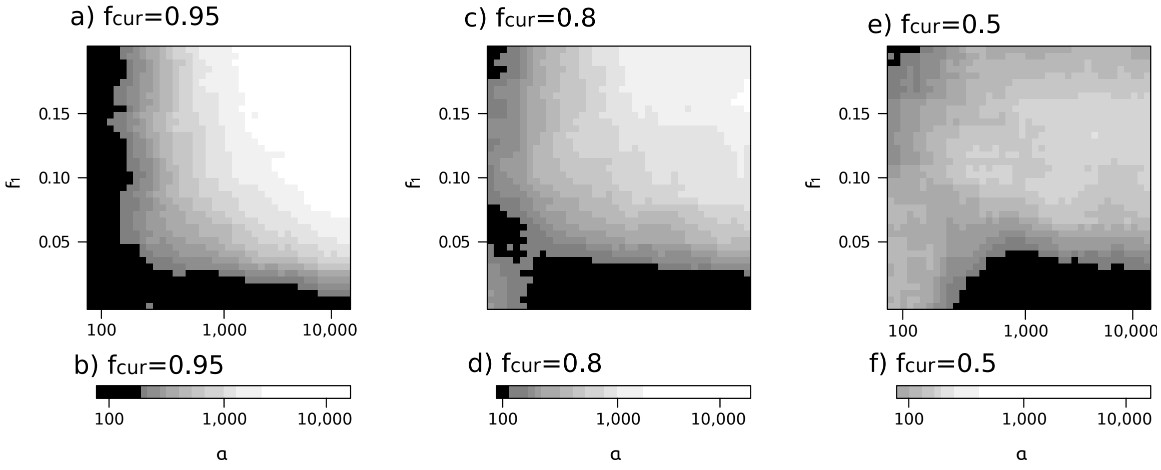

Figure 2.4. Parameter regions where distinction between models is possible. On the x and y axes are the prior ranges for selection coefficient and initial frequency of a selective sweep, respectively. Panels a, c and e give simulations under the SSV model, panels b, d and f for the SDN model. The different panels represent different current frequencies: In Panels a, b f_{cur} is 0.95, in c, d $f_{cur} = 0.8$ and in panels e and f $f_{cur} = 0.5$. Color gives the proportion of simulated data sets that were assigned to the correct model, when compared to the two alternative models. Black areas correspond to regions where this proportion is less than 50%, white areas to parameter regions where 95% or more of the data sets are correctly assigned. Each shade of grey corresponds to a 5% increase in the number of correctly assigned data sets.

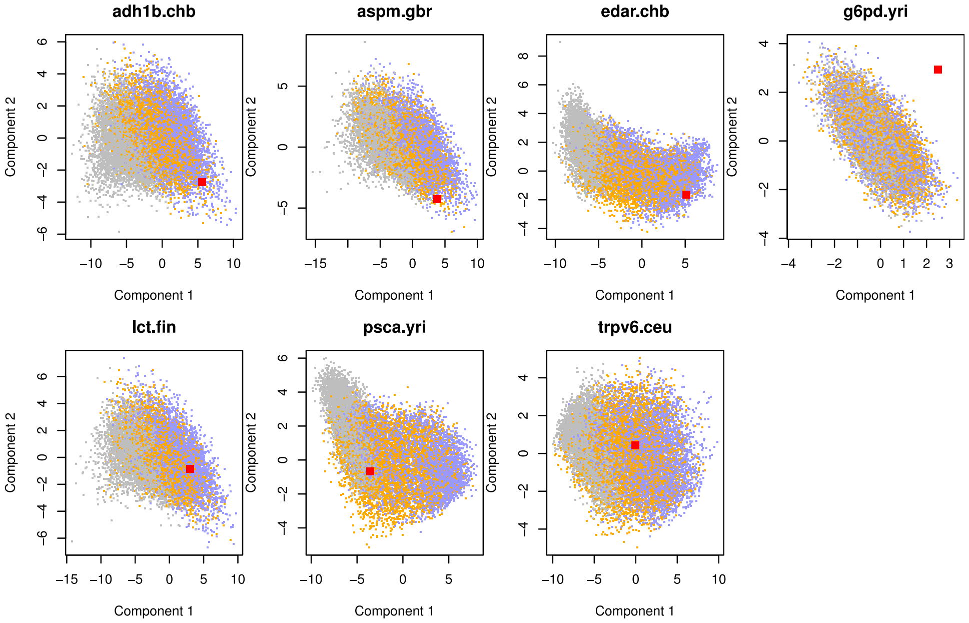

Figure 2.5. Distribution of summary statistics of 7 genes.
This figure shows the observed (red) and prior predictive distribution of the first two PLS-DA components. Neutral simulations are shown in grey, SSV in orange and SDN in blue. For G6PD we show components 2 and 3 to highlight the finding that none of the three models analyzed is able to model the data for this gene.

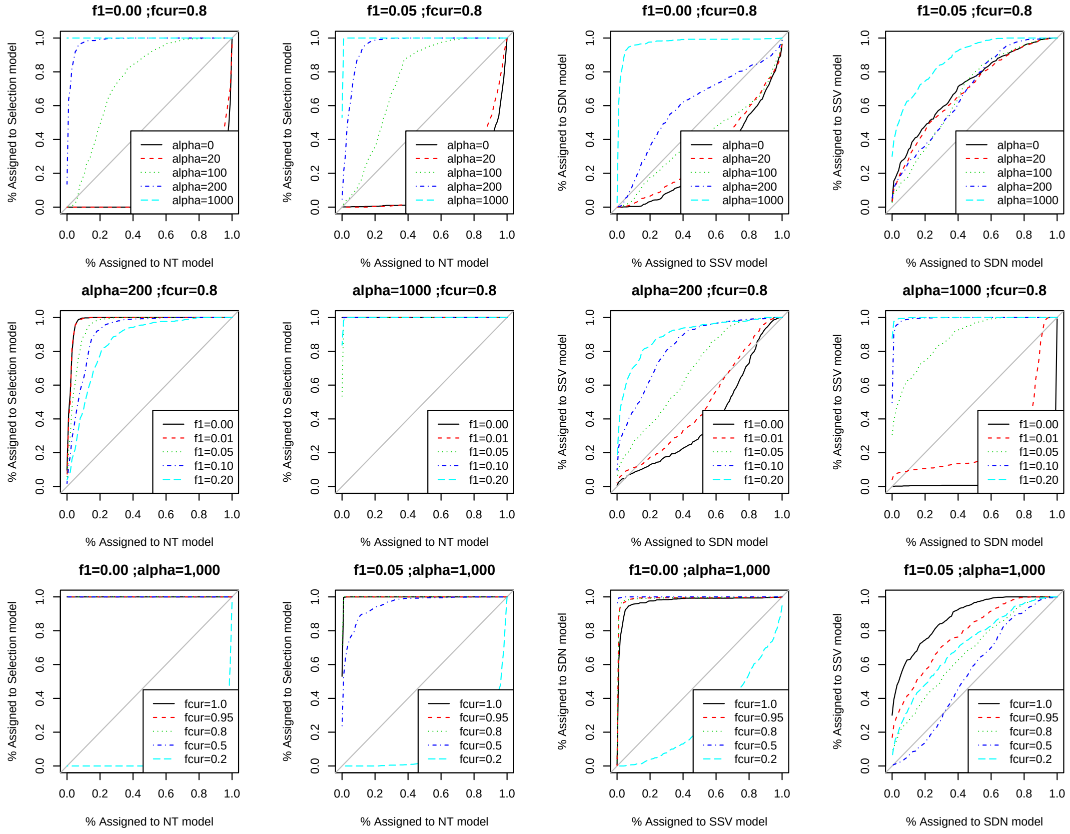

Figure 2.6. ROC plots.

This figure gives ROC plots for the same data as in Figure 3. As we have three models, the first two columns compare both selection models with a neutral model, and the last two columns compare the two selection model, with the model better characterized as the null model plotted on the x-axis. The lines give the percentage of simulation assigned to the model on the y-axis (sensitivity), given a proportion of models assigned to the x-axis (specificity). Parameters used for the simulations are given above the plot and in the legend box.

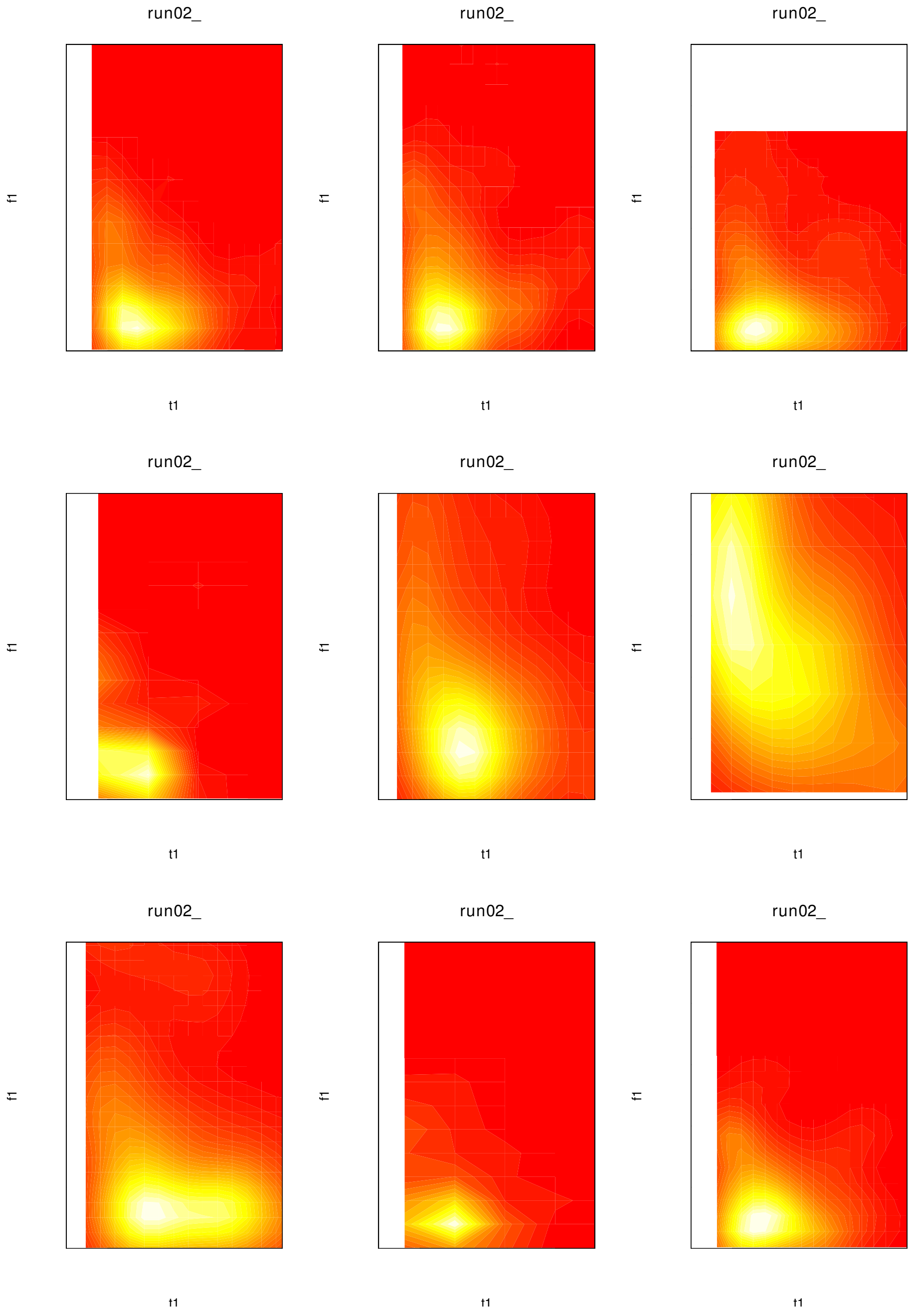

Figure 2.7. Joint posteriors of f_1 and t_1 of simulations under SDN model. Inferred joint posterior distribution of nine replicate simulation with parameters of $\alpha = 400$, $\mu = 2.5 \times 10^{-8}$ are shown. Red and blank areas correspond to areas with zero probability, yellow areas indicate high probability densities. Notice that for most simulation the inferred initial frequency is below 2%.

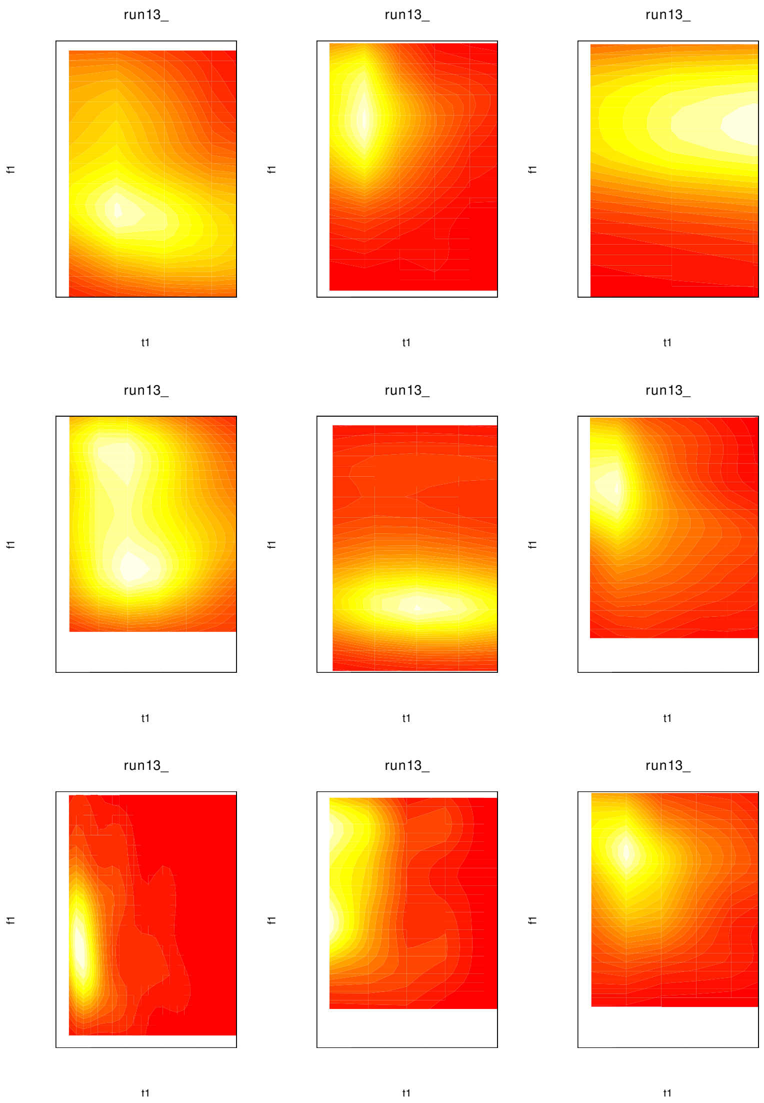

Figure 2.8. Joint posteriors of f_1 and t_1 of simulations under SSV model Inferred joint posterior distribution of nine replicate simulation with parameters $f_1 = 0.1$, of $\alpha = 400$, $\mu = 2.5 \times 10^{-8}$ are shown. Red and blank areas correspond to areas with zero probability, yellow areas indicate high probability densities. Notice that for most simulation the inferred initial frequency is above 5%, but the inferred probability of f_1 is often very inaccurate.

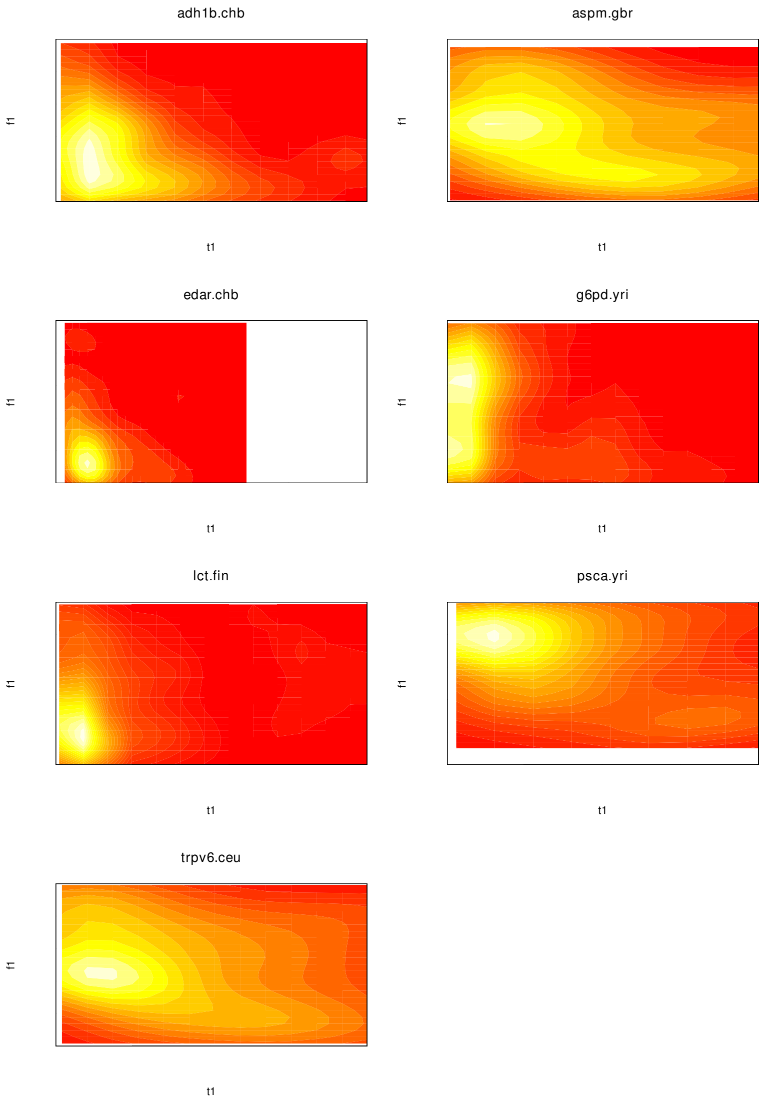

Figure 2.9. Joint posteriors of f_1 and t_1 for analyzed genes.
Inferred joint posterior distribution of all seven genes analyzed in this paper. Red and blank areas correspond to areas with zero probability; yellow areas indicate high probability densities.

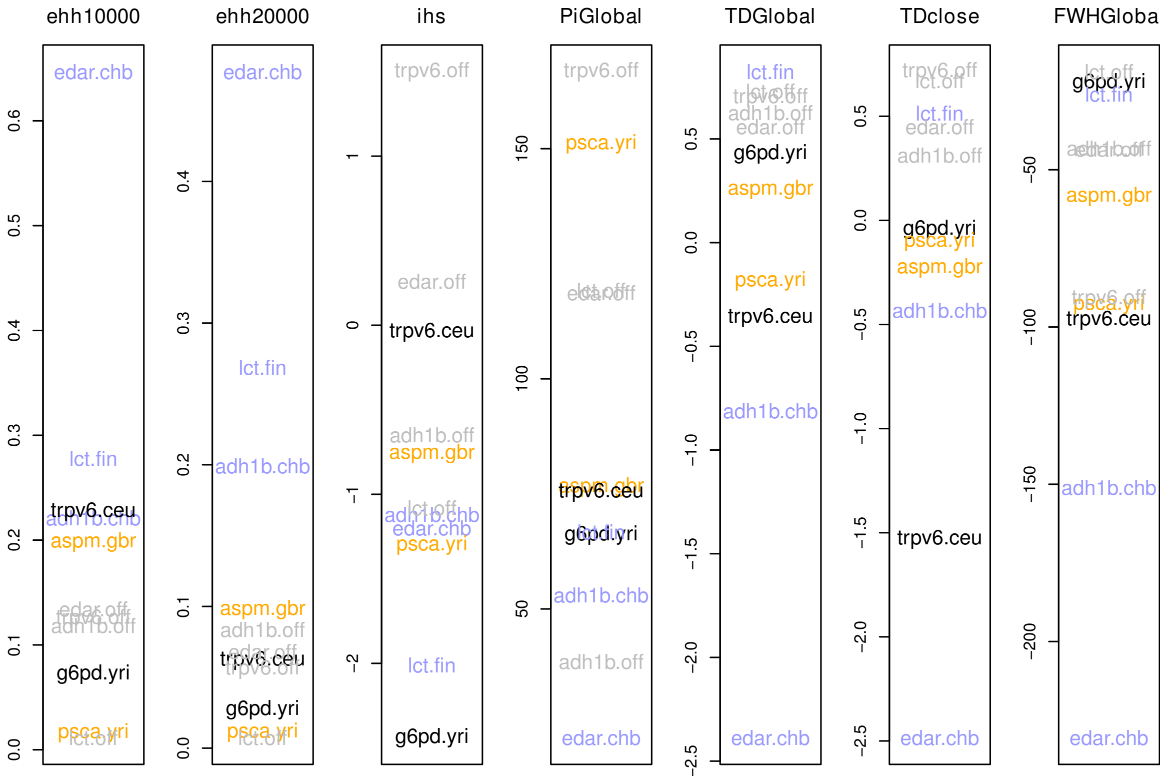

Figure 2.10. Observed summary statistic distributions.
We show the observed untransformed summary statistics for all genes and genomic regions we analyzed in this study (see Table 2.3). Colors indicate the most likely mode of evolution: neutral evolution (grey), SDN (blue), SSV (orange) and undetermined (black). TD=Tajima's D, FWH=Fay and Wu's H. The suffix "global" indicates that the statistic was calculated for the entire gene, the suffix "close" indicates the statistic calculated on a 20kb window around the selected site.

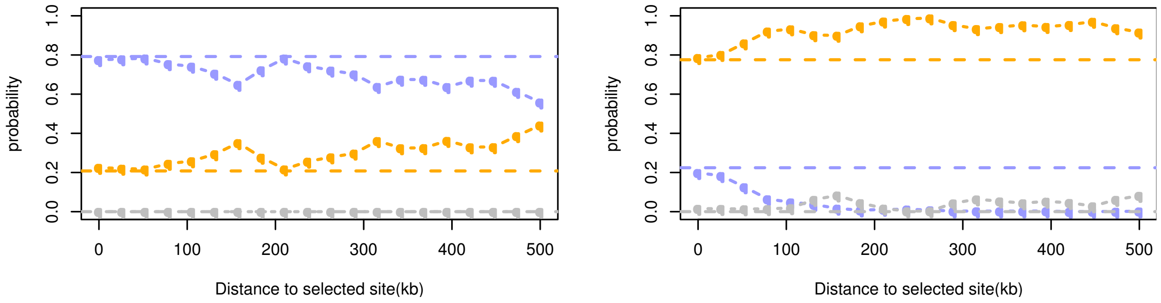

Figure 2.11. Effect of misidentification of selected site.
We show the posterior probabilities for SSV (orange), SDN (blue) and NT (grey) for simulations done from a *de novo* mutation(left panel) and standing variation (right panel), if we misidentify the selected allele. Simulations were done with selection strength $\alpha = 1,000$, sample size $n = 100$, mutation rate $\mu = 2.5 \times 10^{-8}$ and recombination rate $\rho = 3 \times 10^{-8}$. For the SSV simulation, f_1 was set to 0.1 X-axes give the distance between the "true" selected allele from the site for which the summary statistics were calculated. If the distance is larger than 50kb, we find a bias towards inferring SSV.

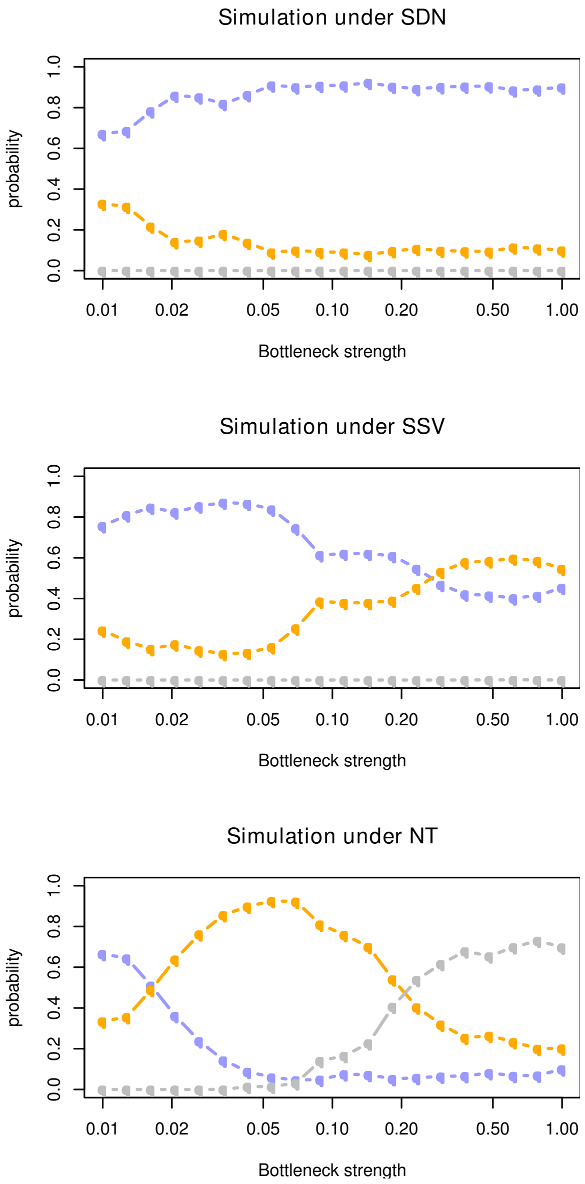

Figure 2.12. Bias in model choice due to a population bottleneck. We show the inferred posterior probabilities for SSV (orange), SDN (blue) and NT (grey) under a constant size model for simulations done under a bottleneck model. The bottleneck started 400 generations ago and lasted for 2,000 generations, which might be similar to the human out-of-Africa bottleneck Simulations were done with selection strength α=1,000, sample size $n = 100$, mutation rate $\mu = 2.5 \times 10^{-8}$ and recombination rate $\rho = 3 \times 10^{-8}$. For the SSV simulation, f_1 was set to 0.1. X-axes give the strength of the bottleneck as a proportion of the current effective population size. Unaccounted demographic history results in a bias towards estimates of stronger selection.

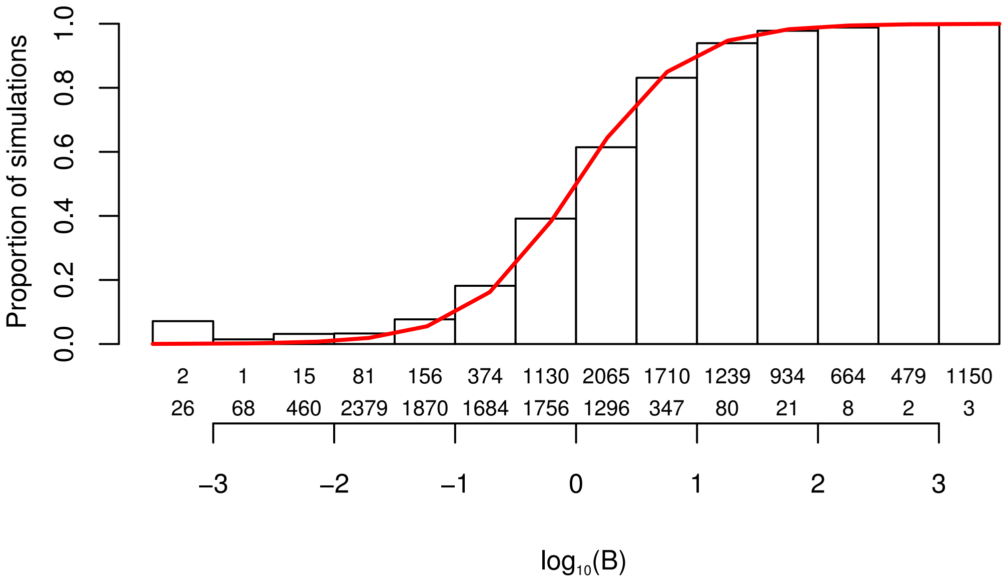

Figure 2.13. Model choice bias.
B denotes the Bayes factor in favor of the SSV model, $B = \mathbb{P}(SSV)/\mathbb{P}(SDN)$. We simulated 10,000 data sets under both the SSV and SDN model, and performed our model choice procedure on each data set, and divided the distribution into discrete bins. The figure gives the observed (bars) and expected (red line) proportion of simulations from the SSV in each bin. As can be seen, there is a slight excess of simulations from the SSV on the lower end of the graph. The leftmost bin contains only 28 simulations, two of which were simulated under the SSV model. Both of these simulations had a f_1 below 0.005, corresponding to a parameter region where the SSV and SDN models are very similar. The first and second row of numbers below the figure denote the number of simulations simulated under the SSV and SDN model, respectively.

Chapter 3

Detecting Range Expansions from Genetic Data

3.1 Introduction

Range expansions are ubiquitous in natural populations, and they are responsible for numerous biological phenomena. Range expansions result in a series of founder events that cause the newly founded populations to differ genetically from the source population. Some well-known examples are biological invasions (Handley et al., 2011), the post-ice age patterns of migration in several European taxa (François et al., 2008; Hewitt, 1999; Schmitt, 2007), and the colonization of Eurasia, North and South America by modern huamans (Cavalli-Sforza et al., 1994; Ramachandran et al., 2005; Tishkoff et al., 2009). Some of the descendants of an ancestral source population may remain near the location of that ancestral population. For example, the European population of the brown bear *Ursus arctos* most likely survived the last ice age in refugia in Spain and Greece. Brown bears followed the receding glaciers to colonize most of Europe, but populations at the locations of the former refugia persisted until the populations were driven to the verge of extinction by humans in the 20th century (Taberlet et al., 1998). Humans provide another example; derived populations are found all over the world, but there are also descendants of the first humans still living in Africa.

Sometimes, the routes of migration are known from direct observations, historical records and archaeological evidence. Frequently, however, the exact history of a species is unknown, and we want to use population genetic methods to gain more information. In this paper, we use genetic data to address two related problems: detecting whether a range expansion has occurred and inferring the geographic origin of a range expansion.

Characterizing the influence of geographic structure on genetic diversity has been one of the major goals of population genetics theory, with important contributions from Wright (1943), Malécot (1950), Kimura (1964) and many others. While there are many statistics designed to infer differentiation between populations (Balakrishnan and Sanghvi, 1968; Goldstein et al., 1995; Nei, 1972; Reynolds et al., 1983), the most widely used statistic to detect differentiation between populations is the fixation index F_{ST}, which traces to Wright (1949). A variety of estimators of F_{ST} have been developed (e.g. Reynolds et al., 1983; Weir and Cockerham, 1984). Roughly speaking, F_{ST} measures how much diversity exists between subpopulations compared to the diversity in the entire population; an F_{ST} value of 0 indicates that the two subpopulations are indistinguishable, whereas a value of 1 indicates that two populations are maximally differentiated. F_{ST} has been directly linked to the migration rate in several models, including the finite island (Slatkin and Voelm, 1991) and stepping-stone models (Cox and Durrett, 2002). Although F_{ST} can be used to estimate the amount of gene flow between equilibrium populations, it cannot be used to infer directionality of gene flow.

Two other methods that are widely used to detect geographic patterns are clustering algorithms and ordination methods. Clustering algorithms (Corander et al., 2004; François et al., 2008, 2010; Pritchard et al., 2000) such as STRUCTURE (Pritchard et al., 2000) classify individuals into discrete groups, which can then be used for further analysis. Ordination techniques (Cavalli-Sforza and Edwards, 1967), such as principal components analysis and multidimensional scaling, summarize data by indicating the overall similarity of populations. For example, principal component analysis has shown that genetic diversity is correlated with the geographic distribution of humans on a continental (Novembre et al., 2008) and global (Cavalli-Sforza et al., 1996; Wang et al., 2012) scale.

It is also possible to use likelihood methods to infer past features of population history. For example, the program IM (Hey, 2010) estimates the time of separation of populations and migration rates between them using data from multiple unlinked loci, and the program dadi (Gutenkunst et al., 2009) estimates past rates of population growth from the joint allele frequency spectrum of two or three populations. Both of these programs are computationally intensive and analysis for more than a few populations is infeasible.

Most statistics applied to subdivided populations do not provide information about asymmetries. F_{ST} and most genetic distances are defined in such a way that they are commutative (i.e. F_{ST} between populations A and B is the same as F_{ST} between B and A), and hence the value depends only on the amount of migration, not whether migrants moved mostly from A to B or from B to A. Clustering algorithms can produce groupings of populations that can be interpreted as describing an expansion, but expansion-specific information is lost in the process and the results of clustering analysis is often sensitive to tuning parameters such as the number of clusters. For principal components analysis, the view that the first principal component axis follows the direction of expansion (Menozzi et al., 1978) has recently been challenged (DeGiorgio and Rosenberg, 2012; François et al., 2010; Novembre et al., 2008), and it has recently been shown that, depending on the parameter values and the locations of the populations sampled, the first principal component axis may be parallel to or orthogonal to the axis of expansion, or at an angle in between.

Population genetics theory has shown that a range expansion can be detected from the characteristic reduction in genetic diversity with increasing distance from the origin of the expansion (Austerlitz et al., 1997; DeGiorgio et al., 2009; Edmonds et al., 2004;

Hallatschek et al., 2007; Ramachandran et al., 2005; Slatkin and Excoffier, 2012). The reason is that the succession of founder events during the expansion causes the progressive loss of genetic variants. In extreme cases, this can lead to relatively rapid fixation of neutral or even deleterious alleles, a process called allele surfing (Hallatschek et al., 2007; Klopfstein et al., 2006). The prediction of decreasing diversity has been confirmed by comparing the numbers of mtDNA haplotypes found in Southern European refugia and in central Europe (Taberlet et al., 1998). The same pattern can also been seen in humans where both a reduction in heterozygosity and an increase in linkage disequilibrium with increasing distance from the presumed origin of the expansion in Africa can be seen (Ramachandran et al., 2005).

In addition to creating a gradient in genetic diversity, range expansions tend to create clines in the frequencies of neutral alleles, with the frequency increasing on average in the direction of the expansion (Slatkin and Excoffier, 2012). An intuitive reason for this pattern is that each founder event results in additional genetic drift, and populations further away from the origin of expansion will therefore have experienced more drift. This can be seen from the following argument: The expected frequency of a neutral allele in the newly founded population is the same as in the source population. But some alleles will have zero frequency in the new population. Therefore, the average frequency of alleles in the newly founded population, given that they have non-zero frequency is expected to be higher than in the source population, thus creating the cline. This observation provides the foundation for our method of detecting range expansions.

In this paper, we introduce a statistic, the directionality index ψ, defined for pairs of populations. ψ is sensitive to patterns created by range expansions because it detects the allele frequency clines created by successive founder events. We show, using simulations, that the expectation of ψ is zero in an equilibrium isolation-by-distance model, and that its expectation is positive in the direction of the expansion. We also show that, using multiple samples, ψ can be used to infer the origin of a range expansion and the locations of barriers to expansion. We explore the power and robustness of our methods and finally apply it to human genetic data.

3.2 Results

In this section, we define the directionality index, give an intuitive explanation and discuss some of its properties. We will show that the directionality index is sensitive to recent range expansions in a one or two dimensional stepping-stone model, and then explore some more advanced applications.

3.2.1 Definition of the Directionality Index

Consider two samples of size $n, n \geq 2$ taken from two subpopulations S_1, S_2. Each sample consists of L biallelic markers (e.g. SNPs) that are shared between S_1 and S_2. The directionality index is defined as

$$\psi(S_1, S_2) = \frac{1}{n}\left(\bar{f}^{(S_1)} - \bar{f}^{(S_2)}\right) = \frac{1}{nL}\sum_{l=1}^{L}\left(f_l^{(S_1)} - f_l^{(S_2)}\right), \tag{3.1a}$$

where $\bar{f}^{(S)}$ is the average allele frequency of all derived alleles in population S, and $f_l^{(S)}$ is the number of copies of the derived allele at locus l in the sample from population S. It is important that the average and the sum are over only those alleles that are present in both populations; sites where either population is fixed for the ancestral copy are excluded. Equivalently, ψ can also be defined in terms of the two-dimensional site frequency spectrum (2D-SFS):

$$\psi = \sum_{i=1}^{n}\sum_{j=1}^{n}(i - j)f_{ij}. \tag{3.1b}$$

where f_{ij} denotes the proportion of SNP in the sample that are at frequency i in S_1 and at frequency j in S_2, and the SFS is normalized such that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} = 1.$$

This normalization is unusual in that alleles private to either of the two populations are excluded. In the special case where we compare two diploid genomes, $n = 2$ and equation 3.1b reduces to

$$\psi = f_{21} - f_{12}. \tag{3.1c}$$

These three definitions are equivalent and represent different interpretations of the directionality index. To aid intuition, we discuss them briefly. Equation 3.1a corresponds to the model we introduced in the introduction; we compare the average allele frequencies in the two populations. Because the population further away from the expansion origin is expected to have experienced more genetic drift, its alleles are expected to be at a higher frequency on average. Thus a positive ψ indicates that $\bar{f}^{(S_1)} > \bar{f}^{(S_2)}$ and that S_1 is further away from the origin of the expansion than S_2. If both populations have experienced similar amounts of genetic drift, then the average frequencies of shared alleles will be equal, $\psi \approx 0$ and we will not detect an expansion. Equation 3.1b is based on the SFS, and we see that ψ will be positive if f_{ij} is usually greater than f_{ji}. Thus, we are comparing the SFS entries that are reflected along the $x = y$ diagonal, and the directionality index measures the "skew" in the 2D-SFS. If there are more SNP that fall in the upper left triangle of the SFS (where $j > i$), ψ will be negative, and we infer an expansion from S_1 to S_2. The opposite conclusion will be drawn if there is an excess of SNP in the lower right triangle, and if the SNP are distributed symmetrically around the $x = y$ diagonal, ψ will be zero. Much of the paper will be focused on the case where each population is represented by a single genome, a case that will be particularly common in genomic studies. In that case, equation 3.1b reduces to 3.1c and we are simply comparing the abundance of SNP fixed for the derived allele in sample S_1 and heterozygous in S_2 to the number heterozygous in S_1 and fixed in S_2. If either number is significantly larger than the other, we infer expansion in the direction of the population with the larger number. It is also worth comparing the computation times. Equation 3.1a scales proportionally to the number of loci in the sample, whereas equation 3.1b scales with the square of the sample size if the site frequency spectrum has been previously calculated. As this is often required for calculation of other statistics such as F_{ST}, equation (3.1b) should be used for data sets where $L \gg n^2$. It is important to note that we have to assume that the sample sizes are the same in the two populations we are comparing. The reason is that the probability that an allele is absent in a sample is a complicated function of the sample size and the expected site frequency spectrum, and cannot be easily computed. When there are samples of unequal size we downsample the larger sample to the size of the smaller sample.

3.2.1.1 Determining Whether a Range Expansion Occurred

We first test the power of ψ to distinguish pairs of populations sampled from a recent range expansion to pairs of populations sampled under isolation-by-distance at equilibrium in a 1D-model. Figure 3.1 shows that F_{ST} increases at approximately the same rate under an equilibrium stepping-stone model with only isolation-by-distance (Panel A) and a model with a range expansion (Panel B), indicating that the two scenarios are comparable. We see that ψ is nearly zero in the isolation-by-distance model, regardless of the distance between the samples. In contrast, ψ increases with distance under the expansion model. Interestingly, ψ increases almost linearly with the distance between the origin and the population sampled, a fact we exploit later to infer the origin of the expansion. We also plotted the heterozygosity, a statistic that is also expected to be constant under an equilibrium model (Durrett, 2008) and decreasing under an expansion (Austerlitz et al., 1997; Ramachandran et al., 2005). However, our simulations show that heterozygosity is larger in the center of the habitat than near the boundaries because of the boundary effects. This is in contrast to most theoretical results (Durrett, 2008), which either assume either a circular model or an infinitely long stepping stone model, and where the heterozygosity is independent of the deme sampled. This observed gradient in heterozygosity has been observed previously and has been explained by longer coalescence times for a sample taken close to the boundary (Wilkins and Wakeley, 2002). It is also worth noting that this effect is much weaker in a two dimensional population.

Similar results for F_{ST} and ψ were obtained in 2D (Figure 3.2). F_{ST} is slightly larger in the case of a range expansion than in the isolation-by-distance model (Panels A and C), but qualitatively we see an increase of F_{ST} with distance under either model. The pattern for ψ, however, is again different (Panels B and D): under the isolation-by-distance model, $\psi < 0.01$ for almost all comparisons, with the exception of a few demes that are at the boundary of the simulated region. In contrast, the magnitude and sign of ψ nicely illustrate the effect of the range expansion. ψ is zero only for demes that are very close to each other or pairs of demes equally far away from the expansion origin. The latter is due to symmetry: two samples that are an equal distance apart from the origin will have a symmetric SFS, resulting in a ψ close to zero.

In Figure 3.3 we show the effect of the most important parameters on our ability to reject the null hypothesis of equilibrium isolation-by-distance for pairs of samples of size two. For all parameters, we find that using the directionality index results in higher power than comparing differences in heterozygosity, while false-positive rates are low and roughly the same for the two methods. We find that we have comparatively little power to reject the null hypothesis if the two sampled individuals are close to each other(Panel 3.3A). This is expected, since there are fewer founder events separating the two individuals. Therefore we expect ψ to be lower for nearby populations, as shown in Figures 3.1 and 3.2. Panel B shows that a moderate number of shared SNP is necessary, i.e. more than one thousand, to get high power to reject equilibrium isolation-by-distance. In addition, we find that slow expansions are harder to detect than rapid expansions, and more recent expansions are easier to detect than expansions that happened a longer time in the past (Panels C and D). Neither of these findings are unexpected; after an expansion, genetic drift will affect the loci in both populations equally. The number of shared SNP that are due to the range expansion will decrease with time and be partially replaced by SNP that only experienced the equilibrium population structure and hence do not carry a signal of the expansion. Similarly, if the time between expansion events is large, the founder effects caused by the expansion will become less important relative to genetic drift that occurs between expansion events, weakening the signal of the expansion. For these slow expansions, the power of heterozygosity to detect an expansion decays much faster than the power of ψ. Finally, we note that the false positive rate, denoted in grey and pink in Figure 3.3, is independent of both the distance between loci and the number of SNP for both ψ and H.

3.2.2 Inferring the Origin of a Range Expansion

In addition to showing that a range expansion occurred, the results in Figures 1 and 2 suggest that spatial patterns in pairwise values of ψ can indicate the origin of an expansion if more than two populations are sampled. For this purpose, we employ a method commonly used by engineers in problems of localization and navigation (Gustafsson and Gunnarsson, 2003), called Time Difference of Arrival location estimation (TDOA). TDOA methods are used in remote sensing and to locate cell phones (Gustafsson and Gunnarsson, 2003). The key assumption of the TDOA algorithm is that the magnitude of a pairwise statistic between two sample locations i and j is proportional to the difference in distance from i to the origin and the distance from j to the origin. If i is very close to the origin and j far away, the TDOA statistic will be large, but if i and j are at the roughly the same distance from the origin, then the TDOA statistic will be close to zero.

In engineering applications the TDOA statistic is the time difference between the arrival of a signal emitted from different sources (hence the name). In our application, ψ takes on the role of the time difference with the implicit assumption that the magnitude is proportional to the difference in distances of the two samples from the origin. To illustrate, we first consider the special case of $\psi_{ij} = 0$. Assuming that we have already rejected isolation-by-distance in favor of a range expansion, we know that i and j are equally far from the origin and the origin must therefore lie on the line perpendicular to the line through i and j. If we had three or more samples all at the same distance from the origin so that the pairwise ψ values are all zero, we could infer the origin was at the center of the circle passing through the three points.

In general, however, ψ_{ij} will be non-zero. In that case, we know from elementary geometry that the set of candidate points based on a one pair of samples is not a straight line, but a hyperbola with the sample locations as its foci. (see Figure 3.4). For samples from k locations, we calculate ψ for $k(k-1)/2$ pairs and hence obtain $k(k-1)/2$ hyperbolas. In a perfect, noiseless world, all hyperbolas would intersect in a single point, the origin of the expansion. In practice, genetic data is stochastic and we have to estimate the location of origin. To do this, we interpret each hyperbola as a non-linear equation with three unknowns, the sample coordinates x, y and the speed of expansion v. v is a nuisance parameter that describes how much the allele frequency increases per unit distance from the origin. For more than three samples the system is overdetermined and, rather than solving the system of equations explicitly, we use weighted non-linear least squares.

We first illustrate this approach on simulated data, where we sample a regular grid (Figure 3.5). We simulated a range expansion in a 101×101 stepping-stone model. In all simulations, we chose the coordinate system such that each deme corresponds to one unit of distance. The start of the expansion is in deme (25,35), indicated by the grey dotted lines in Figure 3.5. The direction of the arrows plotted in Figure 3.5 indicate the sign of the pairwise ψ-value, between adjacent samples on a grid, and the thickness of each arrow corresponds to the magnitude of ψ. A missing arrow denotes a non-significant ψ value. In Panel 3.5A we performed a simulation under an equilibrium isolation-by-distance model. We see that in this scenario, only 11 out of the 60 pairwise comparisons are significant; all of them point towards the corners and are due to the boundary effects of the simulations. The red ellipse is a 95% confidence ellipse of the inferred origin. Under the isolation-by-distance model, this is located in the center of the population, illustrating that the TDOA approach will yield an answer even if no expansion has occurred, so it is important to first test if an expansion has actually occurred. From Panels B-D we see that the expansion signal is clearly portrayed by the directionality indices and we get high confidence in the estimated origin. In fact, the confidence region is so narrow that the ellipse is barely visible in Panel B. The confidence region becomes larger when we reduce the number of samples. Furthermore, we see in Panels C and D that the origin is slightly biased towards the center of the population. This is again due to a boundary effect, and goes away if we take all samples at least 10 demes away from the boundary of the population.

To assess the properties of this method more systematically, we report the root mean squared error (RMSE) under several scenarios (Figure 3.6). The RMSE is the square of the Euclidean distance between the estimated origin and true origin. We also compare our method to the method of Ramachandran et al. (2005), who used a linear regression of the heterozygosity on the distance to a set of candidate origins. Their inferred origin of the expansion is the point with the highest associated regression coefficient, conditional on the slope of the regression curve being negative. Most data in Figure 3.6 was simulated with a fairly rapid expansion; the time between subsequent expansion events was set to 0.001 coalescence units, so that the complete expansion was completed in 0.13 coalescence units. This speed is roughly that of the out-of-Africa expansion of humans. For these parameters (Figure 3.6A-D) the two methods have similar performance, with only marginal improvements in how the methods perform with different amounts of data. We find that with adequate numbers of samples and data, the RMSE for both method is around four, with less than one distance unit of difference between the two methods. Overall, the ideal amount of data for this method lies around 20 diploid samples and 7,000 independent SNP. Having more data will not substantially improve performance. For the set of simulations with increasing numbers of SNP, we also tested the effects of sampling on a grid versus taking samples from random locations. The latter scenario is probably closer to real sampling schemes. Interestingly, we found only negligible differences, indicating that the sampling locations are only a minor issue unless the sampling locations are very skewed (for example if they all lie on a transect).

Changing the position of the origin has little effect on the RMSE for the first 30 distance units, indicating that the method is accurate if the origin is sufficiently far away from the boundary. If the origin is outside the region sampled, the performance is significantly worse. This has two causes: first, we would expect it to be easier to infer the origin if it lies in the middle of the sample, as compared to an origin that is far from all samples. This part also explains the difference between samples taken on a grid and random samples: In the grid, the corners are systematically sampled (since we force a grid sample to be there), whereas in many random samples there may be fewer samples on one side of the origin than on the other, resulting in a loss of accuracy. A second factor resulting in reduced accuracy are again boundary effects, which skew the effect of the expansion if samples happen to be close to the boundary.

We next focus our attention on the effect of varying the parameters of the expansion (Figure 3.6C-F): The number of founders (Figure 3.6D) has an almost linear effect on the estimation accuracy. Fewer founders imply a stronger founder effect and hence a stronger signal of expansion (Slatkin and Excoffier, 2012), which makes the origin easier to detect. We find the biggest difference in how our method performs in comparison to the Ramachandran method is when the expansion is slower, or when we want to detect an expansion that occurred a longer time in the past. Interestingly, our method has almost the same accuracy for different expansion speeds, whereas the Ramachandran method is less accurate if the expansion is slower. Also, we find that the heterozygosity gradient disappears soon after the expansion has finished (3.6F), whereas the ψ retains the signature of the range expansion for much longer.

3.2.3 Adding Environmental Complexity

The previous section assumes an idealized population in a homogeneous habitat. In practice, however, habitats are heterogeneous and barriers to gene flow and range expansion often exist. In the following sections, we show how our method performs in slightly more complex scenarios. First, we allow demes with different population sizes. While we kept the mean size of demes the same, we followed Wegmann et al. (2006) in drawing deme sizes from a gamma distribution. Next, we include barriers to dispersal that affect both the initial expansion and gene flow following the expansion. We illustrate how we can use algorithms from graph theory to locate barriers. Finally, we model an expansion starting from multiple origins.

3.2.3.1 Heterogeneous Population Sizes

The effect of variance in deme size on demographic expansions was explored by Wegmann et al. (2006). They found that heterogeneous populations have a higher rate of population differentiation between demes, and predicted that detecting range expansion would be more difficult because of the increased noise. Our simulations confirmed this prediction but only if there is substantial variation in deme size (Figure 3.10). We found that heterogeneity in deme size has little effect if the variance in deme size is low, with RMSE only differing slightly from the case with equal deme sizes. A variance of 0.5 in deme size, for example, corresponds to a size difference of around two orders of magnitude. But the average RMSE for the location estimate only increased to 5.43, compared to 4.57 in a comparable model without variation in deme size. However, this value corresponds to some kind of “tipping point”: when we further increase the variance in deme size, some deme sizes will become effectively zero in size and this greatly reduces the accuracy of the estimated origin.

3.2.3.2 Barriers

We can use pairwise directionality indices to gain qualitative information about colonization paths, i.e. the corridors through which the population expanded. To do so, we interpret the matrix of pairwise values of ψ as the adjacency matrix of a graph. A positive ψ between populations S_1 and S_2 is interpreted as meaning “Population S_2 was colonized after population S_1” and can be visually represented by an arrow between S_1 and S_2. To improve the visual representation of the graph, we apply standard algorithms to remove some of the edges. In particular, we apply the transitive reduction algorithm (Aho et al., 1972) to find the graph with the fewest edges that retains the connectivity of the original graph. If ψ is positive between populations S_1 and S_2, but there is also an indirect path with $\psi > 0$ when comparing S_1 and S_3 and S_3 and S_2, we remove the direct connection from S_1 to S_2. This is justified by noting that colonization of S_2 through S_3 is more parsimonious than colonization of S_2 both through S_3 and directly from S_1. We obtained a further reduction by computing a maximum spanning tree (Korte and Vygen, 2008), which reduces the graph to $n-1$ edges, where n is the number of sample locations. The maximum spanning tree identifies major migration paths, and does not cross strong barriers to expansion and gene flow (Figure 3.7). Furthermore, we can obtain an ordering of all samples by simply summing all ψ values that sample is involved in:

$$\psi_i = \sum_{j \in \text{samples}} \psi_{ij}. \tag{3.2}$$

The smallest value of ψ_i indicates the sample taken closest to the origin, and the largest value of ψ_i indicates the sample furthest along the expansion. In Figure 3.7B we show that both the maximum spanning tree and the ordering are useful qualitative tools to identify barriers.

3.2.3.3 Multiple Origins

Range expansions may have more than one origin. A classical example is the colonization of Central Europe after the last glacial maximum. Species with Southern European refugia in the Balkan Penisula, Italy and the Iberian peninsula followed the receding glaciers and explain many biogeographical pattern we observe today (Schmitt, 2007). A straightforward way to apply our method to such expansions is to first estimate which populations were colonized predominantly from each potential origin, and then use only those populations to infer the location of each origin. There are several ways to assign sampled individuals to clusters corresponding to a each origin. In classical studies, often mtDNA haplotypes were used for this purpose (e.g. (Hewitt, 1999; Taberlet et al., 1998)), but programs such as STRUCTURE (Pritchard et al., 2000) or simple clustering based on the observed polymorphism frequencies may yield more accurate results. In our simulations, a simple K-means clustering algorithm was able to correctly identify the number of clusters in all cases, even when the two founder populations were drawn from the same original population. The resulting estimates of the locations of the origins are slightly less precise than with a single origin (Figure 3.8), but that is to be expected because there are fewer samples contributing to the location estimate for each origin. Also, the estimates were worse when the two origins were close together.

3.2.4 Application

3.2.4.1 Human Diversity

We applied our method to a data set from 55 human populations from the Human Genome Diversity Panel and HapMap III (Altshuler et al., 2010; Cann et al., 2002; Fumagalli et al., 2011). The results are given in Figure 3.9. We calculated ψ and its standard error for all pairs of populations and transformed this into a Z-score. As expected from a data set with several hundred thousand loci, the vast majority of comparisons were highly significant, with a median absolute Z-score of 28.1, and a mean absolute Z-score of 41.9 across all comparisons. Globally, we could detect four major clusters: i) Africans, ii) Europeans and Pakistani, iii) East Asians and iv) Native Americans. Here, a cluster is loosely defined as a group of sampled populations that all show the same signal when compared to other groups of populations. For example, all 450 comparisons made between African and Non-African populations showed evidence for expansion out of Africa, consistent with the out-of-Africa hypothesis. Similarly, with few exceptions all comparisons between Europeans and Native Americans showed that Europe was colonized before the Americas.

Within Africa, we found all comparisons to be significant, and all pairwise ψ values agreeing on a single origin of the expansion. The San people were the only population that had positive ψ values when compared to all other populations, indicating that they are closest to the origin. They are followed by the Biaka- and Mbuti-pygmies, which are have negative ψ values with the San. This is followed by the southern Bantu sample, and a cluster consisting of Yerubans, Luhya, Mandenka and Northern Bantu, each having a negative ψ with the others previously mentioned, and positive values for all other populations. The African populations furthest from the origin were the Maasai and Mosabite, the latter being very distinct from the sub-Saharan populations.

The closest outside Africa are the Bedouin and Palestinian populations, both from the Middle East. The third Middle Eastern population present in our data, the Druze, fall in a larger group containing almost all European, Pakistani and Indian populations. Within Europe, the three Italian population all have non-significant ψ scores with one another, but are found to be ancestral to the other European populations. They are followed by the French and French-Basque, which also cannot be distinguished, and the Orcadian, Adygei and Russians. In Pakistan, we find the Makrani to be the most ancestral population, followed by the Brahui and Balochi, Sindhi, Kalash and Burusho. It is noteworthy that this list corresponds to their distances from Africa, with the exceptions that the Brahui and Balochi are switched, and the Hazara are not in the main Pakistani cluster, but rather form a distinct group with the Uygur. Besides the Uygur, all other East Asian populations form a single large cluster with very little resolution. Clearly distinct from this cluster are the Papuans and Melanesians, which are similar with asymmetry between them($\psi = 0.0019$, $SE\psi = 9.2 \times 10^{-4}$, $Z = -2.05$). They are closer to the African populations than to the East Asian populations, but further away than the Pakistani and European populations.

Finally, Native American populations form a distinct cluster, which are strongly separated from all other populations. Within the Native American populations, we find evidence of a North to South colonization pattern with the Pima population being closest to the Eurasian populations, followed by the Maya and Colombians. The most distant populations are the South American Karitiana and Surui, which have a nonsignificant pairwise ψ between them.

We also tested our ability to infer the origin of humans using the TDOA approach. As continents most likely act as strong migration barriers, we did not use the TDOA approach on the entire HGDP data set. Instead, we applied our method to the data set of Henn et al. (2011) which contains 30 African populations. We estimate an origin of the Human expansion at 30° S 13° E, which lies in central South Africa, closest to the location of the San sample (28.5° S 21° E and 22° S 20° E).

3.3 Discussion

We introduce a new statistic, the directionality index ψ, and showed that ψ can be used to test for a range expansion and to characterize it. Although we have focused on range expansions, ψ is sensitive to other deviations from symmetric migration. While a range expansion might be a plausible explanation in many cases, alternative scenarios such as a source-sink population structure or a large differences in effective population sizes should also be considered. One of the main advantages of the directionality index is that the assumptions and limitations of the approach are easy to discern: the directionality index is zero if the 2D-SFS is roughly symmetric about the diagonal. This is certainly true under most equilibrium models considered in theoretical studies, such as island and stepping stone models, particularly as the boundary conditions in the latter are typically chosen such that the model is symmetric. The directionality index can be used to determine how appropriate these models are for a given data set. If ψ differs from zero then care should be taken in applying methods that are based on these theoretical models. On the other hand, if ψ is close to zero, we can interpret this as justification for using the powerful theoretical results for these models (Durrett, 2008).

In this regard, the directionality index can be seen as a “first step” analysis that can be computed very easily, is able to answer very broad questions about a data set and may act as a guide to what parametric models might be employed, e.g. Approximate Bayesian Computation (Beaumont et al., 2002; Wegmann et al., 2010) or dadi (Gutenkunst et al., 2009). We have also shown how we can introduce the physical location of the samples in our inference framework. In many cases, natural populations are well described by a continuous distribution (Guillot et al., 2009; Rosenberg et al., 2005), and as we show in the TDOA analysis, using a simple statistic together with the physical locations can result in a powerful method. Our approach is also different from most other methods dealing with spatial data in that it explicitly assumes a non-stationary population. In this paper, we link the ancestral demographic process of a range expansion to the observed patterns of genetic diversity. While the effect of the expansion on F_{ST} appears to be quite small, our ψ statistic can be used to distinguish between equilibrium and non-equilibrium models. Finally, we show how we can extend our method to deal with more realistic landscapes. Whereas the TDOA analysis is not robust to large barriers of gene flow, interpreting the pairwise ψ statistics as a graph can unmask important details of a species’ history.

3.3.1 Simulation Results

We find that ψ is well suited to distinguishing between isolation-by-distance and range expansion when demes are sufficiently far apart and the range expansion is recent and occurs at a fast rate. These restrictions are not surprising. Geographically close demes will be genetically more similar, regardless of their history, and historical processes should therefore be harder to distinguish. That a recent expansion is easier to detect than an older one is also easily explained by the eventual convergence to equilibrium isolation-by-distance pattern, and similarly, a rapid range expansion leaves less time for genetic drift to blur the patterns created by the range expansion. Lastly, increasing the amount of data will increase the power to distinguish asymmetric from symmetric processes as each SNP contributes only a little information about the history of dispersal. In all cases, our ψ statistic outperforms ΔH. From the analyses of the stepping-stone model we see one of the main differences between ψ and F_{ST}. In an isolation-by-distance model, as the distance between the sampled locations increases, F_{ST} will increase but ψ will remain small. Again, this makes sense intuitively. The number of shared genetic variants decreases with distance, and hence F_{ST} increases. However, this reduction in shared polymorphisms is symmetric, and hence will have no effect on ψ. The pattern is different in the model of a population expansion: when comparing with a sample from the origin of expansion, both F_{ST} and ψ increase with distance. The signal diminishes, when migration rates are high, however. This is apparent from Panel D in Figure 3.1, where ψ is zero for the first ten demes. Here, migration had enough time to undo the effect of the range expansion in the demes that are furthest away from the origin.

We find that we can get surprising estimates of the location of the origin of an expansion from relatively small datasets. 20 samples with around 10,000 SNP yield accurate estimates. This result indicates that our method is not applicable to mtDNA or microsatellite data, but it should be applicable to transcriptome data, which can be assembled for many non-model organisms. It is also worth noting that the error does not go to zero even with larger amounts of data. There are several reasons for this. The linearity assumption we made for the TDOA approach is not completely accurate. ψ does not increase perfectly linearly with distance especially near boundaries. A more subtle reason is the algorithm we use; although least-squares is easy to implement and yields good results, other optimization algorithms might reduce the RMSE. A third reason is the intrinsic stochasticity of genetic processes. We demonstrated how our method can be adapted to incorporate more complex models. We showed that small differences in deme size have little effect on our ability to estimate the location of origin. If however, the habitat is very heterogeneous our method becomes less accurate. This implies that when analyzing species that live in very patchy habitats, , the TDOA method should not been applied, because the assumption that ψ is proportional to physical distance is violated. In that case, while it is not possible to infer an origin that is distinct from the samples, it is nevertheless possible to find the sample that is closest to the origin, which in many cases might suffice. Also, we have shown that we can apply graphical algorithms to get an accurate representation of the expansion pattern.

3.3.2 Human Genetic Diversity

When analyzing the human data set, we found that i) ψ scores are correlated with distance and ii) if population i is closer to Africa than population j, then $\psi(i, j)$ is in most cases negative, a pattern that is expected under a model of expansion from Africa. As explained previously, the directionality index depends not only on the two population compared but also on the history of the other populations. We find the South African San people to be the population closest to the origin of humans both using the TDOA method and when interpreting all pairwise directionality indices. This supports the interpretation that the origin of modern humans is somewhere in Southern Africa (Henn et al., 2011; Tishkoff et al., 2009). Another interesting result is that the Melanesian and Papuan samples, while very similar, show positive ψ values when compared to other East Asian populations, but the directionality index is negative when compared to the Pakistani, European and African populations. This is consistent with a "two-wave" model of colonization of South-East Asia, with a first wave consisting of present-day Papuans and Melanesians, and a second wave consisting of the present day Chinese populations(Rasmussen et al., 2011). Our results are also in agreement to the results obtained by Hofer et al. (2009), who analyzed the HGDP data set and found that neutral processes might be an explanation for large differences in allele frequency between human population groups. Our results support their findings, and extend them by giving an explanation on how the increase in derived allele frequencies might have arisen.

3.4 Methods

3.4.1 Simulations

We implemented a simulator that performs continuous time coalescent simulations on a discrete stepping stone model (Kimura, 1964; Malécot, 1950) of finite size. We assumed that the backward migration rates were equal between all pairs of adjacent demes and that the boundaries were reflecting. We used a modified version of the expansion model of (Slatkin and Excoffier, 2012), where an expansion is modeled with a one-generation bottleneck of reduced size. In our backward-in-time framework, this corresponds to moving all lineages present in a deme being colonized to a randomly chosen neighboring deme. We introduce a founder effect by adding additional coalescence events according to the appropriate backward Wright-Fisher transition probability (Wakeley, 2009, p. 62). Unless noted otherwise, all expansions were done with a founder size of 200. Once the final deme is reached, an regular island model coalescent is run where each island corresponds to a founder population (in most simulation, the number of islands is one).

Throughout this paper, we simulated unlinked SNP using an importance sampling scheme. After generating 1,000 gene trees, we calculate the appropriate multi-dimensional site frequency spectrum, where each sampled population corresponds to a dimension. We can then draw SNP with replacement from this site frequency spectrum.

The parameters used for the majority of the power simulations are as follows: We simulated on a 101×101 stepping stone model, with deme coordinates starting at (-50,50) at the lower left corner and (50,50) in the upper right corner. Each deme exchanges migrants to the neighboring demes to the north, south, east and west at scaled migration rate of $M = 2Nm = 1$. For the power simulation, we sampled a single diploid individual each from two colonies at (-25,-25) and (-25,25). For the TDOA simulations we simulated one individual each from a deme on a quadratic grid between (-30,-30) and (30,30), with 36 samples in total. This corresponds to a distance of 12 demes between any two sampled demes. We usually generated 1,000 independent coalescent trees and then used importance sampling to generate 100,000 SNP from the population, conditioning on them being shared between at least two of the samples. In the case of a range expansion, the standard point of origin was set to (-15,-25) and the expansion occurred at a rate of one expansion event every 0.001 coalescence units, with the expansion being observed 0.13 coalescent units after it started, where coalescent units are measured on the time scale of a local deme. These parameters were chosen to roughly correspond to the human out-of-Africa expansion: if we assume a local human population size of $N \approx 10,000$ and a generation time of 25 years, this corresponds to an expansion that started 65,000 years ago. The directionality index ψ and F_{ST} were calculated in Python; for ψ we used equation (3.1b), and F_{ST} was estimated using Reynold's estimator (Reynolds et al., 1983). Note that these are only baseline parameters, and exploring the effect of changing these parameters was the purpose of our power simulations.

Various significance tests can be used to determine the significance of ψ between two populations; for the case of $n = 2$ in both samples we can simply perform a binomial test on the absolute frequencies f_{21} and f_{12}. If their proportions differ significantly from 0.5, we can reject the null hypothesis of symmetric migration between the two demes. When comparing samples of size $n > 2$, we can generate a null distribution using a permutation test, i.e randomly assigning the allele frequencies for each SNP to either population. However, both these tests will underestimate the variance in the data if SNP are not in linkage equilibrium. In that case the "effective" number of loci will be lower than the actual number. To take linkage into account we use a computationally more computationally intensive block-jackknife approach (Busing et al., 1999; Reich et al., 2009) to analyze the human data.

To generate data for the 1D stepping stone model analyzed in Figure 1, we simulated a 201 x 1 habitat, with scaled migration rates $M = 1, 10$ between adjacent demes. Sampling was done in demes $-i/2$ and $i/2$, with the center deme having coordinate 0. In case of range expansions, the expansion started in deme $-i/2$.

SNP ascertainment may influence our results, because most ascertainment schemes favor high frequency alleles in the populations where the ascertainment was performed. To assess the effect of ascertainment bias on the value of ψ, we performed simulations in an isolation-by-distance stepping stone model with samples at coordinates (0,0), (10,0), (20,0), (30,0), (40,0), (50,0) as well as (0,10) and (15,10) and then computed ψ between the (10,0) and (20,0) sample. We then simulated ascertainment by selecting a set of population, and rejection sampled SNP so their 1D-SFS followed a Beta(2,4/3) distribution, which roughly matches the SFS in the HGDP data set and is very different from the expectation without ascertainment bias. We chose this ascertainment scheme as the original ascertainment scheme for HGDP is unknown. If ψ differs significantly from zero, then we know that ascertainment is important. Results are given in Figure 3.11; ascertainment is important if it is performed in one of the populations that we calculate ψ for. However, the effect of ascertainment is negligible if the population we calculate ψ for are different from the ascertainment population, even if the ascertainment population is much more closely related to some populations than to others.

3.4.2 Estimating the Origin of a Range Expansion

We use a time-difference of arrival (TDOA) approach (Gustafsson and Gunnarsson, 2003) to estimate the origin of a range expansion. TDOA was originally used in naval navigation during the Second World War, and is currently widely used to solve localization and navigation problems. It is based on the assumption that a single source emits a signal that decays with increasing distance from the origin. For range expansions, this signal is the difference in frequency of shared alleles. At the origin, the allele frequency is expected to be lowest (Slatkin and Excoffier, 2012) and to increase approximately linearly with distance. However, since we do not know the allele frequency at the origin, we have to use the indirect approach by comparing pairs of populations. To be precise, if we know that shared alleles have a lower frequency at point S_i compared to point S_j, then we know that S_i is closer to the origin than S_j. If the habitat is two-dimensional, however, this does not tell us the direction of the expansion. Let $||S_i, S_j||$ denote the Euclidean distance between two points S_i and S_j. Then,

$$||S_i, O|| - ||S_j, O|| \approx v\psi_{i,j}, \tag{3.3}$$

where O denotes the unknown origin $\psi_{i,j}$ is the directionality index between samples S_i and S_j and v is a constant that links space to allele frequency (i.e how much does the allele frequency change per unit of space). In words, $\psi_{i,j}$ is approximately proportional to the difference of the distances $||S_i, O||$ and $||S_j, O||$ (see also Figure 3.5). We assume that the sampling locations of S_i and S_j are known without error, and that $\psi_{i,j}$ can be estimated from genetic data, along with its sample variance $\mathrm{Var}(\psi_{i,j})$. We estimate the variance by doing 1,000 bootstrap replicates on the SNP. The unknowns that remain are the coordinates of the origin O and the proportionality constant v. To infer these parameters, we solve for ψ, subtract ψ from the equation and sum over all pairs of samples:

$$\left(\hat{O}, \hat{v}\right) = \underset{O,v}{\operatorname{argmax}} \sum_{i<j} \frac{1}{\mathrm{Var}(\psi_{i,j})} \left(\frac{||S_i, O|| - ||S_j, O||}{v} - \psi_{i,j} \right). \tag{3.4}$$

In most biological application, space will be two-dimensional and therefore we can make this equation more explicit by writing $O = (x, y)$ and $S_i = (x_i, y_i)$. Then,

$$\begin{aligned}(\hat{x}, \hat{y}, \hat{v}) &= \operatorname*{argmax}_{x,y,v} \sum_{i<j} \frac{1}{\operatorname{Var}(\psi_{i,j})} \\ &\times \left(\frac{1}{v}\left(\sqrt{(x_i - x)^2 + (y_i - y)^2} - \sqrt{(x_j - x)^2 + (y_j - y)^2}\right) - \psi_{i,j}\right). \quad (3.5)\end{aligned}$$

The variance terms correspond to weighting terms; terms where ψ has a high variance are weighted down, whereas terms where we can infer ψ with high accuracy are given a larger weight. We can then find a solution to this equation using nonlinear least squares.

3.5 Figures

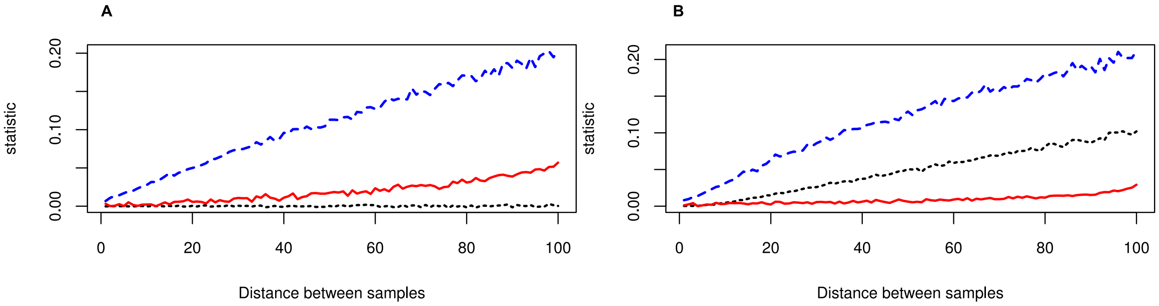

Figure 3.1. Behavior of H, ψ and F_{ST}.
This figure shows the behavior of H (red, full line), ψ (black, dotted) and F_{ST} (blue, dashed) in one-dimensional (A) isolation-by-distance and (B) population-expansion models. Simulations were performed on a 200 demestepping-stone model with scaled migration rate M=100 between adjacent demes, and expansion events every 0.001 coalescence units. F_{ST} increases linearly with distance in both models and ψ is zero in the isolation-by-distance model, but increases approximately linearly in the expansion model. Heterozygosity is plotted for demes from the center of the population (left) to the border of the habitat (right), and given as the difference to the central deme.

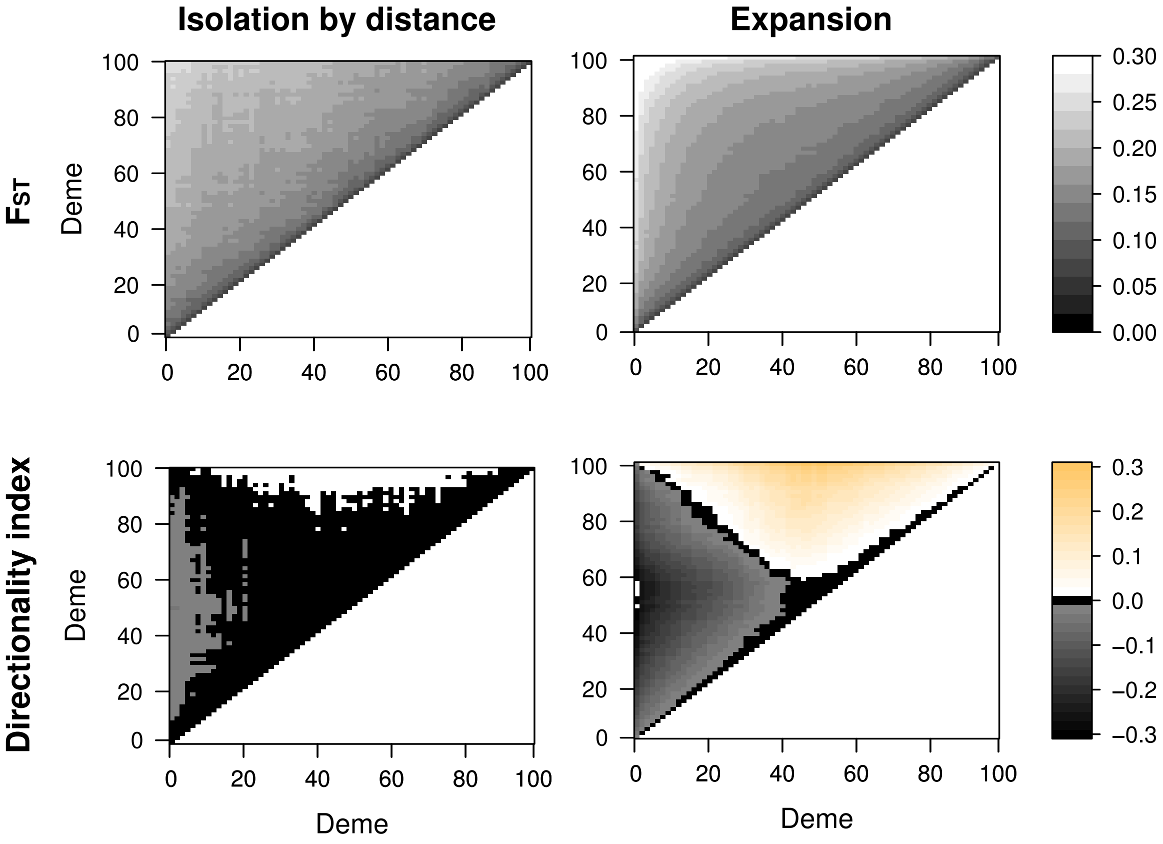

Figure 3.2. Behavior of F_{ST} and ψ

Each panel gives the value of the pairwise statistics F_{ST} and ψ under an isolation-by-distance model and an expansion model with the expansion starting in the central deme (50,50). Simulations were performed on a 101 × 101 deme stepping stone model, and a diagonal transect from demes at coordinates (0,0) to (100,100) was sampled, and all pairwise statistics were calculated. Black regions correspond to regions where F_{ST} and ψ are very low (below 1%). The orange and grey regions denote areas with positive and negative ψ, respectively. Whereas F_{ST} behaves qualitatively similar under both models, the behavior of ψ is very different. Under isolation-by-distance, ψ is very close to zero, with some deviations due to boundary effects. Under an expansion, however, we see a clear signal for all demes, except demes that are very close to each other, or demes that have the same distance to the origin, but in different directions.

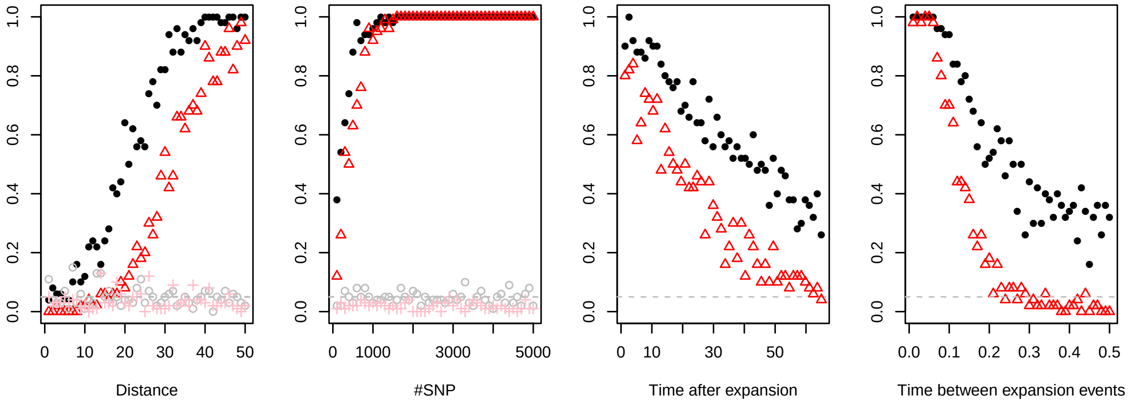

Figure 3.3. True/false positive rates of detecting range expansion. Each panel give the proportion of replicates in which the null model was rejected at the 5% significance level. Black circles correspond to ψ under an expansion model and an isolation-by-distance model, red triangles and plus signs denote simulations correspond to using H to distinguish an expansion model and isolation by distance model, respectively. The grey dashed line at 0.05 gives the expected proportion of false positives under the null hypothesis. Baseline parameters for the simulations were of 2 chromosomes (one diploid individual) at each location sampled, with locations a distance of 50 each other. Fixed parameters used for generating the data sets are 1,000 independent SNP from one diploid individual per sampled deme. Time between expansion events was set to 0.1 (coalescence units) and the data was observed immediately after the expansion ended.

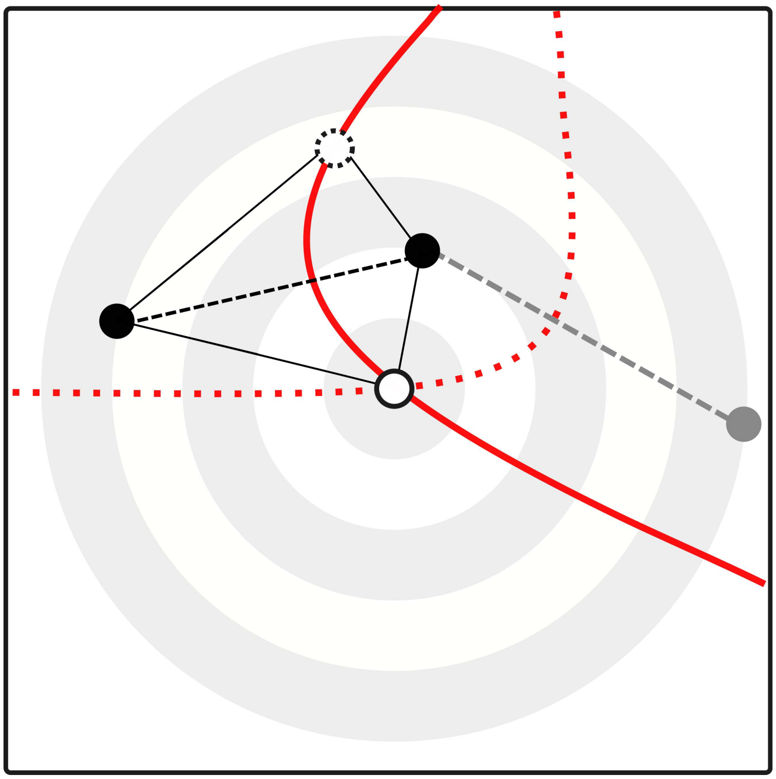

Figure 3.4. Illustration of the method used to infer the origin of a range expansion.

The black and grey points correspond to genetic samples taken, the white point corresponds to the (unknown) origin of the expansion. Using the directionality index ψ, we can infer the difference in distance from the samples to the origin (dashed lines). The set of all points that has the same difference in distance to the origin corresponds to the arm of a hyperbola (red), which comprises all candidate points according to ψ and the location of two points. Using a second pair of points (the grey and top black point), we can identify a second hyperbola (dotted), and find an unique location of the origin. In practice, we use more than three sampling locations. Sampling noise will cause the hyperbolas to not intersect in a single point and we use a least-squares criterion to estimate the location of the origin.

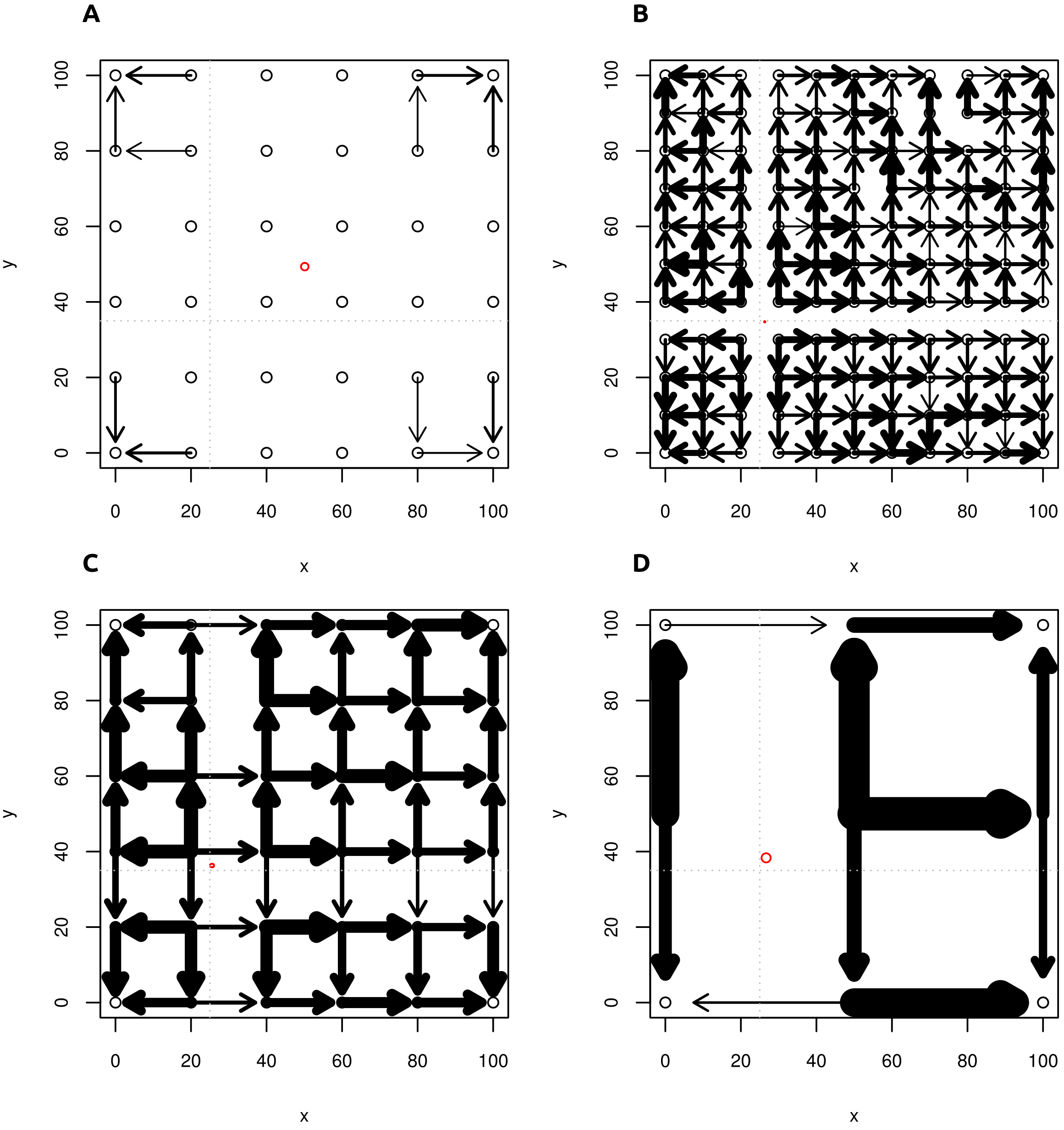

Figure 3.5. Detecting the origin of a range expansion.
Each panel corresponds to a 101 × 101 grid of populations that were simulated. The expansion began at point (25,35) (indicated by gray dotted lines). Black bordered circles indicate sampling locations, black arrows correspond to $\psi > 1\%$ between adjacent samples, with the direction of the arrow indicating the sign of ψ. Thicker arrows correspond to larger ψ. The red ellipse corresponds to the 95% confidence interval of the estimated location of the origin. Panel a: no expansion (isolation-by-distance model). Edge effects cause the estimated origin to be close to the center of the grid of populations. Panels b-d: Expansion with parameters $M = 1$, $t = 0.1$ and samples taken every 10th, 20th and 50th deme. While the confidence region is larger for smaller numbers of samples, we get a very accurate result even when we have only 9 samples.

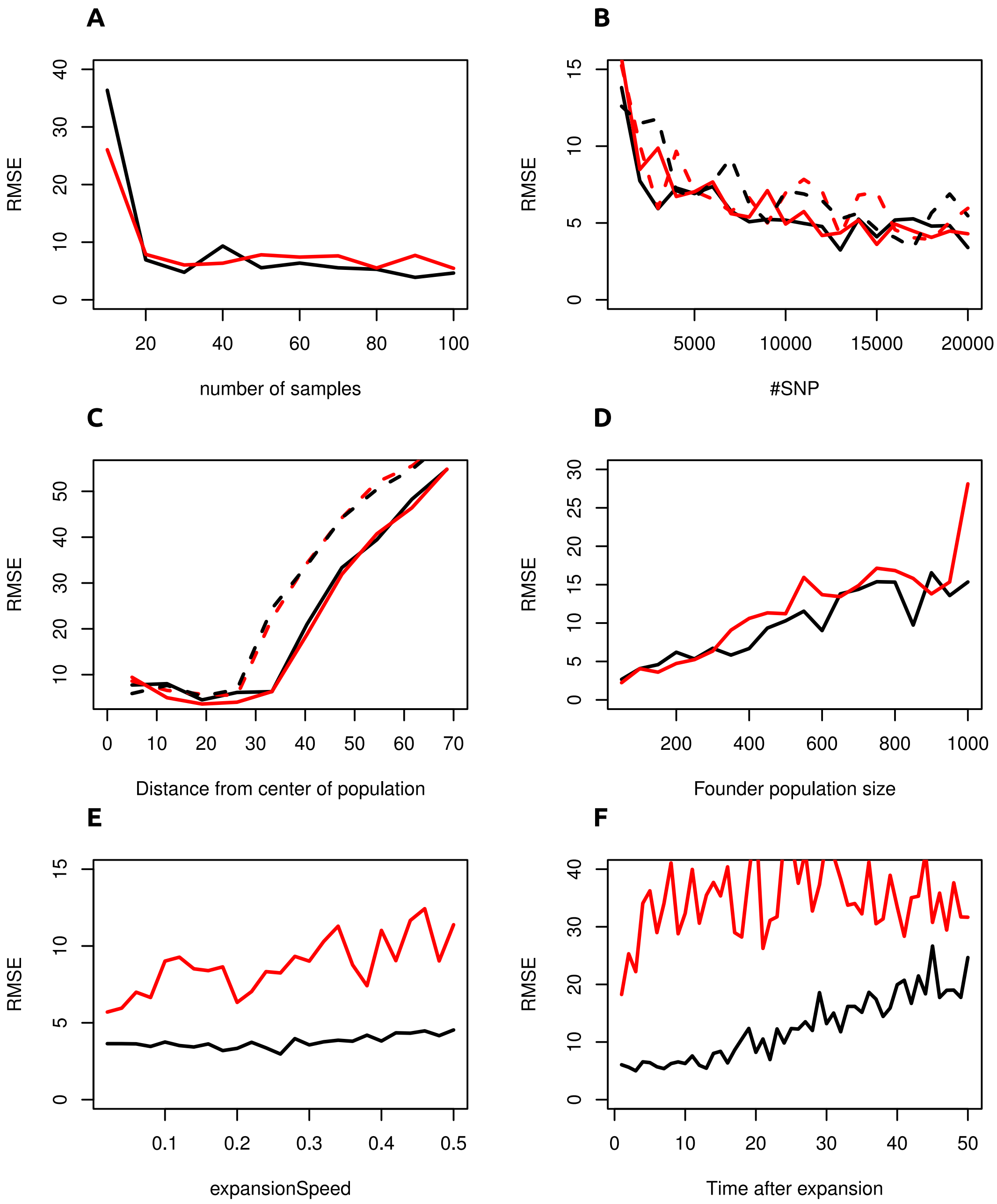

Figure 3.6. Performance of TDOA method.

We present the root mean squared errors (RSME) of our TDOA method (black) compare it with the method of Ramachandran et al. 2005 (red). Samples taken on a grid ware represented by full lines, whereas dashed lines denote samples that were taken from random coordinates in the simulated region. Our method is superior when the expansion occurred slowly or when it finished some time in the past; but the method perform very similar for recent, fast expansions.

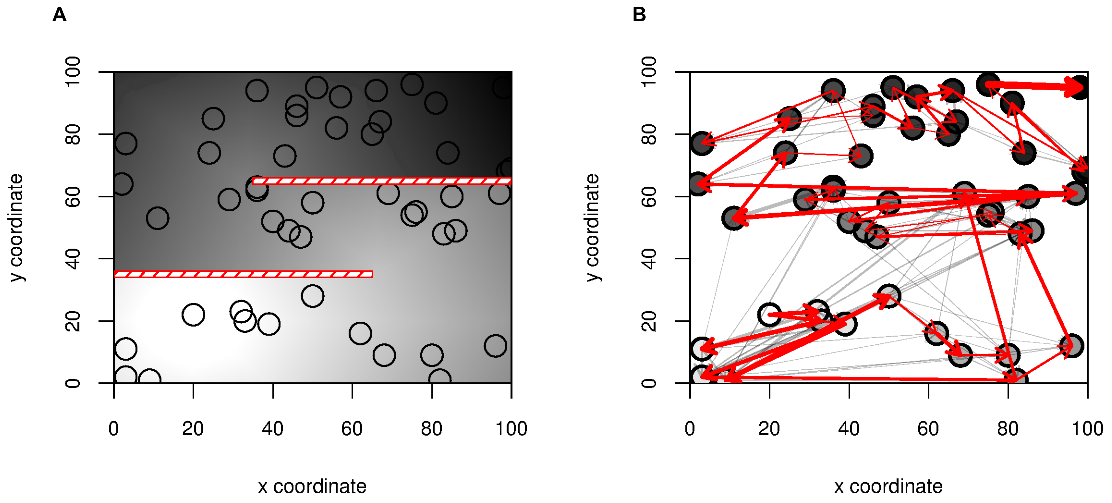

Figure 3.7. Identifying complex patterns of migration.
We simulated data on a S-shaped habitat with two impermeable barriers (Panel A) The darkness of the shading is proportional to the arrival time of the expansion, which began in deme (20,20). Black circles correspond to locations sampled. In Panel B we show the inferred pairwise directionality, with all edges remaining after thinning the graph shown in grey, and a maximum spanning tree in red. We also show the inferred ordering of the samples as a color gradient of the samples from light (closest to origin) to dark. The barriers can be identified from panel B by the absence of any indication of gene flow across the barriers and by examining the ordering of the samples.

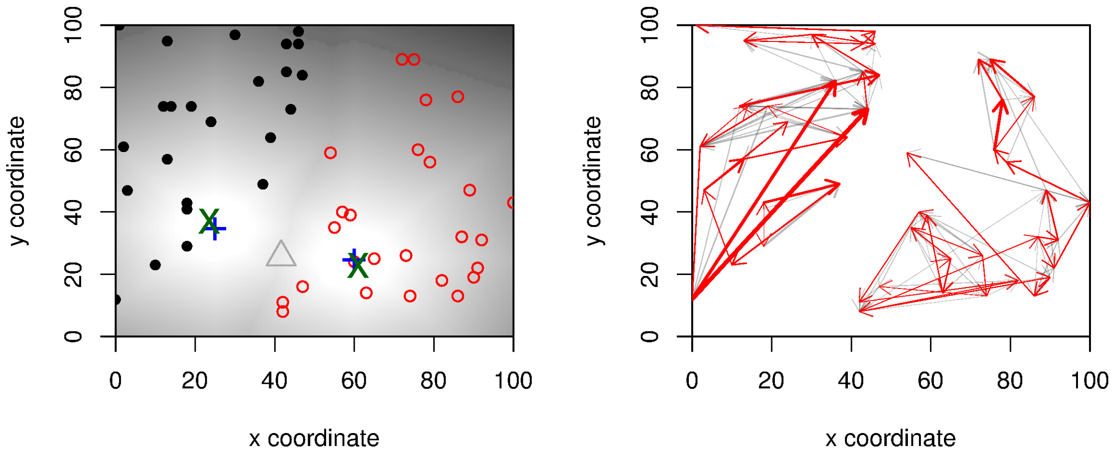

Figure 3.8. Detecting multiple origins.

Panel a: We simulated two expansions that originated at the same time from origins indicated by the blue crosses. The color gradient in the background corresponds to the time of colonization time of each deme. We address the problem of inferring the origin of multiple expansions using a two-step procedure. First, we cluster the samples into discrete clusters (red and black circles, respectively) and then estimate the expansion signal and origins independently for the clusters, resulting in high accuracy for both estimated origins (green X) when compared to the actual origins (blue +). The grey triangle denotes the estimated single origin if we did not do the two step procedure; it lies approximately half way between the two actual origins. The right panel shows the inferred migration patterns after a transitive reduction (grey/red arrows) and a maximum spanning tree (red arrows).

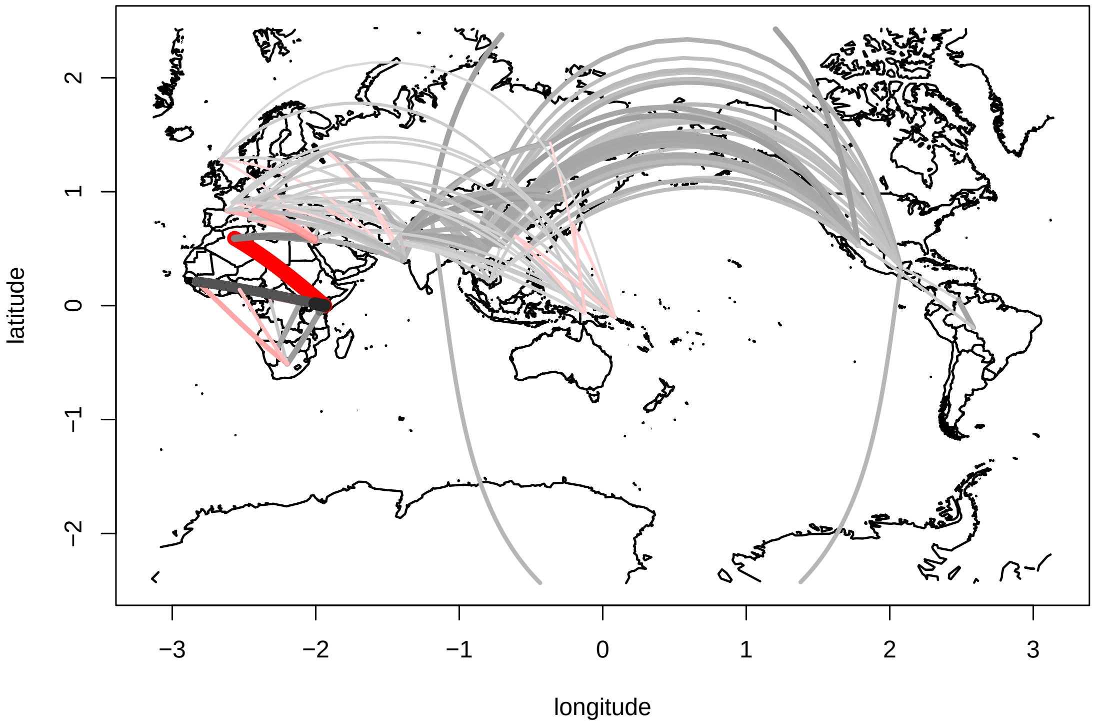

Figure 3.9. Inference of human migration routes.
The figure shows a visual representation of the pairwise directionality indices between human populations in HGDP and HapMap. Each line corresponds to the pairwise ψ statistic, with thicker and brighter lines corresponding to higher values. Grey and red lines denote eastward and westward migration, respectively. Lines with an absolute Z-score below 5 were omitted.

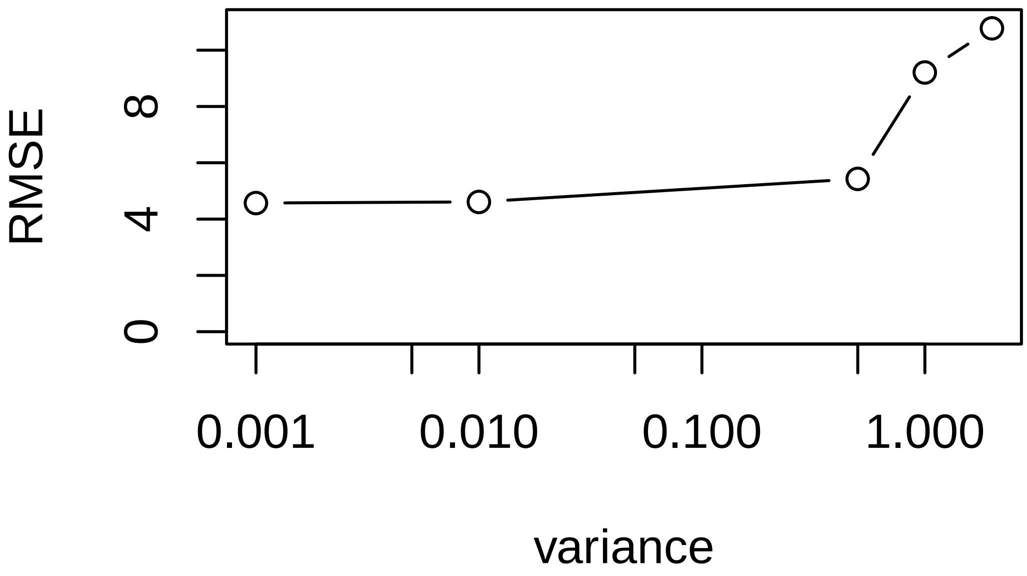

Figure 3.10. Effect of population size variance on the estimation of the origin.
We show the effect of variance in population size on our ability to estimate the origin of a range expansion. RMSE increases only slightly if the variance is below 0.1, but increases quite drastically if the variance is higher than that.

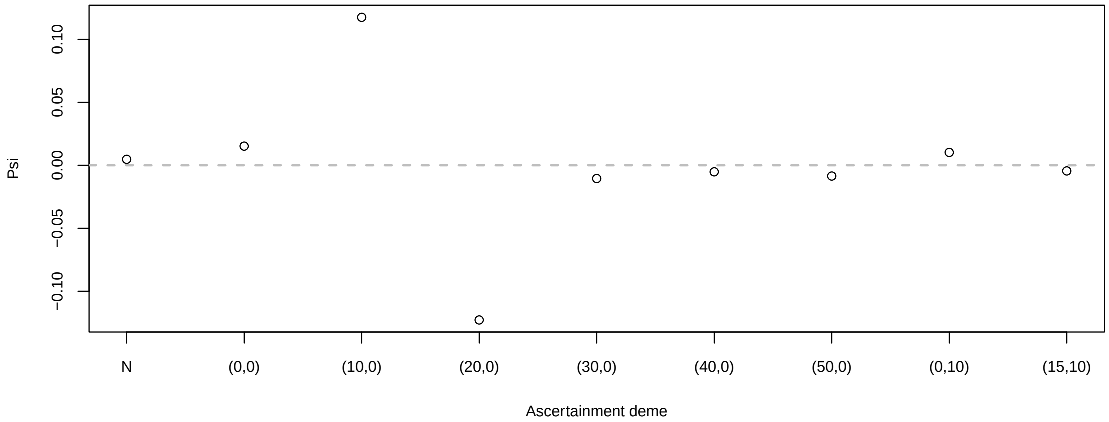

Figure 3.11. Effect of ascertainment bias on ψ.

We show the effect of strong ascertainment bias in different demes given on the x-axis on ψ calculated between samples taken from coordinates (0,10) and (0,20). The first column shows results with no ascertainment. Ascertainment has very little effect if it is performed in a deme that is not used in the comparison.

Chapter 4

The Effective Founder Effect in a Spatially Expanding Population

4.1 Introduction

We may think of a range expansion as the spread of a species or population from a narrow, geographically restricted region to a much larger habitat. Range expansions are a common occurrence in many species and systems, and they happen on time scales that differ by orders of magnitude. Viruses and bacteria may spread across the globe in a few weeks (Brockmann and Helbing, 2013), invasive species are able to colonize new habitats over decades (Davis, 2009); and many species migrated into their current habitat over the last few millennia, following changing environments such as receding glaciers (Hewitt, 1999; Taberlet et al., 1998).

The population genetic theory of range expansion is based on two largely distinct models. The first model, based on the seminal papers of Fisher (1937) and Kolmogorov et al. (1937) and often called the Fisher-KPP model, is based on the diffusive spread of alleles, and has been mostly explored from a statistical mechanics viewpoint. The other model, called the serial founder model, has its roots in the empirically observed decrease in genetic diversity from an expansion origin (Austerlitz et al., 1997; Hewitt, 1999; Ramachandran et al., 2005).

The Fisher-KPP partial differential equation describes the deterministic change in allele frequency at a spatial location due to the individuals with a selected allele having more offspring than wild-type individuals. The model can be applied to range expansions by substituting the selected allele with presence of a species, and the wild-type allele as absence of a species(see e.g. Barton et al., 2013, for a recent review). Its solution is a travelling wave; similar to logistic growth where populations grow to some carrying capacity. This model has received more widespread attention recently due to the empirical tests by Hallatschek et al. (2007), who compared growing colonies of *E. coli* to the predictions from the Fisher-KPP equation.

However, it is also apparent that some of the predictions of the Fisher-KPP model are inconsistent with many macroscopic systems. In particular, the Fisher-KPP model predicts that local populations start with extremely small population sizes. This leads to a very high amount of genetic drift, and, for example, in the experiments of Hallatschek et al. (2007), all local genetic variability was quickly eliminated, so that no polymorphisms were shared between individuals sufficiently far from each other. This is in strong contrast to humans, for example, where an expansion out-of-Africa is well supported (e.g. Ramachandran et al., 2005), but where we find many genetic variants shared between all human populations. This was part of the motivation for the development of the serial founder model, which is typically based on a variation of a stepping stone model (Kimura, 1964) in one or two dimensions. Typically, only a small subset of populations is colonized at the beginning of the process, but over time subsequent populations are colonized, usually by means of some founder effect. There are multiple ways a founder effect has been modelled: Austerlitz et al. (1997) and Ray et al. (2010) chose to model the founder effect by local logistic growth, where each local subpopulation grows according to the logistic equation. A simpler model, favored by DeGiorgio et al. (2011) and Slatkin and Excoffier (2012), is to model the founder effect by a temporary reduction in population size.

A complete serial founder model, including selection, founder events and migration between subpopulations, has, so far, be proven to complex to be of use for analytical research. However, a recursion approach (Austerlitz et al., 1997) and simulations (Klopfstein et al. 2006, Travis & Burton 2010) have been successfully applied to investigate the behavior of the model. Other alternatives are models that do not model expansions explicitly, and make additional simplifications. Perhaps the simplest model of this kind is that of a demographic expansion without any spatial component, which can be fully described by a change in the rate of coalescence (Gravel et al., 2011; Li and Durbin, 2011), with the assumption made that the population is panmictic throughout the expansion. This model resembles a spatially explicit expansion, when migration rates between demes are very large. A more sophisticated model, incorporating spatial structure, is the infinite island approximation model of Excoffier (2004). In this model, an originally small, panmictic population expands instantly into a metapopulation with a large number of subpopulations. In contrast to the demographic expansion, in this model we can compare coalescent time distributions between demes and within demes (called the mismatch distribution), which has been used for inference previously. However, this model assumes that all subpopulations are exchangeable, so that there is no difference in coalescence times with individuals at a wave front, when compared to individuals in the center of a population. A further step were the models of DeGiorgio et al. (2011) and Slatkin and Excoffier (2012). DeGiorgio et al. (2011) derived coalescence times under a serial founder model, using a small bottleneck as a founder event. In the model of Slatkin and Excoffier (2012), the expansion is modelled as a spatial analog of genetic drift, where each founder event corresponds to a generation in a standard Wright-Fisher model.

So far, these theoretical models of range expansions have let to few applications that can be applied to interpret genetic data from non-model organisms. In this paper, we first develop a simple model of a range expansions based on a branching process approximation. The advantage of the simplicity of the model is that it leads to the development of an intuitive understanding of an expansion. We test the model using simulations, and discuss its limitations, and then show how it can be used for inference.

We demonstrate the utility of our approach by re-analyzing SNP data of the model plant species *Arabidopsis thaliana* (L.) Heynh. (Horton et al., 2012). *A. thaliana* is a small, annual plant, thought to be native to Europe, but introduced in North America and other locations (Jorgensen and Mauricio, 2004). The biogeography and population structure of *A. thaliana* has been well studied (Horton et al., 2012; Jorgensen and Mauricio, 2004; Nordborg et al., 2005). While earliest studies showed relatively little population differentiation on a global scale, genome-wide genetic data supports widespread population structure and clear genetic differentiation between populations (Horton et al., 2012; Nordborg et al., 2005). The availability of genome-wide SNP-chip data from more than a thousand individual plants from hundreds of locations make *A. thaliana* an ideal test case for the genetic signatures of range expansions. However, the status of *A. thaliana* as a human commensialist and the fact that *A. thaliana* is a selfing plant, may make the analysis more challenging.

4.2 Results

4.2.1 Overview of Theoretical Results

In this section, we will briefly outline our model and the main theoretical results. Details and full derivations can be found in Section 4.5. A schematic of the model studied is given in Figure 4.1. In brief, we assume a serial founder model on a one dimensional stepping stone grid, where initially only one deme is colonized. We compare the allele frequency of individuals in the same location as the origin of the population at time t, X_t, with the allele frequencies of individuals at the wave front at time t, which we denote by $\tilde{X}_t$. In particular, we are interested in the difference in derived allele frequency between the population at the starting position and the expansion front, which we denote as Z_t. In Section 4.5.1, we show that the expected difference in allele frequency is

$$\mathbb{E}Z_t = f_0 \left(\frac{1}{1-\tilde{L}(t)} - \frac{1}{1-L(t)} \right), \tag{4.1}$$

where f_0 is the initial frequency of an allele, and $L(t)$ and $\tilde{L}(t)$ are the probabilities that an allele is lost by time t at the origin of the expansion and the wave front, respectively.

We can make this result more explicit assuming the populations evolve according to a branching process. A (Galton-Watson) branching process (Harris, 1954) models the evolution of a population by assuming that all individuals produce offspring independently from each other, with some offspring distribution F. In Section 4.5.2 we use standard results from branching process theory to show that if each deme evolves according to a branching process, then (4.1) can be written as

$$\mathbb{E}Z_t = \frac{1}{2}\left(\mathrm{Var}(\tilde{F}) - \mathrm{Var}(F)\right)t + o\left(\frac{1}{t}\right), \tag{4.2}$$

that is, the difference in allele frequency is expected to increase linearly with distance, and the slope is half the difference in the variance of offspring distribution at the expansion front $\mathrm{Var}(\tilde{F})$ and away from the expansion front $\mathrm{Var}(F)$. Since we assume that founder effects occur at the expansion front, we expect it to have a higher offspring variance, corresponding to a lower effective population size. It is worth pointing out that the term of order t in $\mathbb{E}Z_t$ does not depend on f_0, so that we expect the same slope independent of the initial allele frequency. As the higher order terms depend on f_0, we will examine the accuracy of this result using simulations.

In Section 4.5.3 we then use the offspring variance from a Wright-Fisher model to define an effective founder size k_e, and show that

$$\mathbb{E}Z_t = \frac{1}{2}\left(\frac{N_e}{k_e} - 1\right)t, \tag{4.3}$$

where N_e is the effective population size of a deme. In some cases, it might be possible to interpret N_e and k_e directly. For example, if we think of a species colonizing a system of islands, N_e corresponds to the carrying capacity of that species on a given island, and k_e to the number of founders. In most cases, however, subpopulations are not clearly defined and the population is relatively continuously distributed. Under these circumstances, it is not clear what N_e and k_e represent. Therefore, we show in Section 4.5.4 that it makes more sense to think about the distance over which the ratio $\frac{k_e}{N_e}$ has a certain value (e.g. 0.99), that is, how far apart demes need to be so that each founding population is 1% lower than the population at equilibrium. The larger this distance, the weaker the founder effect.

Finally, in Section 4.5.5 we show how we can estimate $\mathbb{E}Z_t$ from genetic data using the ψ statistic defined as

$$\psi = \frac{f_{21} - f_{12}}{f_{12} + f_{11} + f_{21}}, \tag{4.4}$$

where f_{ij} is the (i, j) entry in the allele frequency spectrum. ψ was introduced by Peter and Slatkin (2013), and we show in Section 4.5.5 why ψ is an useful estimator of $\mathbb{E}Z_t$. Taken together, these results suggest that we can define and estimate the effective founder effect that describes the loss of genetic diversity with distance from the expansion origin, and that we can infer the strength of the founder effect using a simple linear regression on the allele frequency of shared alleles.

4.2.2 Simulations

We validate our analytical results by performing extensive simulations under two different models. The first is the forward-in-time model stepping stone model described by Slatkin and Excoffier (2012). The second model is a backward-in-time stepping stone model, based on the Kingman coalescent (Wakeley, 2009).

4.2.2.1 Forward Simulations

We first validate our results using a discrete-time, forward-in-time Wright-Fisher model. In Figure 4.2, we give results for various initial allele frequencies f_0, setting $k_e = 0.1N$ (first row), $k_e = 0.5N$ (middle row) and $k_e = 0.9N$ (bottom row). Using Equation 4.3, we would predict Z_t to be $4.5t$, $0.5t$ and $t/18$, respectively. Those predictions are given by the red lines; the points represent data observed in simulations. We find that we get better estimates when i) the effective founder size is low, ii) the time after the expansion is low and iii) the effective population size is high. In particular, we find that we get a systematic bias when we have a very strong founder effect, and thus allele frequency differences are expected to be very large. In that case, many alleles will become fixed in the population, and the predictions between the Wright-Fisher and the branching process models are quite different.

In Figure 4.3, we investigate the effect of demes growing to their carrying capacity via logistic growth, as opposed to instantaneous growth which we assume in most other cases. Here we can apply the result that under non-constant founder population sizes, the effective founder size is simply the harmonic mean of all founder sizes, divided by the number of generations. In Figure 4.3, simulations were performed with a carrying capacity of 10,000, and growth starting at a size of 1,000 individuals, with subsequent founder effects starting after $k = 1, 2, \dots 10$ generations, corresponding to the different data series.

4.2.2.2 Backwards Simulations

We also performed backward-in-time simulations i) to test the robustness of the branching process predictions to migration, ii) to test the effect of estimation from a subsample and iii) to remove the initial allele frequency as an explicit parameter. Coalescent simulations were performed in a continuous-time model with discrete expansion events. In particular, most of the time lineages are allowed to merge according to the standard structured coalescent. The only exception are the expansion events, which are modelled as a single generation of Wright-Fisher mating, followed by moving all lineages in the newly colonized deme back to the founder deme. Thus, unlike the Kingman-coalescent, this model allows for multiple mergers at the wavefront. Under this model,

$$\mathbb{E}Z_t = \frac{1}{4k_e}, \tag{4.5}$$

since the founder effects result in an increase of the offspring variance by a factor of $(2k_e)^{-1}$. We estimate $\mathbb{E}Z_t$ using the ψ-statistic defined in Peter and Slatkin (2013), justification of this is given in Section 4.5.5. Results are displayed in Figure 4.4. In the top row, we show samples taken immediately after the expansion reached the boundary of the habitat, in the bottom row, we show samples that were taken a very long time ($20N$ generations) after the expansion finished. We find that recent expansions are detected rather easily, almost independent of the migration rate, and the effective founder size is estimated with high accuracy. In the bottom row, we observe that for intermediate migration rates ($M = 0.1$ and $M = 1$), we still get a relatively accurate result, however, we have more noise, indicating that much larger samples would be required to obtain confident estimates, since most SNP will be either fixed or lost after this time. For a low migration rate, we see that we do not have any power for inference, since individuals all coalesce within their demes before they have the opportunity to coalesce with lineages from other locations. For high migration rates ($M = 100$), we find that the signal of the range expansion has almost vanished. Under these conditions, migration is so strong that the population essentially resembles an equilibrium isolation-by-distance population, and the signal of the range expansion has been lost. For $M = 10$, we see an intermediate behaviour, close to the origin we are at equilibrium, but far away the slope of the curve is still the same as we would expect under an expansion.

4.2.2.3 2D Simulations

In addition to the results presented in the previous two sections, where we performed simulations in a one-dimensional habitat, we also performed simulations on a 2D-stepping stone model to investigate the impact of multiple dimensions. We performed simulations by simulating expansions both under a migrant pool model and a propagule pool model (Slatkin and Wade, 1978). In a migrant pool model, all neighbouring populations send migrants at equal rates, whereas under the propagule pool model, one possible founder population is selected to send out a “propagule”, which colonizes the new deme. We find that if the sample axis is parallel to the orientation of the stepping-stone-grid, then the migration model does not matter, and we get the same behavior as in the 1D case. However, if we sample a diagonal we find that the results from the 1D-simulations are applicable under the propagule pool model, but not under the migrant pool model (Figure 4.5). The reason for that is quite simple: under the migrant pool model, there are many different paths on how a deme can be colonized, and the number of paths increases with distance from the origin, which reduces the amount of drift in a non-linear way with distance from the origin. In contrast, under the propagule pool model, always one path is chosen. We also find that under the 2D-model the signal of the expansion disappears faster. Whereas in the 1D-model at a migration rate of $M = 1$ the expansion is still detectable after $20N$ generations, we find that at the same migration rate, the population already approaches equilibrium.

4.2.3 Application to *A. thaliana*

We applied our model to the SNP dataset of Horton et al. (2012). Based on a PCA analysis (Figure 4.6a) and the sample locations, we defined five regions for further analysis: Scandinavia, Americas, as well as Western, Central and Eastern Europe.

In the Americas, (Figure 4.6c) we find a most likely expansion origin in the Great Lakes region, as opposed to the East coast. This might be somewhat surprising, but as we only have one sampling location at the East cost, power might be low. However, we can speculate that traders through the Great Lakes first introduced *A. thaliana* to the US, which may explain the signal. This may be seen as evidence for a recent introduction of *A. thaliana* into the Americas from Europe. Further support for this hypothesis comes from the fact that the American samples cluster together with the Western European samples in the PCA analysis, and from the fact that the North American population shows the strongest founder effect, with a 1% founder effect every 4.26 km.

Scandinavia (green dots in Figure 4.6a, Figure 4.6c), shows the most diversity within a region according to the PCA plot, and the second highest founder effect size. The most distinct samples, in the bottom left of Figure 4.6, are from Northern Sweden and Finland, whereas those samples that cluster with the Central and Eastern European Accessions are predominately found in Southern Sweden. We find evidence of immigration from the East, with the most likely origin of the Scandinavian accessions lying in Finland. Based on the PCA analysis, we might expect the accessions from Southern Sweden to show evidence of a range expansion from the South, and that is indeed the case when we only consider these Southern samples.

If we analyze these Southern Scandinavian samples together with the samples from Austria, Czech Republic, Russia, Lithuania and Tajikistan (Figure 4.6d, pink and brown dots in the PCA), we find evidence of an expansion out of eastern Asia, possibly from a refugium close to the Caspian Sea. For the Central European samples, we find an origin close to the border between Austria and Italy. This is likely a proxy for a refugium in either Southern Italy or the Balkan region, as the inferred origin was covered by an ice sheet during the last glacial maximum. Finally, for the Western European samples we find the weakest founder effect among all analyzed region, with a 1% founder effect at a scale of 38.6 km, almost an order of magnitude weaker than the strongest founder effect we observed in this set of populations, in the Americas. This is however partly due to the aggregation of the British and continental samples; if we just analyze the French, Spanish and Portuguese samples (excluding the British samples), we find a founder effect of 18.7, in line with the other continental European regions. In contrast, if we analyse the British samples separately, we estimate an 1% decrease to occur over 47.8 km, and in fact we cannot exclude equilibrium isolation by distance, as, after Bonferroni correction, as $p > 0.05$.

4.3 Discussion

In this paper, we study range expansions using a serial founder model, with the main goal to develop inference procedures. We use a branching process approximation to approximate the decay of genetic diversity due to the recurring founder effects. We use this approximation to define an effective founder size, which can be estimated using standard linear regression from genetic data.

A linear or approximately linear decline of genetic diversity with distance has been observed previously in humans (DeGiorgio et al., 2009; Ramachandran et al., 2005) and in simulations (DeGiorgio et al., 2009; Peter and Slatkin, 2013). In previous work, we showed, using simulations, that the directionality index ψ, defined in Equation 4.4, increases approximately linearly with distance (Peter and Slatkin, 2013). In this paper, we connect these empirical observations with a theoretical model, that explains this decay in terms of differences in offspring variance. This is justified because in populations with a higher offspring variance, genetic drift occurs faster and therefore the population's effective size becomes smaller.

While branching processes have a long history in population genetics (Ewens, 2004), they differ from other commonly used models such as the Wright-Fisher model and the Coalescent in that the total number of individuals in the population is not constant (or following a predetermined function). Instead, the expected number of individuals is constant, leading to different dynamics. For example, a neutral branching process will eventually die out almost surely, something that cannot happen under the Wright-Fisher model. Therefore, the models presented here are only useful in parameter regions where the branching process model and other population genetic models result in similar dynamics. For example, our model breaks down if there are only few shared variants between populations. However, in this case, phylogeographic methods are arguably more appropriate than population genetic ones. Otherwise, our model appears to be useful as long as a significant fraction of variants has a most recent common ancestor during the expansion or before the expansion started. If that is not the case, as in the simulations with high T and high M in Figure 4.4, we find that ψ will be very close to zero, due to the signal of the expansion vanishing over time. The last parameter region where the model breaks down is when the mean allele frequencies become very large (Figure 4.2. In that case, the increase slows down due to fixations in the Wright-Fisher model, whereas it may further increase under the branching process approximation, explaining the difference.

Similar to the effective founder size defined in Slatkin and Excoffier (2012), the effective founder size k_e we defined here is a variance effective size. This is different from the model proposed by DeGiorgio et al. (2009), where an explicit bottleneck was used to model a founder effect. Using an effective size is less specific than that – there are many models that will lead to the same founder size – but has the advantage that the same formalism can be applied to many different situations. We also showed various rescaling properties. Perhaps counterintuitively, $\mathbb{E}Z_t$ is largely independent of the expansion speed, conditional on k_e. The reason for that is that, even though more segregating variants will be lost in a faster expansion, the difference between the expansion front and the rest of the population remains the same. Similarly, waiting after the expansion finished will not change $\mathbb{E}Z_t$.

Of course, an effective size has its limitations, as in essence, it is just a measure of the speed of genetic drift. Many other models exist that may lead to similar or identical patterns of genetic diversity (DeGiorgio et al., 2009). However, in many cases it is natural to assume a range expansion occurred, often through climatic or historical evidence. In these cases, our framework may provide a starting point for a genetic analysis.

The analysis of the *A. thaliana* data shows both the usefulness and some of the limitations of our approach. We are able to identify expansion origins and infer the strength of the founder effect from genetic data. In the *A. thaliana* data set, we find that the founder effect is much stronger in the Americas than in continental Europe. This is an interesting pattern, and it would be very interesting to see if the same is true for other introduced or invasive species. In Europe, our results are consistent with previous analyses by Nordborg et al. (2005) and François et al. (2008). Nordborg et al. (2005) found that Arabidopsis likely colonized Scandinavia both from the East, through Finland, and from the South. The strong population structure is consistent with this finding a global pattern of an Eastern origin, and evidence for immigration from the South when just analyzing the Southern Swedish samples, or if we jointly analyze them with Eastern European and Asian samples. Overall, we identify a likely ice-age refugium close to the Pyrenees in Southern France or Eastern Spain, a likely refugium near the Caspian sea and a refugium in central Southern Europe, either in the Balkan or Italy, where denser sampling is required for a more accurate picture. In the Americas, we find that Arabidopsis experienced very strong founder events, and we identify a most likely point of introduction near the Great Lakes.

On the other hand, describing the founder effect as a distance over which genetic diversity decreases by a certain amount is not as satisfying as is the inference of an effective founder size, on the same scale as the effective population size. However, it is necessary because of scaling reasons; if a single population spans a larger area, then we necessarily need a strong founder effect to get the same diversity gradient than. On the other hand, if we subdivide the area of the large population into smaller populations, each of those will have its own, smaller founder effect, but the population will experience a larger number of founder events. Thus, if we know the scale of a local population, or can reasonably approximate it (e.g. if we know the dispersal distance of the species). We can obtain an estimate on how much lower the founder size is compared to the effective size at carrying capacity in equilibrium. On the other hand, interpreting the founder effect as a distance allows us to obtain a measure that is independent of how populations actually occupy space, which is more versatile, but somewhat harder to interpret.

4.4 Methods

4.4.1 Forward Wright-Fisher-Simulations

Forward simulations were performed using a simple simulator implemented in R. Simulations were started with a fixed initial frequency f_0, and allowed to evolve for a fixed number of generations. Every t-th generation, the rightmost deme founded a new population, first with a single Wright-Fisher generation of size k_e, which then, in the next generation, expanded to size N. All demes except the newly founded one underwent t generations of Wright-Fisher mating in the same time frame, thus after gt generations, g demes are colonized. $\mathbb{E}Z_t$ was estimated from 10^6 replicate alleles. More complex models were implemented the same framework, i.e. we added migration between all demes at each generation and allowed the population to evolve for additional generations after the expansion finished. We also used a modification that allowed for changes in population size after each expansion event, and we used this modification to study the effect of logistic growth (see Figure 4.3).

4.4.2 Backwards Simulations

Backward-in-time simulations were performed using the standard structured coalescent model Wakeley (2009), with a minor modification. The structured coalescent allows easy inclusion of migration events. The coalescent is usually studied in the continuous limit where the number of generations and population sizes are both very large. We follow this approximation with the exception of expansion events, which are modelled using a single generation of Wright-Fisher-mating. Backward in time, we stochastically merge lineages, the backwards-transition probability for the number of lineages is (Watterson 1975, Wakeley 2008, p. 62):

$$\mathbb{P}(L_t + 1 = j | L_t = i; k_e) = \frac{S_i^{(j)} k_{e[j]}}{k_e^i}, \tag{4.6}$$

where k_e is the effective founder size, L_t is the number of lineages at time t, (time measured backwards in time in coalescence units), $S_i^{(j)}$ is the Stirling number of the second kind and $N_{[j]}$ is the jth falling factorial. If the number of lineages is reduced, we merge lineages uniformly at random. All remaining lineages are then transported to a neighbouring colonized deme. To compare this model to our predictions from the branching process model, we have to consider the excess variance in offspring distribution resulting from these expansion events, which is $\frac{1}{4k_e}$, such that for this coalescent model

$$\mathbb{E}Z_t = \frac{1}{4k}t + O\left(\frac{1}{t}\right). \tag{4.7}$$

Thus, the smaller the effective founder size k_e, the larger the allele frequency gradient will be. 1D-and 2D-simulations were performed using the same simulator. For 1D-simulations, we sampled eleven samples with n lineages every 5th deme, with 20 additional demes to avoid boundary effects. For the 2D-simulations, we sampled both a diagonal and horizontal transect. The horizontal transect, parallel to the demic structure, had length 30. The diagonal transect, where demes were colonized every $\sqrt{2}t$ time units, had length $20\sqrt{2}$, so that both transect are colonized in approximately the same time.

4.4.3 Application

The data set of Horton et al. (2012), along with the coordinates for the accessions was downloaded from the website at `http://bergelson.uchicago.edu/regmap-data/`. Genotypes of the sister species *Arabidopsis lyrata* provided by Matthew Horton were used to determine the ancestral state for each SNP. SNP where we could not determine an ancestral state unambiguously, either because no homolog *A. lyrata* allele was found, or the allele *A. lyrata* was not present in *A. thaliana*, were removed. Similarly, we removed all individuals where we did not have sampling coordinates. Since *A. thaliana* is a selfing plant and highly inbred accessions were sequenced, we only had a single haploid genotype per individual. Since our methodology requires at least two sampled haplotypes, we restricted our analysis to locations with at least two accessions sampled. To avoid bias due to very closely related accessions, we subsequently removed locations where the plants differed at less than 1.5% of sites (average heterozygosity of locations was 7.1%, with a standard deviation of 3.2%). This resulted in a total of 149 locations with at least two samples, representing 855 individuals, with 121,412 SNP genotypes remaining. As a single, uniform expansion throughout Europe seems rather unlikely, we performed a PCA analysis to find the main axes of population differentiation (Figure 4.6a). As the resulting pattern divided the samples broadly into four different groups, we analyzed data from these groups separately. These groups are: Americas (black), Western Europe (blue), Central Europe (red) and Scandinavia (green). For each of these groups, we estimated the origin of the range expansion using Equation 5 of Peter and Slatkin (2013). For visualization, we evaluated eq 5 of Peter and Slatkin (2013) on a grid (with locations not falling on land excluded), and estimated the best fit for the slope parameter (v) using linear regressions, with the location with the highest r^2 corresponding to the least squared estimate of the origin of the expansion (Figure 4.6b-f).

The expected value of ψ depends on the ratio of the effective founder size k_e to the effective population size N_e and the number of demes that the population colonized. The number of demes is relevant, since if we subdivide the population into more demes, it will undergo more (but weaker founder effects) over the same physical distance, or conversely, if we assume that demes are large, then we have few founder events with a very strong founder effect. Using the simple model developed in this paper, we cannot distinguish these cases without additional extraneous information. For example, we may fix the size of each deme based on extraneous information. For example, if the mean dispersal distance is known for a species, we may assume that the spatial extent of each deme is approximately that dispersal distance, and we can calculate k_e relative to that quantity.

In this context, the ratio k_e/N_e has the interpretation as the percentwise reduction in Wright's neighborhood size. Alternatively, if the dispersal distance is unknown, we may fix the ratio $r = k_e/N_e$ to an arbitrary constant, and instead report the required distance x_e over which the effective founder size is k_e. This has the advantage that it provides us with a quantity that is independent of assumptions of the demic structure, and the larger x_e is, the weaker is the founder effect of the population. For illustration purposes, we calculate the ratio k_e/N_e for deme sizes of 1km, 10km and 100km, as well as x_e for all groups and report them in Table 4.1.

4.5 Derivation of Main Results

4.5.1 Discrete Time Expansion Model

We model a range expansion on a one-dimensional stepping stone model with potential deme positions $0, 1, 2, \dots$ labelled $d_i, i = 0, 1, \dots$. All but deme d_0 are not colonized at the start of the process. We denote the frequency of an allele of a biallelic marker in deme d_i at time t as $f_i(t)$, and we assume that $f_0(0) = f_0$, where f_0 is some constant. The population behaves as a Markov process, so that the allele frequencies at time step t only depend on step $t-1$. Each time step, genetic drift will change allele frequencies according to some probability distribution. In addition, deme d_t will become colonized by the offspring of individuals present at time $t-1$ in deme d_{t-1} according to some other probability distribution. For simplicity, we at first assume there is no migration between demes, and test the robustness to this assumption using simulations.

Let $\{X_t\} = \{f_0(0), f_0(1), \dots f_0(t)\}$ and $\{\tilde{X}_t\} = \{f_0(0), f_1(1), \dots f_t(t)\}$ be the processes describing allele frequencies at and away from the wave front, respectively. Since we disallow migration, we can describe the history of "intermediate" demes $d_i, 0 < i < t$ by processes $\{X_t^{(i)}\} = \{f_0(0), f_1(1), \dots f_i(i), f_i(i+1) \dots f_i(t)\}$. In words, demes are colonized when the wave front first reaches them, and the subsequent evolution depends only on the allele frequencies at the time when they first evolved. From this construction, it follows that for $i < j$, $\{X_t^{(i)}\}$ and $\{X_t^{(j)}\}$ are conditionally independent given $f_i(i)$. Together with the Markov property this implies that the difference in allele frequency in two demes is a function of distance, i.e. they obey

$$F(X_t^{(i)}, X_t^{(j)} | f_i(i)) = F(X_{t-i}, \tilde{X}_{t-i} | f_0). \tag{4.8}$$

Throughout this section, we assume that $\mathbb{E}X_t | X_0 = f_0$ is constant, which is satisfied if there are no new mutations and no selection, and we further assume that $\mathrm{Var}(X_t) < \infty$. For example, for the critical branching process model we introduce in the following section, $\mathrm{Var}(X_t) = \sigma t$, where σ is the offspring variance in one generation. Then the autocovariance for $s < t$ is,

$$\mathrm{Cov}(X_s, X_t) = \mathrm{Var}(X_s), \tag{4.9}$$

and similarly for $\tilde{X}$, because $\{X_t\}, \{\tilde{X}_t\}$ are martingales.

Next, we define the conditioned processes $\{Y_t\} = \{X_t | X_t > 0\}$ and $\{\tilde{Y}\} = \{\tilde{X}_t | \tilde{X}_t > 0\}$ which give the allele frequency conditional on the allele not being lost.

Then, we have that

$$\mathbb{E}Y_t = \frac{\mathbb{E}X_t}{\mathbb{P}(X_t > 0)} = \frac{\mathbb{E}X_t}{1 - L(t)} \tag{4.10}$$

since

$$\mathbb{E}X = \mathbb{E}(X|X > 0)\mathbb{P}(X > 0). \tag{4.11}$$

Here, $L(t) = \mathbb{P}(X_t = 0)$ denotes the probability that an allele is at frequency zero in generation t, and we remove the dependency of $L(t)$ from f_0 for notational convenience.

Using the conditional variance formula, we can compute the variance and autocovariance of $\{Y_t\}$:

$$\begin{aligned}
\mathrm{Var}(Y_t) &= \frac{\mathbb{E}(X_t^2)}{1 - L(t)} - \left(\frac{\mathbb{E}(X_t^2)}{1 - L(t)}\right)^2 \\
&= \frac{\mathrm{Var}(X_t)}{1 - L(t)} + L(t)(\mathbb{E}Y_t)^2 \qquad (4.12)
\end{aligned}$$

and covariance for $s < t$

$$\begin{aligned}
\mathrm{Cov}(Y_s, Y_t) &= \mathbb{E}Y_sY_t - \mathbb{E}Y_s\mathbb{E}Y_t \\
&= \mathbb{E}(X_sX_t | X_t > 0) - \mathbb{E}(X_s|X_s > 0)\mathbb{E}(X_t|X_t > 0) \\
&= \frac{\mathbb{E}(X_sX_t)}{\mathbb{P}(X_t > 0)} - \frac{f_0^2}{\mathbb{P}(X_s > 0)\mathbb{P}(X_t > 0)} \\
&= \frac{\mathrm{Var}(X_s)}{1 - L(t)} + L(s)\frac{f_0^2}{(1 - L(s))(1 - L(t))}. \qquad (4.13)
\end{aligned}$$

The last quantity of interest is the difference $Z_t = Y_t - \tilde{Y}_t$, which gives the difference in allele frequency between the wavefront and the origin of the expansion, conditional on an allele surviving in both locations. We find that

$$\mathbb{E}Z_t = f_0\left(\frac{1}{1 - L(t)} - \frac{1}{1 - \tilde{L}(T)}\right), \tag{4.14}$$

$$\begin{aligned}
\mathrm{Var}(Z_t) &= \mathrm{Var}(Y_t) + \mathrm{Var}(\tilde{Y}_t) \\
&= \frac{\mathrm{Var}(X_t)}{1 - L(t)} + \frac{\mathrm{Var}(\tilde{X}_t)}{1 - \tilde{L}(t)} + L(t)\mathbb{E}Y_t + \tilde{L}(t)\mathbb{E}\tilde{Y}_t, \qquad (4.15)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Cov}(Z_s, Z_t) &= \mathrm{Cov}(Y_s, Y_t) + \mathrm{Cov}(\tilde{Y}_s, \tilde{Y}_t) - \mathrm{Cov}(Y_s, \tilde{Y}_t) - \mathrm{Cov}(\tilde{Y}_s, Y_t) \\
&= \frac{\mathrm{Var}(X_s)\,(1 - L(s)) + f_0^2 L(s)}{(1 - L(s))(1 - L(t))} \\
&+ \frac{\mathrm{Var}(\tilde{X}_s)\left(1 - \tilde{L}(s)\right) + f_0^2\tilde{L}(s)}{(1 - \tilde{L}(s))(1 - \tilde{L}(t))}. \qquad (4.16)
\end{aligned}$$

4.5.2 Branching Process

To further specify the moments derived in Section 4.5.1, we need to define $\mathrm{Var}(X_s)$, $L(s)$ and f_0, and the corresponding quantities at the wave front. This is particularly easy using a Galton-Watson branching process. Under this model, each generation individuals leave offspring independent from each other according to some offspring distribution F. Let $L_i(t)$ denote the probability that an allele has been lost by generation t, given that it started with i copies in generation 0. Kolmogorov (1938) showed that when t is large, L_1 is well approximated by

$$L_1(t) \approx 1 - \frac{2}{t\mathrm{Var}(F)}, \tag{4.17}$$

where F is the offspring distribution and $\mathrm{Var}(F)$ is assumed to be finite. We assume that a branching process with offspring distribution F describes neutral genetic drift at the wave front, and that the colonization of new demes occurs according to a branching process with offspring distribution $\tilde{F}$.

If the initial frequency f_0 of the allele is greater than one, the corresponding expression becomes

$$L_{f_0}(t) = (L_1(t))^{f_0}, \tag{4.18}$$

by independence of individuals. Plugging this into on Equation 4.1 and using a Taylor expansion around $t = \infty$ yields

$$\begin{aligned}\mathbb{E}Z_t &= f_0\left(\frac{1}{1-\tilde{L}_{f_0}(T)} - \frac{1}{1-L_{f_0}(t)}\right)\\ &= \frac{1}{2}\left(\mathrm{Var}(\tilde{F}) - \mathrm{Var}(F)\right)t - \frac{(f_0^2-1)\left(\mathrm{Var}(\tilde{F}) - \mathrm{Var}(F)\right)}{6\mathrm{Var}(F)\mathrm{Var}(\tilde{F})}\frac{1}{t} + o\left(\frac{1}{t^2}\right), \end{aligned}\tag{4.19}$$

Thus, we find that the expected difference in allele frequency between the expansion origin and the front of the population increases approximately linear with distance, the slope of the curve being the difference in offspring variance of individuals at the wavefront and expansion origin. From the second term in the Taylor expansion we see that the approximation is suitable when $t > f_0^2$, i.e. the number of demes between the two samples is large, and the frequency of the allele at the founding location is small.

4.5.3 Effective Population Size

The variance effective population size for a Cannings model is defined as

$$N_e = \frac{N}{\mathrm{Var}(F)}, \tag{4.20}$$

where N is the absolute number of individuals per population. The branching process considered above is not a Cannings model, however, the evolution of the offspring of a single individual under a Cannings population is well approximated by a branching process, as long as that offspring only makes up a small fraction of the individuals in a population. Fisher(1930) pioneered the modelling of population genetics using branching processes (Ewens, 2004, p.29). Under a Wright-Fisher model, the offspring distribution of a single individual has mean and variance very close to one. This justified Fisher to approximate the evolution of an individual under the Wright-Fisher model as evolving according to a branching process with a Poisson(1) offspring distribution, which has offspring variance 1, a model we will also use here to model genetic drift away from the wave front.

To incorporate the reduced effective size of a founder effect at the wave front, we use a modified offspring model: with probability $(1-\alpha)$, an individual at the wavefront does not produce any offspring. With probability α, the number of offspring is Poisson distributed with parameter $1/\alpha$ s.t. the overall expected number of offspring is still one and the variance is $\text{Var}(\tilde{F}) = \alpha^{-1}$. This allows us to define an effective founder size k_e

$$k_e = \alpha N, \tag{4.21}$$

which measures the "increase" in genetic drift at the wave front.

Combining Equation 4.21 and Equation 4.19 yields

$$\mathbb{E}Z_t = \frac{1}{2}\left(\frac{N_e}{k_e} - 1\right) t. \tag{4.22}$$

From this, we see immediately that $\mathbb{E}Z_t = 0$ only if $N_e = k_e$, and also that the effective founder size enters the equation only in the ratio $\kappa = \frac{k_e}{N}$, so that it makes sense to further define the relative founder size κ, which measures the strength of the founder effect.

4.5.4 Rescaling

The branching process we used above assume that exactly one generation of genetic drift happens between each founder event. In this section, we show that the expected allele frequency difference between the expansion front and at the origin is (i) invariant of additional generations between expansion events and (ii) invariant to additional generations after the expansion finished.

Both follow from the fact that for a branching process with mean 1, the variances of subsequent generations can simply be added: Consider the generating function of a critical branching process B after t generations, denoted by $p_t(s)$ which has variance $p_t(1)''$. Then, after an additional generation, the generating function becomes $q(p_t(s))$, where $q(s)$ is the generating function of the offspring distribution of that additional generation. Then, the variance in offspring after this additional generation is $q(p_t(1))''$.

$$\text{Var}(B) = (p_t'(1)^2 q''(p_t(1)) + q'(p_t(1))p_t''(1) = q''(1) + p_t''(1), \tag{4.23}$$

since $p_t'(1) = q'(1) = p_t(1) = q(1) = 1$.

Thus, if individuals in the range expansion model have offspring variance v at the expansion front and variance $\tilde{v}$ away from the front, the total variance after t time steps with d expansion events is $(t-d)v + d\tilde{v}$.

Now from Equation (4.19) we have (for $f_0 = 1$),

$$\begin{aligned}\mathbb{E}Z_d &= \frac{1}{2}\left[\mathrm{Var}(X_d) - \mathrm{Var}(\tilde{X}_d)\right] d \\ &= \frac{1}{2}\left[(dv) - (d\tilde{v})\right].\end{aligned}$$

Adding T generations with neutral drift between each founder event and τ generations after the expansion stopped, changes this only to

$$\mathbb{E}Z_d = \frac{1}{2}\left[(dv + (d-1)Tv + \tau v) - (d\tilde{v} + (d-1)Tv + \tau v)\right]$$

which simplifies to Equation (4.19).

We can model more complex expansion models, such as an extended bottleneck or logistic growth similarly. Again, this will result in an increase of $Var(X_d)$ and $Var(\tilde{X}_d)$ by the same amount, which cancels in the difference.

Furthermore, we can also change how we subdivide a population into demes. It is easy to see that a population with expansions at times $0, 1, 2, \ldots$ and offspring variances $\mathrm{Var}(F)$ and $\mathrm{Var}(\tilde{F})$ behaves similarly to a population with expansions occurring at times $0, \delta t, 2\delta t, \ldots$ with offspring variances $\frac{\mathrm{Var}(F)}{\delta}$ and $\frac{\mathrm{Var}(\tilde{F})}{\delta}$ in the sense that $\mathbb{E}Z_t$ will be the same for either population. This suggests that it is not important how we subdivide space into demes, only the relative size of the founder population versus the neutral populations matters. Thus, it is most convenient to report the strength of the founder effect in units of "decrease" in genetic diversity per unit of distance.

4.5.5 Estimation

To estimate $\mathbb{E}Z_t$ from genetic data, we need to take subsampling into account, i.e. we need to estimate $\mathbb{E}\hat{Z}_t = \mathbb{E}\hat{\tilde{Y}}_t - \mathbb{E}\hat{Y}_t$. In particular, the probability that an allele got lost from a population is not the same as it being absent from a sample. To model subsampling, we assume we start with f_0 copies of the derived allele and A_0 copies of the ancestral allele, all evolving as a independent branching processes. The expected number of ancestral alleles will be $\mathbb{E}A_t = A_0$ in all generations, whereas the expected number of the derived allele, conditioned on it not being lost, is $\mathbb{E}Y_t$. We Hence, in generation t, the probability of drawing m copies of the derived allele out of n samples is approximately binomially distributed with parameters n and $\frac{\mathbb{E}Y_t}{\mathbb{E}Y_t + A_0}$. The mean of the expected allele frequency, conditional on sampling at least one derived allele is

$$\mathbb{E}\hat{Y}_t / \frac{n}{2N} = \frac{n\mathbb{E}Y_t}{\mathbb{E}Y_t + A_0} \Bigg/ \left(\frac{\mathbb{E}Y_t}{\mathbb{E}Y_t + A_0}\right)^n = \frac{n\mathbb{E}Y_t(A_0 + \mathbb{E}Y_t)^{n-1}}{A_0^n}, \qquad (4.24)$$

with the $\frac{n}{2N}$ term normalizing $\hat{Y}_t$ to allele counts. Setting $A_0 \approx 2N - \mathbb{E}Y_t$ we obtain the series representation

$$\mathbb{E}\hat{Y}_t = \mathbb{E}Y_t + \frac{n}{2N}(\mathbb{E}Y_t)^2 + o\left(\frac{1}{N^2}\right). \qquad (4.25)$$

Hence,

$$\mathbb{E}\hat{Z}_t = \mathbb{E}\hat{\tilde{Y}}_t - \mathbb{E}\hat{Y}_t = \mathbb{E}\tilde{Y}_t - \mathbb{E}Y_t + \frac{n}{2N}\left((\mathbb{E}\tilde{Y}_t)^2 - (\mathbb{E}Y_t)^2\right) + o\left(\frac{1}{N^2}\right), \tag{4.26}$$

and we see that we have a bias term that increases with sample size. Hence, the easiest way to proceed is to downsample larger samples to a sample size of two, the case that is arguably most important in light of genomic data.

To compare samples of size n_1 and n_2, from a site frequency spectrum $\mathbf{S} = f_{ij}, 0 \leq i \leq n_1, 0 \leq j \leq n_2$ we can calculate a reduced site frequency spectrum matrix S' from the full site frequency spectrum using

$$\mathbf{S'} = \mathbf{P_1 S P_2^T}, \tag{4.27}$$

where $\mathbf{P_1}$ and $\mathbf{P_2}$ are $(2+1) \times (n_1+1)$ and $(2+1) \times (n_2+1)$ matrices (with indices starting at 0), respectively, with entries

$$p_{ji} = \frac{\binom{2}{j}\binom{n_1-2}{i-j}}{\binom{n_1}{i}} \tag{4.28}$$

for $0 \leq i \leq n_1$ and $0 \leq j \leq 2$ for $\mathbf{P_1}$. Entries in $\mathbf{P_2}$, are similar, except n_1 is replaced by n_2.

If we denote the entries of $\mathbf{S'}$ with s_{ij}, we can write $\mathbb{E}\hat{Z}_t$ as

$$\mathbb{E}\hat{Z}_t = \frac{s_{12} - s_{21}}{s_{12} + s_{21} + s_{11}}. \tag{4.29}$$

This statistic is identical to the ψ statistic defined in Peter and Slatkin (2013), where we did not give any theoretical justification.

4.6 Table

region	longitude	latitude	q	r_1	r_{10}	r_{100}	x_e	r^2	p
Scandinavia	24.16	60.32	0.00065	0.99869	0.9871	0.884	7.75	0.252	4.2e-55
USA	-78.63	44.22	0.00118	0.99763	0.9768	0.808	4.26	0.242	6.8e-06
Central Europe	11.40	46.84	0.00035	0.99928	0.9928	0.932	14.05	0.171	8.3e-21
Western Europe	2.47	43.33	0.00013	0.99973	0.9973	0.974	38.60	0.264	4.7e-50
Eastern Range	48.29	46.92	0.00028	0.99943	0.9943	0.946	17.80	0.115	3.6e-22

Table 4.1. Analysis of *A. thaliana* data.

The table shows the inferred latitude and longitude of the origin. q: regression slope in km^{-1}, $r_i = k_e/N_e$, for demes of size ikm. d_i, distance (in km) over which $1 - k_e/N_e = 1\%$. r^2 and p: adjusted coefficient of determination and Bonferroni-corrected p-value.

4.7 Figures

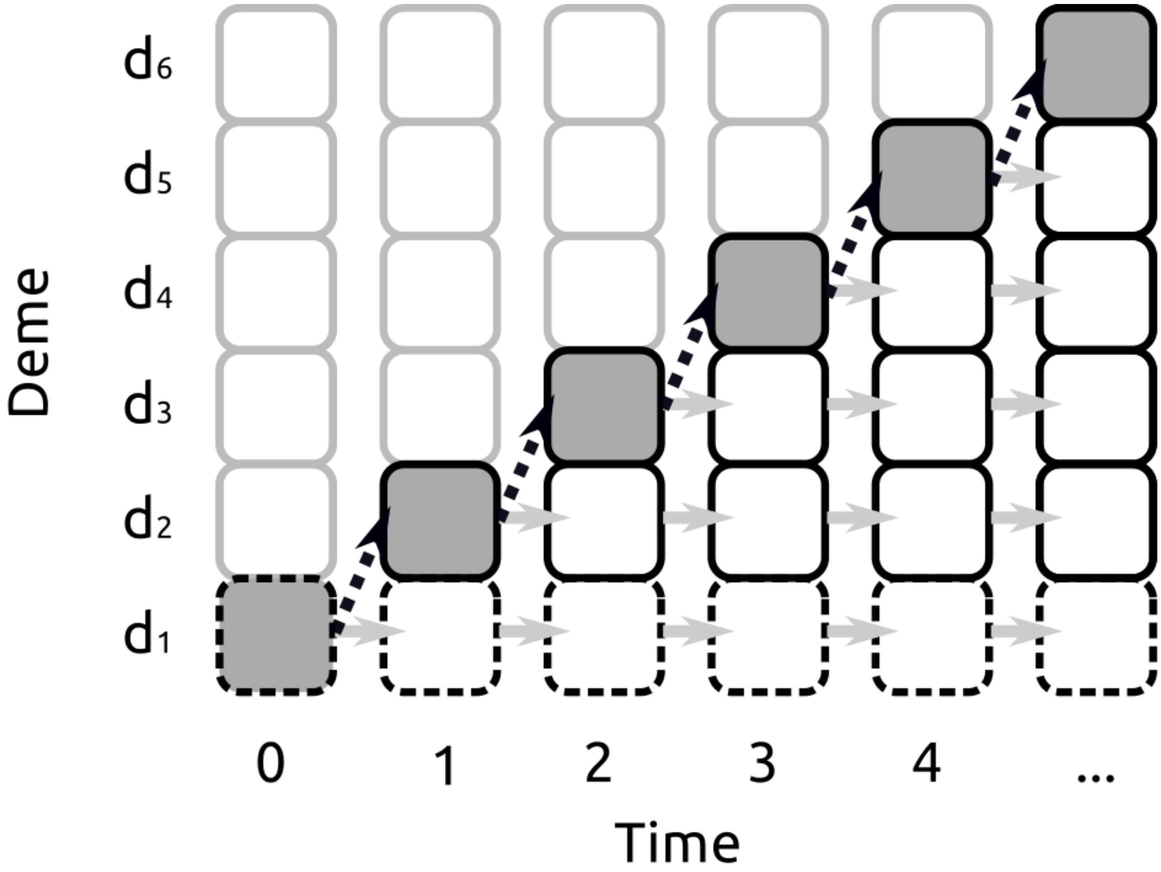

Figure 4.1. Schematic of the expansion models studied.
This figure shows the basic process we study, each square corresponds to a subpopulation, with grey borders indicating subpopulations not colonized at a time step. Each time step, a new deme is colonized (black, dashed arrows), and other demes undergo neutral genetic drift (grey arrows). We compare the allele frequencies $\{X_t\}$ at the expansion origin d_1 (dashed borderes) with the allele frequencies $\{\tilde{X}_t\}$ at the expansion front (dark backgrounds).

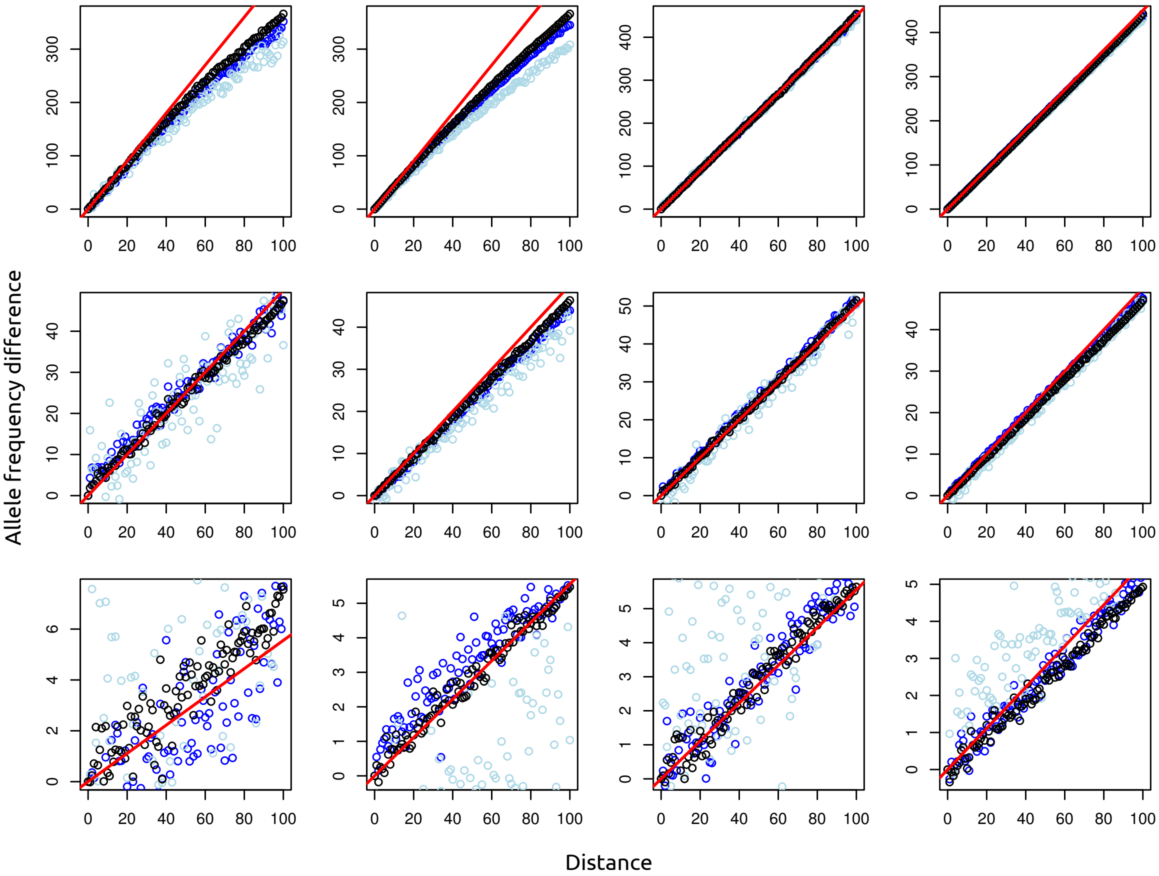

Figure 4.2. This figure shows the expected allele frequency difference between demes compared with simulations.
Top row: $k = 0.1N$, middle row: $k = 0.5N$, bottom row: $k = 0.9N$. first column: $f_0 = 1, N = 1000$, second column: $f_0 = 10, N = 1000$, third column: $f_0 = 10, N = 10000$, Fourth column: $f_0 = 100, N = 10000$, red line: prediction using branching process model. black, blue and lightblue dots correspond to samples right after expansion reached deme 100, 100 generations later and 500 generations later, respectively. Other parameters are $t = 2$, $m = 0$ and 10^6 alleles were generated

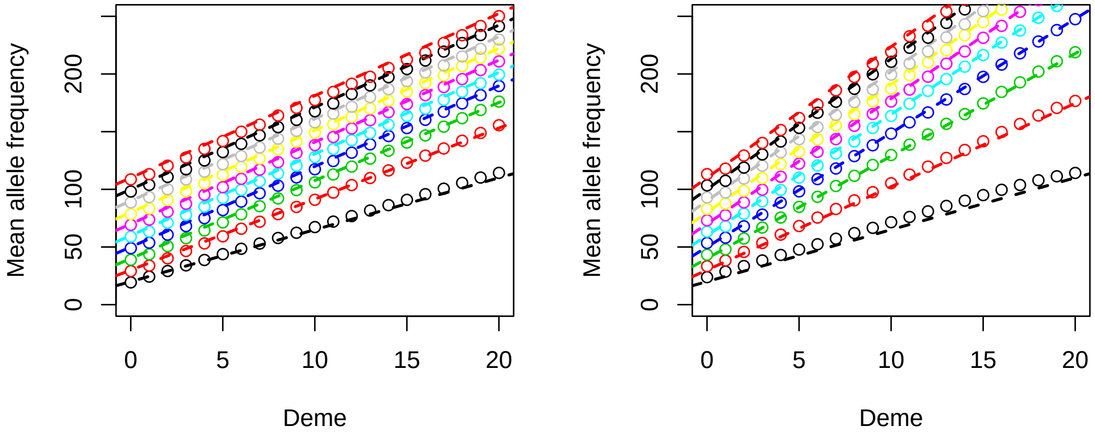

Figure 4.3. Logistic growth

Comparison between WF-simulations and predictions from the branching process model under a logistic growth model. Growth rates were set to 1 (Panel a) and 0.5 (Panel b), respectively the lines correspond to 1-10 generations of logistic growth per expansion step (from bottom to top). Dots correspond to the simulated data, and the dashed lines are the analytical predictions using the harmonic mean of the population sizes.

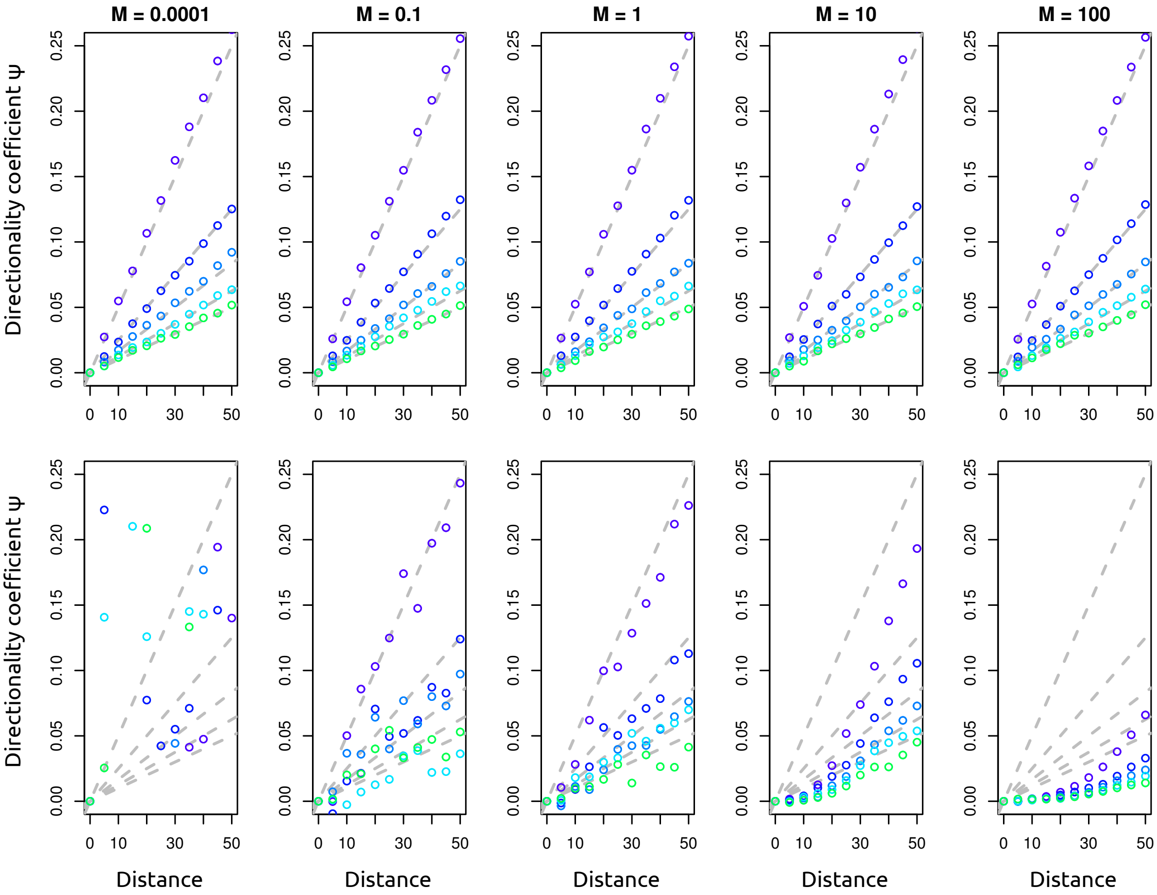

Figure 4.4. Effect of migration rate and subsampling.
Each set of points corresponds to ψ estimated from simulations under a specific k_e value, k_e varies from 100 to 500 in increments of 100 (top to bottom/ blue to green). Grey dashed lines give the expectation from the branching process model. Top row: data sampled immediately after the expansion finished. Bottom row: data sampled a very long time (100 coalescence units) after the expansion finished. Other parameters are: sample size $n = 10$, time between expansion events $t_e = 0.0001$ (in coalescence units).

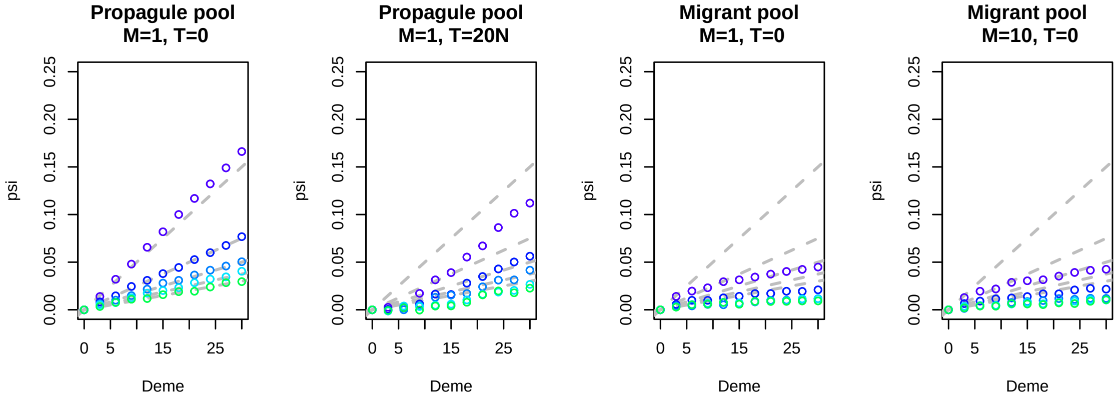

Figure 4.5. Effect of a 2D-geography.
Each set of points corresponds to ψ estimated from simulations under a specific k_e value, k_e varies from 100 to 500 in increments of 100 (top to bottom/ blue to green). Grey dashed lines give the expectation from the branching process model in one dimension.

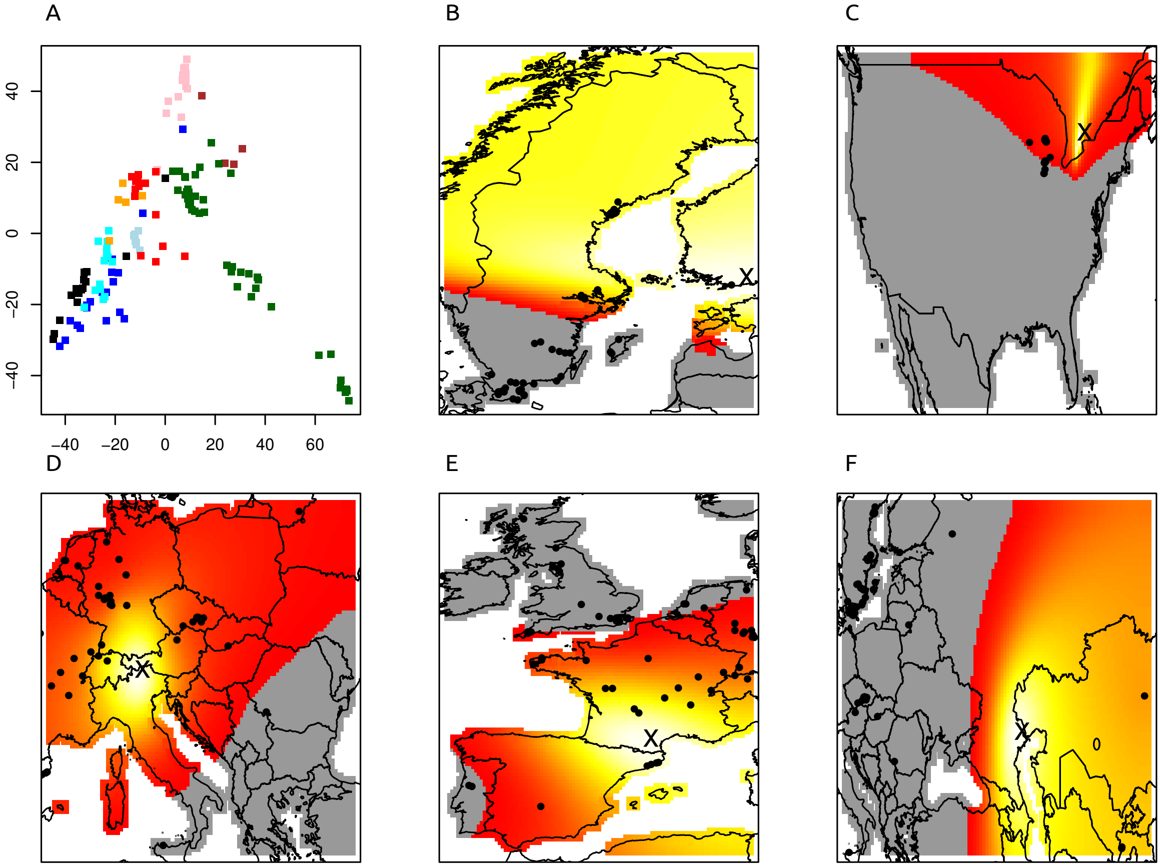

Figure 4.6. Results for *A. thaliana* data set.
Panel a: PCA analysis of the 121,412 SNP. Colors: Green: Scandinavia. Black: Americas. Blue: UK. Cyan: France. Light blue: Span & Portugal. Red: North-Western Europe. Orange: Switzerland & Italy. Pink: Central Europe. Brown: Russia, Lithuania and Western Asia. Panels b-f: Expansion for Scandinavia, USA, Central Europe, Western Europe and British Isles, respectively. Brighter regions indicate more likely origin of expansion.

Chapter 5

Conclusion

While the preceding three chapters address two quite distinct problems, there are several crucial aspects that all projects have in common. First, all projects deal with inference from non-equilibrium models. There are many theoretical results based on equilibrium assumptions (e.g. Durrett, 2008) and the underlying models are comparatively well explored. This is, of course, not because equilibria are biologically common, rather it is their mathematical convenience that is desirable. Many populations of ecological or evolutionary interest will not be in equilibrium, requiring new approaches. In this thesis, I explore three different approaches for the development of inference methods for models with little analytical theory.

In the second chapter, where I develop a method to distinguish selection from standing variation from selection on a *de novo* mutation, I use an Approximate Bayesian Computation (ABC) approach. While we can use ABC to do inference under any model, the statistical properties of the resulting inference method are largely unknown, and we have no guarantees for important quantities, such as the robustness to model violations. This is another common theme between all three chapters, and in all of them I need to explore the properties of the inference methods by performing simulations under scenarios where the model assumptions are violated. For the ABC procedure I apply in Chapter 2, I can go further by using the fact that accuracy is reduced due to the two approximation steps present in ABC. In the first step, I make the approximation that I can substitute the probability of the data given the full model, $\mathbb{P}(\text{data}|\text{model})$, with the probability of some summary statistics from the model. While this makes inference feasible, it nevertheless reduces the confidence in the model choice and parameter inference procedure. The other approximation is the substitution of an exact match of the summary statistics of the observations to the summary statistics of the simulated data with an approximate match. This again introduces some error. If this error is small, this likely implies that the model fit well, however, if theses errors are intermediate or large, we have to be very carful with any interpretation. One important issue is that ABC will always be able to compute a posterior distribution and model choice probabilities, so that it is crucial to check how well the models fit. If they do not, results have to be interpreted with extreme caution.

An example for this is given by our analysis of the G6PD gene, where it can be seen that the genetic patterns can not always be replicated by the simulations, which thus precludes any inference.

Chapters 3 and 4 both address questions pertaining to range expansions, yet the approach I use is quite different between the two chapters. In the first part of Chapter 3, inference is based on a symmetry argument; I try to reject a symmetrical pattern in joint allele frequencies that is expected only if the populations are at a migration equilibrium. Subsequent analyses, inferring the origin of the expansion and the migration history, are primarily based on empirical pattern, without any theory to back them up. However, the diversity gradient correlated with range expansions has been observed in many species (Ramachandran et al., 2005; Taberlet et al., 1998), and is predicted by all theoretical models of range expansions (Austerlitz et al., 1997; Hallatschek and Nelson, 2008; Slatkin and Excoffier, 2012).

Finally, in Chapter 4 I tackle the range expansion model by radically simplifying it, making strong assumptions that are biologically rather unrealistic. In particular, by using a branching process framework, I use a different set of assumptions than that of the commonly used Wright-Fisher model; instead of assuming a predetermined population size function, the population size is changing randomly throughout the life of each allele. This of course rises the question how meaningful the resulting model is for biological applications, and again the easiest way to address this question is by using simulations, which I use to show that our results are meaningful even if the assumptions I am making are violated.

Contrasting the approaches, in Chapter 2 I deal with the difficulties of inference by focusing on inference and model choice between few very specific models, that are necessarily explicitly parametrized and designed to encompass all important patterns of the underlying history of the samples. With the ABC approach I am unable to exactly explain how the inference procedure exactly works, I do not know which details of the data allow the distinction between the SDN and SSV models, but I can assess that, in some situation, this distinction can be made with reasonable certainty. Chapter 3 is motivated by the empirical observation of the diversity gradient that results from an expanding population. This gradient is exploited – first to test whether a range expansion happened, then to infer its origin. Finally, in Chapter 4 I use a bottom-up approach by starting with a very simple, generative model. While the explicit assumptions of this models are almost always violated, I may hope that the model's simplicity and focus on very basic premises adds some robustness, but the extent of this again has to be assessed using computer simulations. The biggest drawback of this approach is that it is much harder to extend. Scenarios that show big differences to the simple model cannot be explored, however, other extensions, for example the inclusion of variable population sizes, are relatively easy to include. As such, I arrive at a model for which a very good understanding of its easiest forms can be developed, but I again need to rely on simulations and power analyses to test its robustness and its statistical properties in more complex cases.

All three chapters also include a real data application. In Chapters 2 and 3 I use human data to illustrate the utility of our approach, and the same is done in Chapter 4 with a data set from *Arabidopsis*. These applications are crucial: They illustrate that the modeling assumptions I make are biological sensible; that sensible results can be obtained using these particular models as the basis for inference. They also advance our knowledge of the biology of these species, while the data sets I am using had previously been used for similar analyses, my novel approaches give a new perspective. For the gene investigated in Chapter 2, I identify allele ages and selection coefficients that are largely concordant to previous analyses using similar methods. In Chapter 3, I identify a Southern African origin for the expansion, which is concordant with some studies (Henn et al., 2011), but disagrees with others (Ramachandran et al., 2005). Finally, for the *Arabidopsis* application in Chapter 4, I confirm the previously identified pattern of an east-to-west colonization (François et al., 2008; Horton et al., 2012; Nordborg et al., 2005), but I also identify a discordant pattern in the Iberian peninsula. Thus, in all chapters, I find some agreement with previous estimates, which increases our confidence in the methods. But I am also able to make a novel contribution to the knowledge of the evolutionary history of the two species we study.

Hopefully the biological importance increases in the future. In ongoing collaborations, I am working on applying modifications of the selection on standing variation methodology to the human genes CPT1A and IFNL4, which are thought to cause cold resistance and Hepatitis C resistance, respectively. Using the range expansion tools I developed in Chapters 3 and 4, I am investigating the colonization history of the Ecuadoran Andes by *Mus musculus*. These projects should further prove the usefulness of the methods described here, and provide further examples on how new and innovative inference procedures help us solve biological problems in complex scenarios.

Bibliography

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature*. 467:1061–1073.

Achaz G. 2009. Frequency spectrum neutrality tests: One for all and all for one. *Genetics*. 183:249 –258.

Aho AV, Garey MR, Ullman JD. 1972. The transitive reduction of a directed graph. *SIAM Journal on Computing*. 1:131–137.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biology*. 2:e286. PMID: 15361935.

Akey JM, Swanson WJ, Madeoy J, Eberle M, Shriver MD. 2006. TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. *Human Molecular Genetics*. 15:2106 –2113.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, De Bakker PI, Deloukas P, Gabriel SB. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*. 467:52.

Austerlitz F, Jung-Muller B, Godelle B, Gouyon PH. 1997. Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*. 51:148–164.

Balakrishnan V, Sanghvi LD. 1968. Distance between populations on the basis of attribute data. *Biometrics*. 24:859–865. ArticleType: research-article / Full publication date: Dec., 1968 / Copyright © 1968 International Biometric Society.

Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. *Trends in Ecology & Evolution*. 23:38–44.

Barton NH, Etheridge AM, Kelleher J, Véber A. 2013. Genetic hitchhiking in spatially extended populations. *Theoretical Population Biology*. 87:75–89.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate bayesian computation in population genetics. *Genetics*. 162:2025–2035.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*. 74:1111–1120. PMID: 15114531 PMCID: 1182075.

Beutler E. 1994. G6PD deficiency. *Blood*. 84:3613 –3636.

Bhatia G, Patterson N, Pasaniuc B, et al. (40 co-authors). 2011. Genome-wide comparison of African-Ancestry populations from CARe and other cohorts reveals signals of natural selection. *The American Journal of Human Genetics*. 89:368–381.

Birnbaumer L, Yidirim E, Abramowitz J. 2003. A comparison of the genes coding for canonical TRP channels and their m, v and p relatives. *Cell Calcium*. 33:419–432.

Blum MGB, François O. 2009. Non-linear regression models for approximate bayesian computation. *Statistics and Computing*. 20:63–73.

Bond J, Roberts E, Mochida GH, et al. (12 co-authors). 2002. ASPM is a major determinant of cerebral cortical size. *Nature Genetics*. 32:316–320. PMID: 12355089.

Boulesteix A, Strimmer K. 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*. 8:32–44. PMID: 16772269.

Box G, Cox D. 1964. An analysis of transformations. *JR Stat. Soc., Ser. B*. 26:211–243.

Brockmann D, Helbing D. 2013. The hidden geometry of complex, network-driven contagion phenomena. *Science*. 342:1337–1342.

Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, Myles S. 2008. Positive selection in east asians for an EDAR allele that enhances NF-κB activation. *PLoS ONE*. 3. PMID: 18493316 PMCID: 2374902.

Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG. 2007. Absence of the Lactase-Persistence-Associated allele in early neolithic europeans. *Proceedings of the National Academy of Sciences*. 104:3736–3741.

Busing FMTA, Meijer E, Leeden RVD. 1999. Delete-m jackknife for unequal m. *Statistics and Computing*. 9:3–8.

Cann HM, De Toma C, Cazes L, et al. (41 co-authors). 2002. A human genome diversity cell line panel. *Science*. 296:261–262.

Carson PE, Flanagan CL, Ickes CE, Alving AS. 1956. Enzymatic deficiency in Primaquine-Sensitive erythrocytes. *Science*. 124:484 –485.

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: Models and estimation procedures. *Evolution*. 21:550–570. ArticleType: research-article / Full publication date: Sep., 1967 / Copyright © 1967 Society for the Study of Evolution.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton university press.

Cavalli-Sforza LLLL, Menozzi P, Piazza A. 1996. The History and Geography of Human Genes: (Abridged Paperback Edition). University Press.

Cook SR, Gelman A, Rubin DB. 2006. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*. 15:675–692.

Corander J, Waldmann P, Marttinen P, Sillanpää MJ. 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*. 20:2363–2369.

Cornuet J, Santos F, Beaumont MA, Robert CP, Marin J, Balding DJ, Guillemaud T, Estoup A. 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate bayesian computation. *Bioinformatics*. 24:2713 –2719.

Cox JT, Durrett R. 2002. The stepping stone model: New formulas expose old myths. *The Annals of Applied Probability*. 12:1348–1377. ArticleType: research-article / Full publication date: Nov., 2002 / Copyright © 2002 Institute of Mathematical Statistics.

Csilléry K, Blum MGB, Gaggiotti OE, François O. 2010. Approximate bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*. 25:410–418. PMID: 20488578.

Currat M, Excoffier L, Maddison W, Otto SP, Ray N, Whitlock MC, Yeaman S. 2006. Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science*. 313:172.

Davis MA. 2009. Invasion Biology. Oxford ; New York: Oxford University Press, 1 edition edition.

DeGiorgio M, Degnan JH, Rosenberg NA. 2011. Coalescence-Time distributions in a serial founder model of human evolutionary history. *Genetics*. 189:579–593. PMID: 21775469.

DeGiorgio M, Jakobsson M, Rosenberg NA. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from africa. *Proceedings of the National Academy of Sciences*. 106:16057–16062.

DeGiorgio M, Rosenberg NA. 2012. Geographic sampling scheme as a determinant of the major axis of genetic variation in principal components analysis. *Molecular Biology and Evolution*. .

Didelot X, Everitt RG, Johansen AM, Lawson DJ. 2010. Likelihood-free estimation of model evidence. *Bayesian analysis*. 6:49–76.

Durrett R. 2008. Probability models for DNA sequence evolution. Springer.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences of the United States of America*. 101:975–979.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*. 30:233–237. PMID: 11788828.

Eng MY, Luczak SE, Wall TL. 2007. ALDH2, ADH1B, and ADH1C genotypes in asians: a literature review. *Alcohol Research & Health: The Journal of the National Institute on Alcohol Abuse and Alcoholism*. 30:22–27. PMID: 17718397.

Ewens WJ. 2004. Mathematical Population Genetics: Theoretical introduction. Springer.

Excoffier L. 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology*. 13:853–864.

Fagundes NJR, Ray N, Beaumont MA, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. *PNAS*. 104:17614–17619.

Fay JC, Wu C. 2000. Hitchhiking under positive darwinian selection. *Genetics*. 155:1405 –1413.

Fisher RA. 1937. The wave of advance of advantageous genes. *Annals of Eugenics*. 7:355–369.

François O, Blum MGB, Jakobsson M, Rosenberg NA. 2008. Demographic history of european populations of arabidopsis thaliana. *PLoS Genet*. 4:e1000075.

François O, Currat M, Ray N, Han E, Excoffier L, Novembre J. 2010. Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*. 27:1257–1268.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*. 147:915 –925.

Fujimoto A, Ohashi J, Nishida N, Miyagawa T, Morishita Y, Tsunoda T, Kimura R, Tokunaga K. 2008. A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in asia. *Human Genetics*. 124:179–185. PMID: 18704500.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*. 7:e1002355.

Goldstein DB, Ruiz A, Cavalli-Sforza LL, Feldman MW. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics*. 139:463–471.

Gravel S, Henn BM, Gutenkunst RN, et al. (560 co-authors). 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*. p. 201019276. PMID: 21730125.

Grossman SR, Shylakhter I, Karlsson EK, et al. (13 co-authors). 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*. 327:883 –886.

Guillot G, Leblois R, Coulon A, Frantz AC. 2009. Statistical methods in spatial genetics. *Molecular Ecology*. 18:4734–4756.

Gustafsson F, Gunnarsson F. 2003. Positioning using time-difference of arrival measurements. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. volume 6, p. VI–553.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 5:e1000695.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR. 2007. Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences*. 104:19926 –19930.

Hallatschek O, Nelson DR. 2008. Gene surfing in expanding populations. *Theoretical Population Biology*. 73:158–170.

Handley LL, Estoup A, Evans DM, Thomas CE, Lombaert E, Facon B, Aebi A, Roy HE. 2011. Ecological genetics of invasive alien species. *BioControl*. 56:409–428.

Harris TE. 1954. The theory of branching processes. Courier Dover Publications.

Henn BM, Gignoux CR, Jobin M, et al. (19 co-authors). 2011. Hunter-Gatherer genomic diversity suggests a southern african origin for modern humans. *Proceedings of the National Academy of Sciences*. .

Hermisson J, Pennings PS. 2005. Soft sweeps. *Genetics*. 169:2335–2352. PMID: 15716498 PMCID: 1449620.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Project G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science*. 331:920 –924.

Hewitt GM. 1999. Post-glacial re-colonization of european biota. *Biological Journal of the Linnean Society*. 68:87–112.

Hey J. 2010. Isolation with migration models for more than two populations. *Molecular Biology and Evolution*. 27:905–920.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *persimilis*. *Genetics*. 167:747–760.

Hofer T, Ray N, Wegmann D, Excoffier L. 2009. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Annals of Human Genetics*. 73:95–108.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM. 2001. Lactase haplotype diversity in the old world. *American Journal of Human Genetics*. 68:160–172. PMID: 11095994.

Horton MW, Hancock AM, Huang YS, et al. (13 co-authors). 2012. Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*. 44:212–216.

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics*. 120:831 –840.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics*. 116:153 –159.

Hurst LD. 2009. Genetics and the understanding of selection. *Nat Rev Genet*. 10:83–93.

Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences of the United States of America*. 101:10667 –10672.

Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG. 2009. The origins of lactase persistence in europe. *PLoS Comput Biol*. 5:e1000491.

Jobin MJ, Mountain JL. 2008. REJECTOR: software for population history inference from genetic data via a rejection algorithm. *Bioinformatics*. 24:2936–2937.

Jorgensen S, Mauricio R. 2004. Neutral genetic variation among wild north american populations of the weedy plant arabidopsis thaliana is not geographically structured. *Molecular Ecology*. 13:3403–3413.

Kaplan NL, Hudson RR, Langley CH. 1989. The “hitchhiking effect” revisited. *Genetics*. 123:887–899. PMID: 2612899 PMCID: 1203897.

Kimura M. 1964. Diffusion models in population genetics. *Journal of Applied Probability*. 1:177–232.

Kimura M. 1985. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Kimura M, Ohta T. 1973. The age of a neutral mutant persisting in a finite population. *Genetics*. 75:199–212.

Kimura R, Yamaguchi T, Takeda M, et al. (13 co-authors). 2009. A common variation in EDAR is a genetic determinant of Shovel-Shaped incisors. *Am J Hum Genet*. 85:528–535.

Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*. 23:482 –490.

Kolmogorov A, Petrovskii I, Piscounov N. 1937. A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem. *Univ., Math. Mech.* 1:1–25.

Korte BBH, Vygen J. 2008. Combinatorial Optimization: Theory and Algorithms. Springer London, Limited.

Kouprina N, Pavlicek A, Mochida GH, et al. (13 co-authors). 2004. Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. *PLoS Biol*. 2:e126.

Kuokkanen M, Enattah NS, Oksanen A, Savilahti E, Orpana A, Järvelä I. 2003. Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. *Gut*. 52:647–652. PMID: 12692047 PMCID: 1773659.

Lê Cao K, González I, Déjean S. 2009. integrOmics: an r package to unravel relationships between two omics datasets. *Bioinformatics*. 25:2855 –2856.

Leuenberger C, Wegmann D. 2010. Bayesian computation and model selection without likelihoods. *Genetics*. 184:243 –252.

Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*. 74:175 –195.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. 27:718 –719.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*. .

Li H, Mukherjee N, Soundararajan U, et al. (11 co-authors). 2007. Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western asia. *American Journal of Human Genetics*. 81:842–846. PMID: 17847010.

Malécot G. 1950. Quelques schémas probabilistes sur la variabilité des populations naturelles. In: Annales de l'Université de Lyon A. volume 13, p. 37–60.

Malmström H, Linderholm A, Lidén K, Storå J, Molnar P, Holmlund G, Jakobsson M, Götherström A. 2010. High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern europe. *BMC Evolutionary Biology*. 10:89.

Marin J, Pillai N, Robert CP, Rousseau J. 2011. Relevant statistics for bayesian model choice. *arXiv:1110.4700*. .

Marjoram P, Molitor J, Plagnol V, Tavaré S. 2003. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*. 100:15324 –15328.

Maruyama T. 1974. The age of an allele in a finite population. *Genetics Research*. 23:137–143.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the adh locus in drosophila. *Nature*. 351:652–654.

McGovern PE, Zhang J, Tang J, et al. (13 co-authors). 2004. Fermented beverages of pre- and proto-historic china. *Proceedings of the National Academy of Sciences of the United States of America*. 101:17593 –17598.

Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, Hudson RR, Tishkoff SA, Lahn BT. 2005. Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens. *Science*. 309:1720 –1722.

Mekel-Bobrov N, Lahn BT. 2007. Response to comments by timpson et al. and yu et al. *Science*. 317:1036.

Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in europeans. *Science*. 201:786–792.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A Fine-Scale map of recombination rates and hotspots across the human genome. *Science*. 310:321 –324.

Nei M. 1972. Genetic distance between populations. *American Naturalist*. 106:283–&. WOS:A1972M475000002.

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics*. 158:885–896. PMID: 11404349.

Nielsen R, Williamson SH, Kim Y, Hubisz MJ, Clark AG, Bustamante CD. 2005. Genomic scans for selective sweeps using SNP data. *Genome research*. 15:1566–75.

Nkhoma ET, Poole C, Vannappagari V, Hall SA, Beutler E. 2009. The global prevalence of glucose-6-phosphate dehydrogenase deficiency: A systematic review and meta-analysis. *Blood Cells, Molecules, and Diseases*. 42:267–278.

Nordborg M, Hu TT, Ishino Y, et al. (24 co-authors). 2005. The pattern of polymorphism in arabidopsis thaliana. *PLoS Biol*. 3:e196.

Novembre J, Johnson T, Bryc K, et al. (12 co-authors). 2008. Genes mirror geography within europe. *Nature*. 456:98–101.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*. 23:263–286.

Olds LC, Sibley E. 2003. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Human Molecular Genetics*. 12:2333–2340. PMID: 12915462.

Osier MV, Pakstis AJ, Soodyall H, et al. (14 co-authors). 2002. A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *The American Journal of Human Genetics*. 71:84–99.

Peng Y, Shi H, Qi Xb, Xiao Cj, Zhong H, Ma Rl, Su B. 2010. The ADH1B Arg47His polymorphism in east asian populations and expansion of rice domestication in history. *BMC Evolutionary Biology*. 10:15.

Pennings PS, Hermisson J. 2006a. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genetics*. 2:e186. PMID: 17173482.

Pennings PS, Hermisson J. 2006b. Soft sweeps II—Molecular population genetics of adaptation from recurrent mutation or migration. *Molecular Biology and Evolution*. 23:1076 –1084.

Peter BM, Slatkin M. 2013. Detecting range expansions from genetic data. *Evolution*. 67:3274–3289.

Peter BM, Wegmann D, Excoffier L. 2010. Distinguishing between population bottleneck and population subdivision by a bayesian model choice procedure. *Molecular ecology*. 19:4648–4660. PMID: 20735743.

Plantinga TS, Alonso S, Izagirre N, Hervella M, Fregel R, Meer JWvd, Netea MG, R|[uacute]|a Cdl. 2012. Low prevalence of lactase persistence in neolithic South-West europe. *European Journal of Human Genetics*. .

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945–959.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution*. 59:2312–23.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proceedings of the National Academy of Sciences of the United States of America*. 102:15942–15947.

Rasmussen M, Guo X, Wang Y, et al. (58 co-authors). 2011. An aboriginal australian genome reveals separate human dispersals into asia. *Science*. 334:94–98.

Ray N, Currat M, Foll M, Excoffier L. 2010. SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics*. 26:2993–2994.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing indian population history. *Nature*. 461:489–494.

Reynolds J, Weir B, Cockerham C. 1983. Estimation of the Co-Ancestry coefficient - basis for a Short-Term genetic distance. *Genetics*. 105:767–779. WOS:A1983RN08900018.

Robert CP, Cornuet J, Marin J, Pillai N. 2011. Lack of confidence in ABC model choice. *1102.4432*. .

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet*. 1:e70.

Ruwende C, Khoo SC, Snow RW, et al. (14 co-authors). 1995. Natural selection of hemi- and heterozygotes for G6PD deficiency in africa by resistance to severe malaria. *Nature*. 376:246–249.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 419:832–837.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science*. 312:1614 –1620.

Sabeti PC, Varilly P, Fry B, et al. (12 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*. 449:913–918. PMID: 17943131 PMCID: 2687721.

Saunders MA, Hammer MF, Nachman MW. 2002. Nucleotide variability at g6pd and the signature of malarial selection in humans. *Genetics*. 162:1849–1861. PMID: 12524354.

Schmitt T. 2007. Molecular biogeography of europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*. 4:11. PMID: 17439649.

Simonsen K, Churchill G, Aquadro C. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*. 141:413.

Sisson SA, Fan Y, Tanaka MM. 2007. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*. 104:1760 –1765.

Slatkin M, Excoffier L. 2012. Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics*. 191:171–181. PMID: 22367031.

Slatkin M, Voelm L. 1991. FST in a hierarchical island model. *Genetics*. 127:627–629.

Slatkin M, Wade MJ. 1978. Group selection on a quantitative character. *Proceedings of the National Academy of Sciences of the United States of America*. 75:3531–3534. PMID: 16592546 PMCID: PMC392812.

Spencer CCA, Coop G. 2004. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*. 20:3673–3675.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*. 22:63 –73.

Taberlet P, Fumagalli L, Wust-Saucy A, Cosson J. 1998. Comparative phylogeography and postglacial colonization routes in europe. *Molecular Ecology*. 7:453–464.

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

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics*. 145:505–518. PMID: 9071603.

Tenenhaus M. 1998. La régression PLS: théorie et pratique. Editions TECHNIP.

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

Teshima KM, Innan H. 2009. mbs: modifying hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC bioinformatics*. 10:166.

Timpson N, Heron J, Smith GD, Enard W. 2007. Comment on papers by evans et al. and Mekel-Bobrov et al. on evidence for positive selection of MCPH1 and ASPM. *Science*. 317:1036.

Tishkoff SA, Reed FA, Friedlaender FR, et al. (25 co-authors). 2009. The genetic structure and history of africans and african americans. *Science*. 324:1035–1044.

Tishkoff SA, Reed FA, Ranciaro A, et al. (19 co-authors). 2007. Convergent adaptation of human lactase persistence in africa and europe. *Nature Genetics*. 39:31–40. PMID: 17159977.

Tishkoff SA, Varkonyi R, Cahinhinan N, et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*. 293:455 –462.

Troelsen JT, Olsen J, M?ller J, Sjöström H. 2003. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology*. 125:1686–1694. PMID: 14724821.

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA. 2002. Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *American Journal of Human Genetics*. 71:1112–1128. PMID: 12378426.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS biology*. 4:e72.

Wakeley J. 2009. Coalescent theory: an introduction. Roberts & Co. Publishers.

Wang C, Zöllner S, Rosenberg NA. 2012. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet*. 8:e1002886.

Waters MR, Stafford TW. 2007. Redefining the age of clovis: Implications for the peopling of the americas. *Science*. 315:1122–1126.

Wegmann D, Currat M, Excoffier L. 2006. Molecular diversity after a range expansion in heterogeneous environments. *Genetics*. 174:2009–2020.

Wegmann D, Leuenberger C, Excoffier L. 2009. Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics*. 182:1207–1218. PMID: 19506307 PMCID: 2728860.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. 2010. ABCtoolbox: a versatile toolkit for approximate bayesian computations. *BMC Bioinformatics*. 11:116.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the analysis of population structure. *Evolution*. 38:1358–1370. ArticleType: research-article / Full publication date: Nov., 1984 / Copyright © 1984 Society for the Study of Evolution.

Wilkins JF, Wakeley J. 2002. The coalescent in a continuous, finite, linear population. *Genetics*. 161:873–888.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *PNAS*. 102:7882–7887.

Wright S. 1943. Isolation by distance. *Genetics*. 28:114–138. PMID: 17247074 PMCID: PMC1209196.

Wright S. 1949. The genetical structure of populations. *Annals of Human Genetics*. 15:323–354.

Wu X, Ye Y, Kiemeney LA, et al. (48 co-authors). 2009. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat Genet*. 41:991–995.

Yu F, Hill RS, Schaffner SF, Sabeti PC, Wang ET, Mignault AA, Ferland RJ, Moyzis RK, Walsh CA, Reich D. 2007. Comment on ”Ongoing adaptive evolution of ASPM, a brain size determinant in homo sapiens”. *Science*. 316:370.

Zhang J. 2003. Evolution of the human ASPM gene, a major determinant of brain size. *Genetics*. 165:2063 –2070.