UNIVERSITY OF CALIFORNIA

Los Angeles

Addressing challenges for population genetic inference from next-generation sequencing

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Eun-Jung Han

2014

ABSTRACT OF THE DISSERTATION

Addressing challenges for population genetic inference from next-generation sequencing

by

Eun-Jung Han

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2014

Professor John Novembre, Co-chair

Professor Janet S. Sinsheimer, Co-chair

Next-generation sequencing (NGS) data provides tremendous opportunities for making new discoveries in biology and medicine. However, a structure of NGS data poses many inherent challenges - for example, reads have high error rates, read mapping is sometimes uncertain, and coverage is variable and in many cases low or completely absent. These challenges make accurate individual-level genotype calls difficult and make downstream analysis based on genotypes problematic if genotype uncertainty is not accounted for. In this dissertation, I present recent works addressing challenges that arise in the analysis of NGS data for population genetic inferences and and provide recommendations and guidelines to interpret such data with precision. Throughout this dissertation, I focus on estimating the site frequency spectrum (SFS). The distribution of allele frequencies across polymorphic sites, also known as the SFS, is of primary interest in population genetics. It is a complete summary of sequence variation at unlinked sites and more generally, its shape reflects underlying population genetic processes.

First, I characterize biases that can arise inferring the SFS from low- to medium-coverage sequencing data and present a statistical method that can ameliorate such biases. I compare two approaches to estimate the SFS from sequencing data: one approach infers individual genotypes from aligned sequencing reads and then estimates the SFS based on the inferred genotypes (call-based approach) and the other approach directly estimates the SFS from aligned sequencing reads by maximum likelihood (direct estimation approach). I find that the SFS estimated by the direct estimation approach is unbiased even at low coverage, whereas the SFS by the call-based approach becomes biased as coverage decreases. The direction of the bias in the call-based approach depends on the pipeline to infer genotypes. Estimating genotypes by pooling individuals in a sample (multisample calling) results in underestimation of the number of rare variants, whereas estimating genotypes in each individual and merging them later (single-sample calling) leads to overestimation of rare variants. I characterize the impact of these biases on downstream analyses, such as demographic parameter estimation and genome-wide selection scans. This work highlights that depending on the pipeline used to infer the SFS, one can reach different conclusions in population genetic inference with the same data set. Thus, careful attention to the analysis pipeline and SFS estimation procedures is vital for population genetic inferences.

Next, I describe a development of a novel algorithm that can speed-up the existing direct estimation method with the EM optimization. The existing method directly estimates the SFS from sequencing data by first computing site likelihood vectors (i.e. the likelihood a site has a each possible allele frequency conditional on observed sequence reads) using a dynamic programming (DP) algorithm. Although this method produces an accurate SFS, computing the site likelihood vector is quadratic in the number of samples sequenced. To overcome this computational challenge, I propose an algorithm we call the adaptive K-restricted algorithm, which is linear in the number of genomes to compute the site likelihood vector. This algorithm works because in a lower triangular matrix that arises in the DP algorithm, all non-negligible values of the site likelihood vector are concentrated on a few cells around the best- guess allele counts. I show that this adaptive K-restricted algorithm has comparable accuracy but is faster than the original DP algorithm. This speed improvement makes SFS estimation practical when using low coverage NGS data from a large number of individuals.

Finally, as an application, I analyze high-coverage sequencing data of two dogs and three wolves to detect genetic signatures of adaptation during early dog domestication. This work is part of a larger research effort, called the Canid Genome Project, where I take the lead in the selection scans. We identify the importance of dietary evolution in early dog domestication, supported by our top selection hit, a CCRN4L gene. Moreover, we observe that genes affecting brain function, metabolism, and morphology show signatures of selection in the dog lineage.

The dissertation of Eun-Jung Han is approved.

Steve Horvath

Donatello Telesca

Janet S. Sinsheimer, Committee Co-chair

John Novembre, Committee Co-chair

University of California, Los Angeles

2014

This dissertation is dedicated to my husband Wonho Park

and my son Sangwook Park for all of their love and support.

Table of Contents

List of Figures

ACKNOWLEDGMENTS

I would like to thank my advisors, Professor John Novembre and Professor Janet Sinsheimer, for their continuous guidance and patience throughout the course of my research and thesis. I also thank my committee members Steve Horvath and Donatello Telesca for their support and encouragement. Finally, I would like to thank my lab colleagues Alex Platt, Charleston Chiang, Darren Kessner, and Diego Ortega Del Vecchyo for helpful discussion and advice.

This work was supported by the National Institutes of Health (Training Grant in Genomic Analysis and Interpretation T32 HG002536 for EH, R01 HG007089 for JN).

Vita

2008-2009	MS, Biostatistics, UCLA, Los Angeles, California
2005-2008	PhD Program, Physiological Science, UCLA, Los Angeles, California
1999-2004	BS, Biotechnology, Yonsei University, Seoul, Korea
2012-2014	NIH Training Grant in Genomic Analysis and Interpretation, UCLA Human Genetics
2012-2014	UCLA Graduate Student Fellowship, UCLA Biostatistics
2009-2011	NIH Training Grant in AIDS Research, UCLA Biostatistics
2007-2008	NIH Training Grant in Physiological Science Research, UCLA Physiological Science

Publications

Han E, Sinsheimer JS, Novembre J. Fast and accurate site frequency spectrum estimation from low coverage sequence data. *In preparation*

Han E, Sinsheimer JS, Novembre J (2014). Characterizing bias in population genetic inferences from low coverage sequencing data. *Molecular Biology & Evolution.* 31 (3): 723-35.

Freedman AH, Gronau I, Schweizer RM, Vecchyo DO, **Han E**, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, Beale H, Ramirez O, Hormozdiari F, Alkan C, Vil C, Squire K, Geffen E, Kusak J, Boyko AR, Parker HG, Lee C, Tadigotla V, Siepel A, Bustamante CD, Harkins TT, Nelson SF, Ostrander EA, Marques-Bonet T, Wayne RK, Novembre (2014). Genome Sequencing Highlights the Dynamic Early History of Dogs. *PLOS Genetics.*

Novembre J, **Han E** (2012). Human population structure and the adaptive response to pathogen-induced selection pressures. *Phil. Trans. R. Soc. B.* 367 (1590): 878-86.

Vonholdt BM, Pollinger JP, Lohmueller KE, **Han E**, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A, Reynolds A, Bryc K, Brisbin A, Knowles JC, Mosher DS, Spady TC, Elkahloun A, Geffen E, Pilot M, Jedrzejewski W, Greco C, Randi E, Bannasch D, Wilton A, Shearman J, Musiani M, Cargill M, Jones PG, Qian Z, Huang W, Ding ZL, Zhang YP, Bustamante CD, Ostrander EA, Novembre J, Wayne RK (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature.* 464 (7290): 898-902.

Franois O, Currat M, Ray N, **Han E**, Excoffier L, Novembre J (2010). Principal component analysis under population genetic models of range expansion and admixture. Characterizing bias in population genetic inferences from low coverage sequencing data. *Molecular Biology & Evolution.* 27 (6): 1257-68.

CHAPTER 1

Introduction

1.1 Motivation

Next-generation sequencing (NGS) technologies over the past several years led to the tremendous growth in making new discoveries in biology and medicine. The ability to generate an enormous volume of genomic data at low cost makes it possible to perform large-scale population genetic studies that were unimaginable just a few years ago.

Despite its big promises, the structure of NGS data also poses many inherent challenges. First, NGS data suffers from high error rates at multiple stages, including base-calling errors on short reads and read mapping errors onto the available reference genome. These errors often mimic single-nucleotide polymorphisms (SNPs), leading a true homozygote to be misclassified as a heterozygote. Second, NGS data have variable depth of coverage because short sequence reads are randomly generated from the individual genome. For those sites with low depth of coverage ($<$ 5X per site per individual), there is a high probability that all sequence reads are sampled from only one of the two chromosomes in a diploid individual. This sampling variation may lead a true heterozygote to be miscalssified as a homozygote. For those sites with no sequencing data, we have missing data problems.

Under such circumstances, accurate individual-level genotype calling is problematic, and there is often considerable uncertainty associated with the genotype calls for low ($<$5X) to medium (5-20X)-coverage sequencing data. Uncertainty in genotype calls is an important consideration in population genetic studies, in which many inferences are based on summary statistics, such as allele frequencies or the distribution of the allele frequencies. Ignoring genotype call uncertainty can lead to biased estimates of those summary statistics and spurious conclusions in many population genetic analyses, such as demographic inference based on the frequency spectrum, empirical selection scans, the identification of rare mutations, and association mapping. These have been identified as key limitations for performing population genetic studies from NGS data.

One method for handling uncertainty associated with genotype calls in sequencing data is to sequence target regions at high coverage (>20X) to obtain more reliable genotypes (with more data, more confidence in genotype calls). However, cost constraints leads to difficult choices between increasing sequencing coverage and increasing sample sizes. With limited budgets, we expect a category of experimental work will continue in which it is most advantageous to maximize the number of individuals by using low coverage - for example, identification of low-frequency variants and association mapping.

Alternatively, reducing and quantifying the uncertainty associated with genotype calling can be accomplished using a probabilistic framework and this area has recently been the subject of extensive research. The key quantity used in the probabilistic methods is genotype likelihoods, which incorporate base-calling and alignment errors. Based on the genotype likelihoods in conjunction with the genotype prior, one can infer the individual-level genotypes, and the resulting genotype is assigned with a measure of genotype uncertainty (called a genotype quality score). Moreover, in population genetic studies, one can directly compute population genetic summary statistics based on the genotype likelihoods without an intermediate step of calling genotypes.

Although there are a number of statistical methods and associated computational tools available for genotype calling (GATK, SAMtools, SOAPsnp etc.) and for directly computing population genetic summary statistics (ANGSD, etc.), there is no established guideline and recommendation in the field for conducting population genetic studies using next-generation sequencing data to avoid bias and spurious conclusions. Furthermore, there have been no studies that compare biased pattern of population genetic summary statistics and the impact on the downstream analysis when one uses genotype calls and computes summary statistics (call-based approach) vs. one directly computes summary statistics from the sequencing data (direct approach). Hence, in this dissertation, we focus on addressing the challenges of using NGS data for population genetics inferences and suggest the best NGS data analysis pipeline to minimize the bias. Furthermore, we optimize computational tools that compute population summary statistics, called the Site Frequency Spectrum, from low- to medium-coverage NGS data after taking account for genotype uncertainty, because the existing method is computationally intractable for a large sample size. Then, we consider an application study, in which we analyze high-coverage next-generation sequencing data to detect genetic signatures of adaptive evolution during early dog domestication.

1.2 Organization

The outline of the thesis is as follows. The dissertation starts with an introductory chapter. Then, chapter 2 gives a background on NGS technologies (section 2.1) and addresses challenges in computing summary statistics for population genetic inferences from NGS data (section 2.2). Among many summary statistics, we focus our interests on estimating the site frequency spectrum (SFS). In section 2.2, we define the SFS, discuss its importance in population genetic studies, and then review several statistical methods that were proposed to account for genotype uncertainty in computing the SFS. Finally, in section 2.3, we give a background on population genetic theory of adaptive evolution, because we will analyze NGS data to detect genetic signatures of adaptive evolution during early dog domestication in chapter 5 as an application study. We describe selective sweep events, and methods to detect genetic signatures of selective sweeps in the genome.

In the following three chapters, we present our recent works in the analysis of NGS data in population genetic studies. In chapter 3, we characterize bias in population genetic inferences from NGS data by using detailed, realistic simulations. We compare the called-based approach (first inferring genotypes from the aligned short-read sequencing data and then computing summary statistics based on the inferred genotype calls) to the direct estimation approach (computing summary statistics directly from the aligned short-read sequencing data), and conclude the chapter with guidelines and recommendations for conducting population genetic inference from NGS data. In chapter 4, we discuss the computational challenges of using the direct estimation approach when a sample size is large, and conduct exploratory analysis to find computational burdens to implement the direct estimation approach. Based on the exploratory analysis, we develop the new algorithm by which we can run the direct estimation method even for a large sample of low- to medium-coverage sequencing data. In chapter 5, we analyze high-coverage next-generation sequencing data of dogs and wolves to detect of genetic signatures of adaptive evolution during dog domestication. In fact, this study was the motivation to consider all of the analysis in chapter 3 and 4 to decide the best NGS data analysis pipeline given conditions of our data set. Based on the results of chapter 3 that we can trust the SFS above 10X, we computed summary statistics used in selection scans based on the observed SFS, rather than using the direct estimation method to infer the SFS.

Finally, chapter 6 concludes this dissertation with discussions for future works. We describe the extension of our works to estimate the multi-dimensional SFS (for example, 2-dimensional SFS for a pair of populations) and discuss further computational advantages of using the new algorithm we develop for estimating the multi-dimensional SFS.

CHAPTER 2

Background

2.1 Next Generation Sequencing

The advent of next-generation sequencing (NGS) technologies provides tremendous opportunities in population genetics to perform large-scale comparative and evolutionary studies in not only model organisms but also non-model organisms. Next-generation techniques were first developed in 2005 (Margulies *et al.* (2005) describes the development of the first NGS technology using the pyrosequencing method) in response to the limitations of the automated Sanger sequencing methods, such as low throughput and high cost (Table 2.1). The major advance offered by NGS is the ability to produce an enormous amount of genomic data at low cost. While the specific methods vary by platforms (see Metzker (2010) for a review of NGS technologies and their applications), they generally use massively parallel sequencing to obtain millions of short reads from random locations in the genome. Table 2.1 compares some characteristics of the Sanger sequencing methods and the NGS methods (Shendure and Ji, 2008).

Table 2.1: Comparisons of the Sanger sequencing and NGS methods.

	Length	Error rate	Throughput	Cost
Sanger	500-1000 nt	10^{-4}-10^{-5}	6 Mb/day	\$500/Mb
NGS	75-400 nt	10^{-2}-10^{-3}	750-5000 Mb/day	\$0.5-\$20/Mb

Despite the promise of NGS for population genomic studies, there are some computational and statistical challenges associated with NGS data analysis.

2.1.1 Base calling

The first step in NGS is template preparation (Shendure and Ji, 2008). A whole genome or targeted regions of the genome (for example, exomes) is randomly digested into small fragments. Then, these small fragments get sequenced. During sequencing, base-calling algorithms infer the actual base from the fluorescence intensity, and then assign a measure of uncertainty to each base call using noise estimates from image analysis (Nielsen *et al.*, 2011). This measure is called a per-base quality score and reported in a Phred scale given by

$$Q = -10 \log_{10} P(\text{base calling error}).$$

NGS techniques have a high per-base sequencing error rate on short reads, ranging from 1 out of 1000 bases to 1 out of 100 bases (Table 2.1). In the presence of sequencing errors, it is often hard to distinguish true genetic variations from sequencing errors, in particular when coverage is low. Moreover, NGS data have variable depth of coverage due to the stochastic nature of a data generation procedure. This leads to different levels of confidence in the genotype calls, especially low confidence in the genotype calls for sites with low coverage.

2.1.2 Read mapping

The next step in NGS is aligning the resulting short reads onto an available reference genome (called read mapping)(Shendure and Ji, 2008). Because NGS techniques produce short reads with a length of 75 to 400 nucleotides (Table 2.1), there might be uncertainty in read mapping in a region of repeats or of structural variation, such as copy number variation, insertion and deletion, and chromosome rearrangement. When short reads are misaligned, we might misclassify alignment errors as genetic variations.

2.1.3 Genotype calling

Finally, from a set of aligned short-read sequencing data, genotypes for each individual are inferred (Nielsen *et al.*, 2011). Variable coverage, sequencing errors and alignment errors associated with NGS data often make accurate individual-level genotype calling problematic, and there is considerable uncertainty associated with the genotype calls. To reduce uncertainty associated with the genotype calls, researchers may either increase sequencing coverage or incorporate genotype uncertainty in a probabilistic framework:

Increasing coverage With more sequencing data at a a particular site, we can be more confident about the genotype call, as more data leads to a higher probability that two alleles at heterozygous sites are both sampled and a lower probability that sequencing errors are misidentified as mutant alleles. However, sequencing is still expensive and cost constraints lead to difficult choices between increasing sample size and increasing coverage. There are certain cases that we prefer the experimental design of large samples of low- to medium-coverage sequencing rather than small samples of high-coverage sequencing. For example, in genome-wide association studies, we obtain more power by sequencing many individuals at low coverage rather than sequencing fewer individuals at high coverage (Kim *et al.*, 2010). For the identification of rare variants (variants whose allele frequency is less than 5%), we prefer a large sample size at low coverage (1000 Genomes Project Consortium, 2010). Moreover, cost constraints will not disappear even though sequencing cost keeps dropping down, because users will continue to push limits of a large sample size with low coverage, especially with non-model organisms.

Probabilistic methods The alternative for reducing uncertainty associated with the genotype calls is using a probabilistic framework. The key quantity in this framework is *genotype likelihood* which can be calculated using the per-base quality scores for each short read.

Consider aligned short-read sequencing data X for a particular individual at a particular site. Let X_i be the ith read among the aligned short reads and let G denote a genotype for that individual. The probability $P(X_i|G)$ is given by a simple function of the per-base quality score of read X_i. For example, in the the Genome Analysis Toolkit (GATK) (DePristo *et al.*, 2011), assuming an equal chance of sampling one base (B) out of two bases in a genotype (G), we can express $P(X_i|G)$ as follows:

$$\begin{aligned} P(X_i|G = B_1B_1) &= P(X_i|B = B_1) \\ P(X_i|G = B_1B_2) &= \frac{1}{2}\{P(X_i|B = B_1) + P(X_i|B = B_2)\} \end{aligned}$$

where

$$P(X_i|B) = \begin{cases} 1 - P(\text{base calling error}) & \text{if B is same as the base in } X_i \\ P(\text{base calling error})/3 & \text{otherwise.} \end{cases}$$

Assuming independence among aligned short reads, the genotype likelihood $P(X|G)$ can be calculated as follows:

$$P(X|G) = \prod_{i=1}^{\text{coverage}} P(X_i|G).$$

It has been suggested to take correlated errors into account when computing the genotype likelihood (Li *et al.* (2008) for MAQ, Li *et al.* (2009a) for SAMtools). For example, SAMtools assumes that errors among short reads are correlated (Li *et al.*, 2009a).

Based on the genotype likelihoods $P(X|G)$ and a genotype prior $P(G)$, we can compute the posterior probability of genotype G, $P(G|X)$, by the Bayes theorem (Nielsen *et al.*, 2011)

$$P(G|X) = \frac{P(X|G)P(G)}{\sum_{G'} P(X|G')P(G')}.$$

Then, we assign the genotype with the highest posterior probability to the individual. Either this highest genotype posterior probability or the ratio between the highest and the second highest genotype posterior probability is used as a measure of confidence in genotype calls (called a genotype quality score).

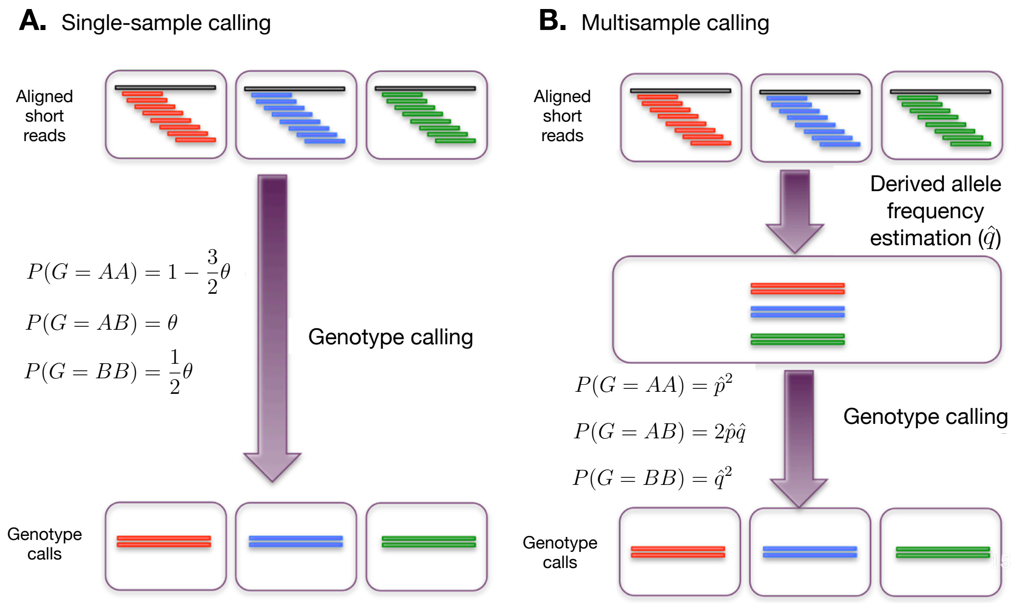


Figure 2.1: Comparisons of the single-sample and multisample calling methods.

The specification of the genotype prior is related to two modes of genotype calling, i.e. single-sampling calling and multisampling calling (Figure 2.5)(Han *et al.*, 2014). With the single-sample calling pipeline (Figure 2.5A), aligned sequencing read data are analyzed for one individual at a time and then the most likely genotypes for that individual alone are determined. Single-sample calling uses the constant prior across all sites that is based on the population mutation rate θ (more description on θ in section x):

$$P(G = AA) = 1 - \frac{3}{2}\theta$$
$$P(G = AB) = \theta$$
$$P(G = BB) = \frac{1}{2}\theta$$

where A represents an ancestral allele and B represents a derived allele. In contrast, with the multisample calling pipeline (Figure 2.5B), aligned sequencing read data are analyzed for all individuals in a sample simultaneously and then the most likely genotype configurations for all individuals are determined. Multisample calling first estimates the derived allele frequency $\hat{q}$

from sequencing data of all of the individuals at each site and then computes the genotype prior assuming Hardy-Weinberg equilibrium.

$$P(G = AA) = \hat{p}^2$$
$$P(G = AB) = 2\hat{p}\hat{q}$$
$$P(G = BB) = \hat{q}^2$$

where $\hat{p} + \hat{q} = 1$.

Finally, one can gain more precision in the genotype calls by utilizing the the pattern of linkage disequilibrium (LD) at nearby sites (Nielsen *et al.*, 2012). For example, one can impute the missing genotype of a given individual by using the haplotype information of other individuals in a reference panel (called genotype imputation; see Marchini and Howie (2010) for an overview of the statistical methods for imputing genotypes). By the same token, one can gain more confidence in the genotype call even at low coverage by utilizing the LD pattern of other individuals in a reference panel. This approach has been taken in the 1000 Genomes Project to analyze sequencing data and it has shown that this approach can lead to a significant improvement in genotype calling accuracy (1000 Genomes Project Consortium, 2010).

2.2 Estimation of Population Genetic Summary Statistics from NGS data

Population genetic inferences often proceed by compressing large-scale genetic variation data into simple and informative summary statistics, such as allele frequencies, heterozygosity, and nucleotide diversity. Due to genotype uncertainty associated with NGS data, the computation of population genetic summary statistics based on the genotype calls can lead to serious biases and possibly spurious conclusions if the coverage is not so large that the genotypes are known with absolute certainty for each individual.

To correct for such bias in estimating population genetic summary statistics from NGS data, several methods have been proposed. The simplest method is using strict filters to account for uncertainty associated with the genotype calls. A common practice is to use the genotype calls that exceed some threshold for genotype quality (GQ) or depth of coverage (DP) and treat less confident genotype calls as missing data. However, these filters can adversely affect summary statistics estimation based on the genotype calls (Johnson and Slatkin, 2008; Kim *et al.*, 2011).

The other method is developing statistical methods, in which summary statistics are directly inferred from aligned short-read sequencing data. This approach makes an implicit assumption that inferred genotypes from sequencing data are inaccurate and models this uncertainty using genotype likelihoods and some prior information. Several approaches have been developed in this framework (Johnson and Slatkin, 2008; Lynch, 2008, 2009; Liu *et al.*, 2009, 2010; Kang and Marjoram, 2011; Keightley and Halligan, 2011; Kim *et al.*, 2011).

Among many summary statistics, we focus on the site frequency spectrum (SFS), as it is a sufficient statistic for data from independent sites, and the shape of the spectra is indicative of underlying population genetic processes, such as population growth, bottlenecks, and selection. Hence, estimating the SFS is a major entry-point into many population genetic analyses, and a number of population genetic inferences can proceed directly from the inferred SFS.

2.2.1 Site Frequency Spectrum (SFS)

The SFS is defined as a distribution of allele frequencies in a sample across many unlinked loci. Assuming that mutations are rare enough that the observed SNPs are biallelic (this model is called an infinite site model), one can polarize the alleles at any particular segregating site as either being ancestral (original) or derived (mutant) with sequence data from an outgroup. For a sample of n haplotypes from a panmictic population, the SFS is an $(n-1)$-dimensional vector, given by $\xi = (\xi_1, \ldots, \xi_{n-1})'$, where the i-th entry ξ_i represents the proportion of polymorphic sites with i copies of the derived allele and $(n-i)$ copies of the ancestral allele in the sample (Figure 2.2A). Note that the SFS is sometimes defined by the absolute number of polymorphic sites rather than the proportion of polymorphic sites. Furthermore, one can specify the SFS as an

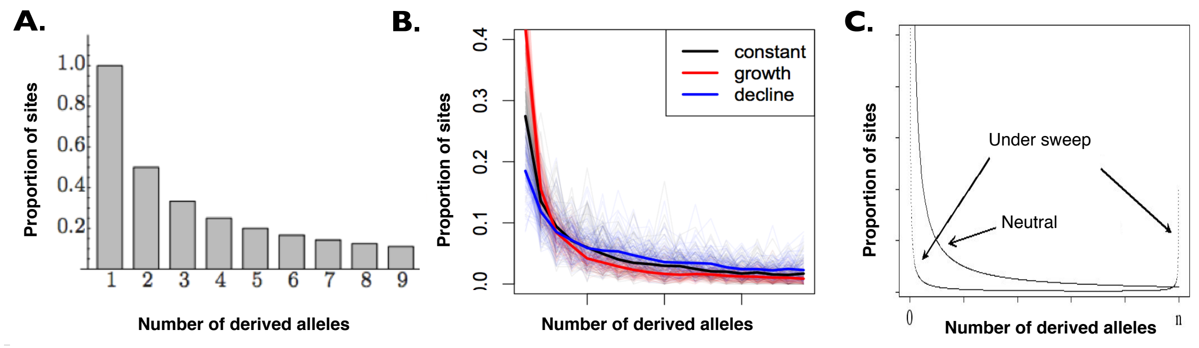


Figure 2.2: Site frequency spectrum. Panel A shows the SFS for a sample of 10 haplotypes. Panel B shows the expected SFS under population growth (shown in red) and under population decline (shown in blue) compared to the SFS under the constant size model (shown in black). Panel C shows the expected SFS under selective sweeps. Under this scenario, we expect negative values of Tajima's D.

an $(n+1)$-dimensional vector, denoted by $\xi = (\xi_0, \xi_1, \ldots, \xi_n)'$, where ξ_0 represents a proportion (or the absolute number) of monomorphic sites fixed for the ancestral allele and ξ_n represents a proportion (or the absolute number) of monomorphic sites fixed for the derived allele. In chapter 3 and 4, we define the SFS as the $(n+1)$-dimensional vector considering both monomorphic and polymorphic sites.

2.2.2 Importance of the SFS for Population Genetic Inferences

Test for Selective Sweeps First, the SFS allows us to develop statistical tests to detect selection. The standard model of population genetics assumes that the population is at mutation-drift equilibrium, evolves according to the Wright-Fisher model with a constant population size, and all mutations are selectively neutral. This standard model constitutes null hypothesis in statistical testing and rejection of the null hypothesis supports alternative hypotheses, such as selection or demographic history. These tests have been referred to as *neutrality tests*.

To develop statistical tests given a set of DNA sequences, population geneticists have derived sampling properties of summary statistics assuming that the null hypothesis holds. Under the standard model, Fu (1995) showed that

$$E(\xi_i) = \frac{\theta}{i}, \quad \text{for } 1 \leq i \leq n-1$$

where $\theta = 4N_e\mu$ is a scaled population mutation rate where N_e is the effective population size, and μ is the locus neutral mutation rate. This shows that $E(i\xi_i) = \theta$ and suggests any summary statistic $\hat{\theta}_w$ that is a weighted linear combination of the SFS

$$\hat{\theta}_w = \frac{1}{\sum_{i=1}^{n-1} w_i} \sum_{i=1}^{n-1} w_i i \xi_i$$

can form a basis for getting unbiased estimator of θ (Achaz, 2009). The choice of the weight w_i allows one to look at different parts of the SFS when estimating θ. For example, Watterson's θ estimator (Watterson, 1975) is given by taking weights $w_i = \frac{1}{i}$:

$$\hat{\theta}_S = \frac{1}{\sum_{i=1}^{n-1} \frac{1}{i}} \sum_{i=1}^{n-1} \xi_i = \frac{S}{a_n}$$

where S represents the number of segregating sites and $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$. Tajima's θ estimator (Tajima, 1983) is given by using weights $w_i = (n-i)$:

$$\hat{\theta}_\pi = \frac{1}{\sum_{i=1}^{n-1}(n-i)} \sum_{i=1}^{n-1} i(n-i)\xi_i = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i)\xi_i = \pi$$

where π represents the average pairwise differences between all sequences in the sample.

Neutrality tests based on the SFS compare two different estimators of the population mutation rate θ to determine whether the observed SFS deviates from that expected under the standard model. Under the standard model, all θ estimators should be consistent regardless of the choice of weights w_i, whereas under the alternative model (either selection or demographic events) two different θ estimators will be inconsistent. Hence, a test statistic T in the neutrality test is the difference between two θ estimators, normalized by its standard deviation (Achaz, 2009):

$$T = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{Var(\hat{\theta}_1 - \hat{\theta}_2)}}.$$

Under the null hypothesis, we expect that $E(T) = 0$. Large deviations from a null distribution of T have been used to detect local gene regions under selection, and this approach is used in many empirical genome-wide selection scans (Andolfatto, 2007; Begun *et al.*, 2007; Andersen *et al.*, 2012; Axelsson *et al.*, 2013).

One of the neutrality tests based on the SFS is Tajima's D test (Tajima, 1989). This test compares θ estimator based on the number of segregating site ($\hat{\theta}_S$) and θ estimator based on the average pairwise differences($\hat{\theta}_\pi$):

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{\alpha_n \theta + \beta_n \theta^2}}$$

where

$$\alpha_n = \frac{1}{a_n}\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n}\right) - \beta_n,$$
$$\beta_n = \frac{1}{a_n^2 + b_n}\left(\frac{2(n^2+n+3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2}\right),$$
$$a_n = \sum_{i=1}^{n-1} \frac{1}{i} \text{ and } b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}.$$

The rationale of the Tajima's D test is that $\hat{\theta}_S$ is more sensitive to low-frequency alleles compared to $\hat{\theta}_\pi$ (Figure 2.3A). Hence, a negative value of D indicates too many low-frequency sites and a positive D indicates too many intermediate-frequency sites than expected under the standard model (Figure 2.3B) (adopted from Achaz (2009)).

Demographic Inference Moreover, the SFS is very informative about population history. For example, an excess of low-frequency mutations is consistent with recent population growth, as the increase in population size in recent past distorts a coalescent tree such that there will be more and shorter branches near the root and longer branches near the tips than expected under a constant size model. Therefore, many researchers have made use of the observed SFS of putatively neutral SNPs (for example, synonymous sites) to learn about demographic history. For example, several recent large-sample sequencing studies (Coventry *et al.*, 2010; Nelson

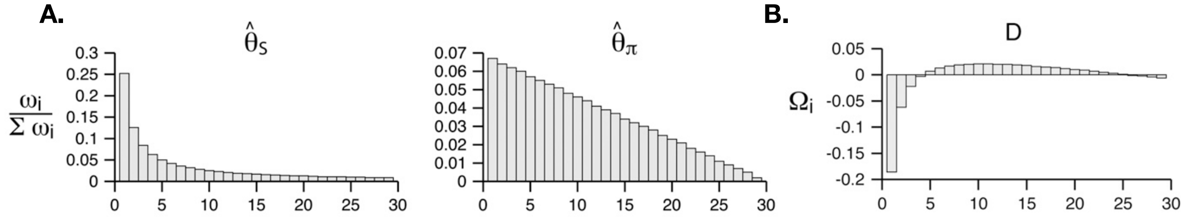


Figure 2.3: A graphical view of the weight vectors of θ estimators (A) and of neutrality tests (B) for $n = 30$ (Achaz, 2009).

et al., 2012; Keinan and Clark, 2012; Tennessen *et al.*, 2012) have found that humans have an excess of rare variants compared to the expectation from a constant population size model and these studies have inferred demographic models with recent exponential population expansion.

2.2.3 Methods for Computing SFS from NGS data

A number of methods have been proposed to compute the SFS directly from DNA sequence data in the presence of sequencing errors. Yi et al. proposed an empirical Bayes approach to estimate the joint 2-dimensional SFS between two populations (Yi *et al.*, 2010). Li et al. and Nielsen et al. separately proposed a maximum likelihood approach to compute the SFS for a single population (Li, 2011; Nielsen *et al.*, 2012). We provide a detailed review of these methods in this section, in particular for a maximum likelihood approach with an EM algorithm, because we use this in chapter 3 and 4.

Empirical Bayes Approach (Yi *et al.*, 2010) Let π_j be the posterior probability that a biallelic SNP has a derived allele frequency of $j/(2n)$ in a sample of n diploid individuals. In an empirical Bayes approach, the derived allele frequency $\hat{p}$ in a population is first estimated from all of the individuals in the sample, and then $\hat{p}$ is used for providing genotype priors assuming Hardy-Weinberg equilibrium. Based on genotype likelihoods and genotype priors, a dynamic programming algorithm is used for calculating the posterior probability π_j for each site. The estimated values of π_j can then be used for computing the SFS, either by averaging over π_j or by using a maximum a posteriori probability estimate of j.

Maximum Likelihood Approach with an EM algorithm (Li, 2011; Nielsen *et al.*, 2012) Consider a sample of n diploid individuals and a genetic region of length l sites. The SFS for n individuals across l sites is denoted by a $(2n+1)$-dimensional vector, $\xi = (\xi_0, \xi_1, \ldots, \xi_{2n})'$, where ξ_i represents a proportion of sites with the allele frequency in a sample of $i/(2n)$ and $\sum_{k=0}^{2n} \xi = 1$.

Let D_{ij} represent aligned short reads sequencing data for an individual i at a site j, and consider a matrix D to denote the observed sequencing data for all n individuals in a region of l sites:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D_1} \\ \vdots \\ \mathbf{D_l} \end{pmatrix} = \begin{pmatrix} D_{11} & \ldots & D_{1n} \\ \vdots & & \vdots \\ D_{l1} & \ldots & D_{ln} \end{pmatrix}, \; i = 1, \ldots, n, \; j = 1, \ldots, l.$$

Let G_{ij} denote a genotype for the individual i at the site j, defined as the number of derived alleles ($G_{ij} \in \{0, 1, 2\}$). Consider a matrix G to denote the genotypes for n individuals across l sites:

$$\mathbf{G} = \begin{pmatrix} \mathbf{G_1} \\ \vdots \\ \mathbf{G_l} \end{pmatrix} = \begin{pmatrix} G_{11} & \ldots & G_{1n} \\ \vdots & & \vdots \\ G_{l1} & \ldots & G_{ln} \end{pmatrix}, \; i = 1, \ldots, n, \; j = 1, \ldots, l.$$

Let X_i denote the total number of the derived allele for the sample of n diploid individuals at the site i:

$$X_i = \sum_{j=1}^{n} G_{ij}, \;\; 0 \leq X_i \leq 2n$$

and let X denote the vector of X_i across l sites:

$$\mathbf{X} = (X_1 \ldots X_l)'.$$

The E step of the EM algorithm is computing the expectation of the complete-data log likelihoods given the previous estimate of a parameter. In this example,

$$
\begin{aligned}
Q(\xi|\xi^{(\mathbf{m})}) &= E[\,\ln P(D,X|\xi)|D,\xi^{(\mathbf{m})}] \\
&= \sum_{i=1}^{l} E[\,\ln P(D_i,X_i|\Phi)|D_i,\xi^{(\mathbf{m})}] \text{ assuming independence among loci} \\
&= \sum_{i=1}^{l}\sum_{x=0}^{2n} P(X_i=x|D_i,\xi^{(\mathbf{m})}) \ln P(D_i,X_i=x|\xi) \\
&= \sum_{i=1}^{l}\sum_{x=0}^{2n} P(X_i=x|D_i,\xi^{(\mathbf{m})}) \ln\left[P(D_i|X_i=x)P(X_i=x|\xi)\right] \\
&= \sum_{i=1}^{l}\sum_{x=0}^{2n} P(X_i=x|D_i,\xi^{(\mathbf{m})}) \ln P(X_i=x|\xi) + C, \\
&= \sum_{i=1}^{l}\sum_{x=0}^{2n} P(X_i=x|D_i,\xi^{(\mathbf{m})}) \ln \xi_x + C
\end{aligned}
$$

Next, the M step of the EM algorithm is maximizing the expectation, $Q(\xi|\xi^{(\mathbf{m})})$ with respect to a parameter ξ. Define

$$\tilde{Q} = Q(\xi|\xi^{(\mathbf{m})}) + \lambda\left(1-\sum_{k=0}^{2n}\xi_k\right)$$

Because the partial derivative of $\tilde{Q}$ with respect to ξ_x leads to

$$\frac{\partial\tilde{Q}}{\partial\xi_x} = \sum_{i=1}^{l}\frac{1}{\xi_x}P(X_i=x|D_i,\xi^{(\mathbf{m})}) - \lambda = 0,$$

we have

$$\hat{\xi}_x = \frac{\sum_{i=1}^{l} P(X_i=x|D_i,\xi^{(\mathbf{m})})}{\lambda}$$

Moreover, because the partial derivative of $\tilde{Q}$ with respect to λ leads to

$$\frac{\partial\tilde{Q}}{\partial\lambda} = 1-\sum_{k=0}^{2n}\xi_k = 0,$$

we have $\hat{\lambda} = l$. Hence, we can update each element of the SFS ξ as follows:

$$\xi_x^{(m+1)} = \frac{1}{l}\sum_{i=1}^{l} P(X_i=x|D_i,\xi^{(\mathbf{m})})$$

Because by Bayes theorem, we have

$$P(X_i = x|D_i, \xi) = \frac{P(D_i|X_i = x)P(X_i = x|\xi)}{\sum_y P(D_i|X_i = y)P(X_i = y|\xi)} = \frac{h_{ix}\xi_x}{\sum_y h_{iy}\xi_y}$$

where $h_{ix} = P(D_i|X_i = x)$ denotes a site likelihood for the allele frequency $x/(2n)$, we can update each element of the SFS, ξ, as follows:

$$\hat{\xi}_x^{(m+1)} = \frac{1}{l}\sum_{i=1}^{l} \frac{h_{ix}\xi_x^{(m)}}{\sum_{j=0}^{2n} h_{ij}\xi_j^{(m)}} \quad x = 0, 1, .., 2n$$

To run the EM algorithm, we need to compute the site likelihood function $h_{ix} = P(D_i|X_i = x)$ for all sites ($i = 1, \ldots, l$) and all derived allele counts ($x = 0, 1, \ldots, 2n$), and then store them in $l \times (2n+1)$ matrix $\mathbf{H}$:

$$\mathbf{H} = \begin{pmatrix} \mathbf{h_1} \\ \vdots \\ \mathbf{h_l} \end{pmatrix} = \begin{pmatrix} h_{1,0} & h_{1,1} & \ldots & h_{1,2n} \\ \vdots & & & \vdots \\ h_{l,0} & h_{l,1} & \ldots & h_{l,2n} \end{pmatrix}$$

where

$$\begin{aligned} h_{i,x} &= P(\mathbf{D_i}|X_i = x) \\ &= \sum_{g_1=0}^{2} \cdots \sum_{g_n=0}^{2} P(\mathbf{G_i} = (g_1, ..., g_n)|X_i = x)P(\mathbf{D_i}|\mathbf{G_i} = (g_1, ..., g_n)) \end{aligned}$$

where $\mathbf{G_i}$ is a genotype configuration for n individuals at site i.

Assuming independence among individuals, we have

$$P(\mathbf{D_i}|\mathbf{G_i} = (g_1, ..., g_n)) = \prod_{k=1}^{n} P(D_{ik}|G_{ik} = g_k) = \prod_{k=1}^{n} L_{ik}(g_k)$$

where $L_{ik}(g_k) = P(D_{ik}|G_{ik} = g_k)$ represents a genotype likelihood for a genotype g_k for individual k at site i. Also, we can derive

$$P(\mathbf{G_i} = (g_1, ..., g_n)|X_i = x) = \frac{\prod_{k=1}^{n} \binom{2}{g_k}}{\binom{2n}{x}}$$

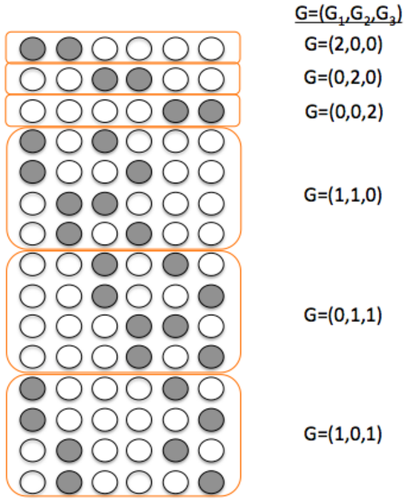


Figure 2.4: $P(G|X)$

by a probabilistic argument (Figure 2.4).

Hence, we have the following equation for the site likelihood, $h_{i,x}$, for the derive allele count x at site i:

$$\begin{aligned} h_{i,x} &= P(\mathbf{D_i}|X_i = x) \\ &= \frac{1}{\binom{2n}{x}} \sum_{g_1=0}^{2} \cdots \sum_{g_n=0}^{2} I\left(\sum_{k=1}^{n} g_k = x\right) \prod_{k=1}^{n} \binom{2}{g_k} L_{ik}(g_k). \end{aligned}$$

Nielsen and co-workers implemented a dynamic programming algorithm that was originally proposed by Li in the software package ANGSD (Li et al. 2012 and Nielsen et al. 2012). The dynamic programming algorithm updates the site likelihood vector by adding one individual at a time. For ease of notation, we drop a subscript i from the site likelihood vector $\mathbf{h_i}$ at site i. To compute the site likelihood vector $\mathbf{h} = (h_0, h_1, \ldots, h_{2n})'$ where

$$h_x = \frac{1}{\binom{2n}{x}} \sum_{g_1=0}^{2} \cdots \sum_{g_n=0}^{2} I\left(\sum_{k=1}^{n} g_k = x\right) \prod_{k=1}^{n} \binom{2}{g_k} L_k(g_k),$$

let's define the $(2j+1)$-dimensional site likelihood vector for j individuals $\mathbf{z^j} = (z_0^j, z_1^j, \ldots, z_{2j}^j)'$

where

$$z_x^j = \sum_{g_1=0}^{2} \cdots \sum_{g_j=0}^{2} I\left(\sum_{k=1}^{j} g_k = x\right) \prod_{k=1}^{j} \binom{2}{g_k} L_k(g_k)$$

where $j = 1, .., n$ and $x = 0, 1, ..., 2j$.

Given $\mathbf{z^{j-1}}$, each element of $\mathbf{z^j}$ is computed by the following recurrence:

$$\begin{aligned}
z_x^j &= \sum_{g_1=0}^{2} \cdots \sum_{g_{j-1}=0}^{2} \left[\prod_{k=1}^{j-1} \binom{2}{g_k} L_k(g_k)\right] \sum_{g_j=0}^{2} \binom{2}{g_j} L_j(g_j) I\left(\sum_{k=1}^{j-1} g_k + g_j = x\right) \\
&= \sum_{g_1=0}^{2} \cdots \sum_{g_{j-1}=0}^{2} \left[\prod_{k=1}^{j-1} \binom{2}{g_k} L_k(g_k)\right] \times [L_j(0) I\left(\sum_{k=1}^{j-1} g_k = x\right) + \\
&\quad 2L_j(1) I\left(\sum_{k=1}^{j-1} g_k = x-1\right) + L_j(2) I\left(\sum_{k=1}^{j-1} g_k = x-2\right)] \\
&= L_j(0) z_x^{j-1} + 2L_j(1) z_{x-1}^{j-1} + L_j(2) z_{x-2}^{j-1}
\end{aligned}$$

Finally, each element of $\mathbf{z^n}$ is rescaled by a corresponding factor $\binom{2n}{x}$ in order to obtain the site likelihood vector $\mathbf{h}$

$$h_x = \frac{z_x^n}{\binom{2n}{x}}$$

This algorithm requires runtime of $O(n^2)$ for each site, because updating the site likelihood vector by adding one individual at a time is done in a triangular fashion, i.e. first computing 3 elements for one individual, then computing 5 elements for two individuals, etc. $(3 + 5 + 7 + .. + (2n + 1) = \sum_{i=1}^{n}(2i + 1) = n^2 + 2n)$.

2.3 Detection of Positive Selection Using Genetic Data

Now, we turn our interests to an applied problem of detecting selective sweeps. In chapter 5, we will present our recent works of analyzing high-coverage NGS data to find genetic signatures of adaptive evolution during early dog domestication and this section provides a detailed review of the population genetic theory of the selective sweeps and existing methods to detect selective sweeps in the genome. We want to emphasize that the section 2.3.1 is a part of the previously published review in Philosophical Transactions of the Royal Society Biological Sciences (Human population structure and the adaptive response to pathogen-induced selection pressures. Novembre J and Han E, Phil. Trans. R. Soc. B (2012) 367, 878-886, doi: 10.1098/rstb.2011.0305).

2.3.1 Selective Sweeps: Hard Sweeps vs. Soft Sweeps

In evolutionary genetic studies of natural selection, two major outcomes of natural selection are now distinguished, hard sweeps and soft sweeps (Pritchard *et al.*, 2010). Hard sweeps are the outcome of a single instance of an advantageous mutation arising and spreading through a population. Because it arises from a single instance of mutation, the advantageous variant necessarily begins on a single chromosome, which defines a unique ancestral advantageous haplotype (figure 2.5a). As this ancestral haplotype increases in frequency, or 'sweeps' through the population, it brings along neutral or nearly neutral genetic variants found on the ancestral haplotype. This impact on linked variation owing to selection has been understood for some time and was first called genetic hitchhiking by Maynard Smith & Haigh (Smith and Haigh, 1974).

Recombination plays an important role in determining the chromosomal extent of genetic hitchhiking during a hard sweep. As one moves along the chromosome in either direction away from the advantageous mutation, the chromosomes carrying the advantageous mutation will show less and less of the ancestral haplotype because of the shuffling effects of recombination (figure 2.5). As a result, after the advantageous allele reaches fixation (a frequency of 100%), one finds the regions around the selected locus have a complete lack of genetic diversity (even at neutral sites) and as one looks further from the selected locus variation is restored to background levels. This trough of diversity leaves a strong observable pattern in the genome, and is one hallmark of a hard sweep as opposed to a soft sweep. Regions with an excess of rare variants might indicate regions where a hard sweep has finished and new mutations are entering the population. Regions where one haplotype has low diversity relative to all others might indicate regions partway through the hard sweep process so called partial sweeps. The first generation of selection scan methods was built to capture these signatures (e.g. Tajimas D (Tajima, 1989), integrated haplotype score (iHS) (Voight *et al.*, 2006), cross-population extended haplotype homozygosity (XP-EHH) (Sabeti *et al.*, 2007), the composite likelihood ratio test (Nielsen *et al.*, 2005), composite of multiple signals test (Grossman *et al.*, 2010)).

In contrast, models of soft sweeps are defined by the occurrence of the advantageous mutation on several haplotypes (Pennings and Hermisson, 2006b). This can occur via two major routes 1) when selection first begins to act, the advantageous mutation may be pre-existing in the population on multiple haplotypes or 2) while selection is taking place on the first instance of an advantageous mutation, additional mutations may arise on different haplotypes and likewise begin to spread. Pennings and Hermison (Pennings and Hermisson, 2006b,a; Hermisson and Pennings, 2005) have explored models of soft sweeps in depth. Major factors influencing the rate are the total mutation rate and selection coefficient in favor of the advantageous allele and the effective population size. After the completion of a soft sweep at a single locus, the advantageous allele is fixed in the population, but neutral variants at nearby locations are often not fixed (figure 2.5b). This process leads to much less obvious signatures of selection the signature in single nucleotide polymorphism (SNP) data is weak because one does not expect a big reduction in diversity at nearby loci.

2.3.2 Tests of Selective Sweeps

By using patterns of genetic variation within species (polymorphism) and between species (divergence), we can detect genomic regions that have been a target of recent positive selection. The fixation of a beneficial allele in a population distorts the patterns of neutral variation at linked loci, thereby leaving distinct signatures in a region around the locus under selection. These include reducing nucleotide diversity around the selected locus, increasing the fraction of rare and high-frequency derived alleles in the site frequency spectrum (SFS), and increasing the extent

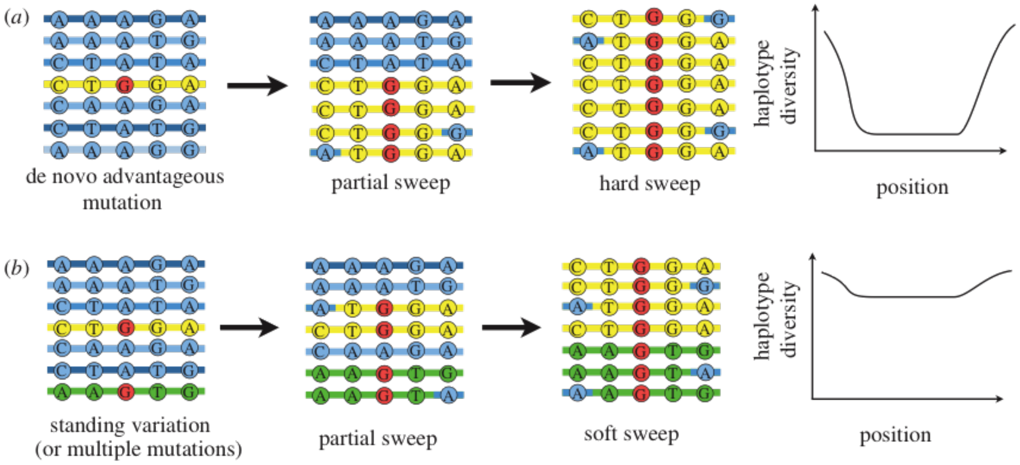


Figure 2.5: Hard and soft sweeps. (a) Hard sweeps. A de novo advantageous mutation (red G allele) arises on a single haplotype (marked in yellow). The advantageous allele increases in frequency and neutral or nearly neutral genetic variants at nearby locations also increase in frequency (genetic hitchhiking). As selection proceeds, recombination shuffles alleles off the ancestral haplotype (marked in yellow), and as a result after the completion of hard sweeps (fixation of the advantageous mutation), there is a trough in diversity around the selected locus with a sizeable reduction in genetic diversity at the location of the selected locus. (b) Soft sweeps. The advantageous mutation (red G allele) is found on multiple haplotypes (marked in yellow and green). After the completion of a soft sweep, the advantageous allele is fixed, but its less likely nearby neutral variants are also fixed. This leads to a small reduction in diversity at locations near the selected locus.

of allele frequency differences between populations, and extended linkage disequilibrium (LD) segments (Przeworski 2002, Pritchard et al. 2010). Hence, we can identify the targets of recent or ongoing selective sweeps by searching the genome for regions that show these signatures (Nair et al. 2003, Wright et al. 2005).

Diversity-based test: Local Reduction in Genetic Variations An important signature of selective sweeps is a local reduction in genetic variation around the selected site relative to its chromosomal neighbourhood or genomewide average (Smith and Haigh, 1974). Several studies have used this feature to look for loci under selection (Oleksyk et al. 2008, and others). For example, with a simple scan method, the genome is divided by sliding windows within which the average polymorphism (often measured as heterozygosity or average pairwise differences) is computed, and then regions with local dip in genetic diversity are proposed as candidate regions under selection. While the simple scan method is easy to implement, it is often difficult to distinguish this signature from the pattern generated by demographic history, such as population bottlenecks or recent founder effects, which can also reduce variation across the genome. For example, SNP analyses of domestic dogs and cats often show long stretches of homozygous regions as a result of strong bottleneck during domestication and artificial selection during breed formation (Lindblad-Toh et al. 2005, Pontius et al. 2007).

SFS-based test: Changes in the shape of SFS After the sweep is completed, new mutations gradually appear in the region. These mutations are initially present at low frequency, as their chances of increasing frequency in a population under drift are very low. Hence, in the region under selection, the expected shape of the SFS is characterized by a relative increase in the proportion or either low- or high-frequency mutations (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000). This shift in the SFS can be used in selection tests (see Neutrality tests).

F_{st}**-based test: Allele Frequency Differences between Populations** Differential selection pressures between populations can generate unusually high allele frequency differences between populations than expected under drift. The classic measure of allele frequency differentiation between populations is Wrights F_{ST} statistic (Wright, 1951). We can compute F_{ST} values across the genome, construct an empirical distribution of F_{ST} , and then look for outliers representing a set enriched for loci under selection.

CHAPTER 3

Characterizing biases in population genetic inferences from low coverage sequencing data

The work described in this chapter has been previously published in Molecular Biology and Evolution (Characterizing bias in population genetic inferences from low-coverage sequencing data. Han E, Sinsheimer J, Novembre J, Mol Biol Evol. (2014) 31: 723-35. doi: 10.1093/molbev/mst229).

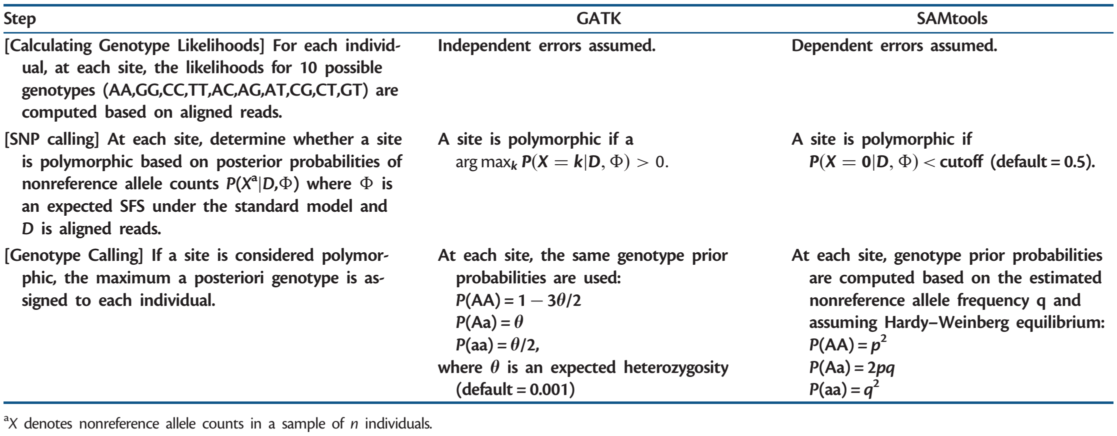
3.1 Introduction

The availability of full-genome sequence data promises to increase understanding of molecular evolution in a broad array of organisms. These large-scale data sets also raise statistical challenges because inferred genotypes from sequencing data are often inaccurate due to high error rates (e.g., base-calling and alignment errors) (Bentley *et al.*, 2008; Nielsen *et al.*, 2011). If these errors not accounted for, population genetic inference based on the genotype calls could be misleading (Pool *et al.*, 2010).

Population genetic inference often proceeds by compressing large-scale variation data into simple and informative summary statistics, such as allele frequencies, heterozygosity, or nucleotide diversity. The distribution of allele frequencies across sites, the so-called site frequency spectrum (SFS), is of primary interest, as many summary statistics are simple functions of the SFS and a number of population genetic inferences can proceed directly from the SFS. For example, a family of unbiased estimators of the population mutation rate θ, called θ estimators, is a simple function of the SFS (Achaz, 2009). These include Wattersons θ estimator that uses the number of segregating sites (Watterson, 1975) and Tajimas θ estimator that is based on the average number of pairwise nucleotide differences between two sequences (Tajima, 1983). Inferring demographic history (such as rates of ancestral population growth) can proceed from the SFS directly (Gutenkunst *et al.*, 2009) or using approximate Bayesian computation approaches (Beaumont, 2010) that often rely on summary statistics of the SFS. Another use of the SFS is in testing neutrality based on the frequency spectrum (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Achaz, 2008, 2009). Neutrality tests based on the SFS compare different estimators of θ to determine whether the observed SFS deviates from that expected under the standard constant-size equilibrium mutation-drift model. Large deviations from a background distribution have been used to detect local gene regions under selection, and this approach is used in many empirical genome-wide selection scans (Andolfatto, 2007; Begun *et al.*, 2007; Andersen *et al.*, 2012; Axelsson *et al.*, 2013).

A number of approaches can be taken to infer the SFS from NGS data. These can be classified into two broad categories. The first of these is a call-based approach, in which individual genotypes are first inferred from aligned short reads and then the SFS is estimated based on these inferred genotypes by allele counting. To infer genotypes from short-read sequencing data, a number of programs have been developed, which identify single-nucleotide variants (SNVs) and call genotypes. Among them, two of the most popular tools are the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010; DePristo *et al.*, 2011) and SAMtools (Li *et al.*, 2009a; Li, 2011). The details of the differences in the implementation of SAMtools and GATK are presented in table 3.1. Both programs determine whether a site is polymorphic based on the pileup of reads at a given site (SNV calling) and estimate individual genotypes if the site is variable (genotype calling). Each program has two different SNV and genotype calling pipelines, a single-sample and a multisample calling mode. With the single-sample calling pipeline, aligned short-read sequencing data is analyzed for one individual at a time and then the most likely genotypes for that individual alone are determined. In contrast, with the multisample calling pipeline, aligned short-read sequencing data is analyzed for all individuals in a sample simultaneously and then the most likely genotype configurations for all individuals are determined. Imputation methods represent an extension of multisample calling in which a reference panel is used and often linkage disequilibrium (LD) from multiple variant sites is integrated into making calls at any one variant (Li *et al.*, 2009b). In practice, imputation methods are generally restricted to well-studied species with reference samples such as the 1000 Genomes panel in humans (1000 Genomes Project Consortium, 2012) and the Drosophila Genome Reference panel in *Drosophila melanogaster* (Mackay *et al.*, 2012).

Table 3.1: Comparison of a GATK and SAMtools's multisample calling pipeline.



Step	GATK	SAMtools
[Calculating Genotype Likelihoods] For each individual, at each site, the likelihoods for 10 possible genotypes (AA,GG,CC,TT,AC,AG,AT,CG,CT,GT) are computed based on aligned reads.	Independent errors assumed.	Dependent errors assumed.
[SNP calling] At each site, determine whether a site is polymorphic based on posterior probabilities of nonreference allele counts $P(X^a \| D,\Phi)$ where Φ is an expected SFS under the standard model and D is aligned reads.	A site is polymorphic if a $\arg\max_k P(X = k \| D, \Phi) > 0$.	A site is polymorphic if $P(X = 0 \| D, \Phi) <$ cutoff (default = 0.5).
[Genotype Calling] If a site is considered polymorphic, the maximum a posteriori genotype is assigned to each individual.	At each site, the same genotype prior probabilities are used: $P(AA) = 1 - 3\theta/2$ $P(Aa) = \theta$ $P(aa) = \theta/2$, where θ is an expected heterozygosity (default = 0.001)	At each site, genotype prior probabilities are computed based on the estimated nonreference allele frequency q and assuming Hardy–Weinberg equilibrium: $P(AA) = p^2$ $P(Aa) = 2pq$ $P(aa) = q^2$

[a] X denotes nonreference allele counts in a sample of n individuals.

The second approach is a direct estimation approach, in which the SFS or summary statistics are directly inferred from aligned short reads. This approach makes an implicit assumption that inferred genotypes from sequencing data are inaccurate and models this uncertainty. Several approaches have been developed in this framework (Johnson and Slatkin, 2008; Lynch, 2008, 2009; Liu *et al.*, 2009, 2010; Kang and Marjoram, 2011; Keightley and Halligan, 2011; Kim *et al.*, 2011). Recently, Li (2011) proposed an EM algorithm and Nielsen et al. (2012) proposed an approach using Broyden-Fletcher-Goldfarb-Shanno (BFGS) steps to obtain the maximum likelihood estimate (MLE) of the SFS based on individual genotype likelihoods across all individuals and all sites. Both of these methods are implemented in the ANGSD software (Li, 2011; Nielsen *et al.*, 2012).

In this chapter, we use detailed, realistic simulations to investigate the accuracy of these approaches to infer the SFS from NGS data and the impact of bias in the inferred SFS on the downstream analysis, such as genome-wide selection scans based on rank statistics and parameter estimates for a given demographic model. Motivated by an interest in populations and species that have nonexistent or poor imputation panels, we focused here on two-stage approaches that use single-sample and multisample calls to infer the SFS. On the basis of our findings, we conclude with guidelines and recommendations for conducting population genetic inference using low-coverage sequencing data to avoid spurious conclusions.

3.2 Materials and Methods

To compare different approaches for estimating the SFS from sequencing data, we first conducted population genetic simulations to produce haplotype data and then overlaid sequencing errors assuming a paired-end short read sequencing approach.

3.2.1 Population Genetic Simulations

We simulated phased haplotypes for individuals by coalescent simulations under three different scenarios: the standard model (a neutral model with a constant population size) and two deviations from the standard models: a neutral model with an exponential population growth and positive selection on a new beneficial allele (a hard sweep model where a newly arisen beneficial allele increases in frequency and ultimately is fixed in a population). All coalescent simulations were performed using MSMS (Ewing and Hermisson, 2010) with an effective population size of 10,000 diploid individuals, a mutation rate per-base per-generation of 2.5×10^{-8}

and a recombination rate of 1×10^{-8}. To simulate exponential population growth, we assumed that the population began with an initial population size of 10,000 to reach a present size of 40,000 in 16,000 generations (i.e., growth rate of 0.01%). To simulate exponential population decline, we used the initial population size of 40,000 that reached a present size of 10,000 in 16,000 generations (i.e., growth rate of -0.01%). To simulate positive selection, we introduced a new advantageous mutation with a selective advantage of 0.01 in the middle of the simulated region and conditioned the simulations on the allele just reaching fixation in a population. Under each scenario, we simulated 100 replicates of 100 kilobase pair (kb) genomic regions for a sample size of 10 diploid individuals to evaluate the accuracy of the estimated SFS. To perform genome-wide selection scans and parameter estimation for the exponential population growth, we simulated 10 megabase pair (Mb) genomic regions for a sample size of 10 diploid individuals. Finally, we randomly combined pairs of haplotypes to create genotype data, an assumption of panmixia.

3.2.2 Sequencing Experiment Simulations

To simulate 100-bp paired-end short read sequencing data for a given individual, we first sampled one of two haplotypes with an equal probability and then picked a starting position of the first read uniformly and a starting position of the second read by adding a paired-end distance from the last position of the first read. The paired-end distance was chosen according to a Poisson distribution with a rate set to 204 bp based on analysis of an Illumina 100 bp paired-end library of Drosophila melanogaster sequences (results not shown). On the basis of the two starting positions for the paired reads, we generated each read based on the underlying haplotype but with errors introduced according to the empirical distribution of base quality scores (after recalibration) from the same sequence library. The distribution of observed error rates from sequencing experiment simulations is shown in figure 3.11.

3.2.3 Estimating the SFS

We assessed two ways to infer the SFS: the call-based and direct estimation approaches. With the call-based approach, we first inferred individual genotypes from aligned sequencing data and then computed the SFS from genotype calls by simple allele counting. In this case, we ignored uncertainty associated with genotype calls. To infer individual genotypes, we used one of two freely available programs, GATK (version 2.1.11) and SAMtools (version 1.4), and in each program we used either their single-sample or multisample calling procedures. Through this article, we refer to the results of these procedures as Single-GATK, Single-SAMtools, Multi-GATK, and Multi-SAMtools. To reconstruct the SFS from genotype calls by allele counting, we only used fully observable sites: the sites in which all individuals in a sample have at least one short read covering the site (hence, a genotype is observable for all individuals). With the direct estimation approach, we directly estimated the SFS from aligned sequencing data without inferring genotypes (Nielsen et al. 2012). We used the freely available program ANGSD (version 0.522) with an EM algorithm option to obtain the MLE of the SFS (Nielsen et al. 2012). We refer to results of this procedure as Direct.

3.2.4 Computing Summary Statistics for Population Genetic Inference

On the basis of the estimated SFS, we computed θ estimators and neutrality test statistics. We computed four θ estimators: 1) two original θ estimators, Wattersons θ estimator ($\hat{\theta}_s$) based on the number of segregating sites (S) and Tajimas θ estimator ($\hat{\theta}_\pi$) based on the average pairwise differences (π), and 2) two more recent θ estimators that ignore singletons to increase robustness to sequencing error, one derived from Wattersons θ estimator ($\hat{\theta}_{s-1}$) and one derived from Tajimas θ estimator ($\hat{\theta}_{\pi-1}$) (Achaz, 2008, 2009). In the absence of sequencing errors and under a strict neutral model, these θ estimators are unbiased estimators of a population mutation rate $\theta = 4N_e\mu$, where N_e is an effective population size and μ is a mutation rate per-site per-generation. For neutrality tests based on the SFS, we used Tajimas D as it is a well used and powerful test of neutrality (Simonsen *et al.*, 1995; Fu, 1997) and Achazs Y (Achaz, 2008), which is derived from Tajimas D by ignoring singletons. Without sequencing errors and under the standard model with a constant population size, the expected value of D and Y are near zero regardless of sample size (Tajima, 1983; Achaz, 2009). The variance of D is expected to be one, but recombination reduces the variance in D to be smaller than one (Tajima, 1989).

3.2.5 Quantification of Accuracy of the SFS Estimation

To evaluate the accuracy of the SFS estimated from sequencing data as a function of coverage, we computed the KL divergence of the estimated SFS from the ground-truth SFS (computed from genotype data) for each SFS estimation method. We also evaluated the accuracy of the estimated SFS in each nonreference allele frequency bin $i/(2n)$ in a sample of n diploid individuals. For each nonreference allele frequency bin, we computed a relative deviation of the fraction of sites with frequency $i/(2n)$ in the estimated SFS $f_{seq}(\frac{i}{2n})$ from that in the ground-truth SFS $f_{true}(\frac{i}{2n})$:

$$\text{Relative deviation } (\frac{i}{2n}) = \frac{f_{seq}(\frac{i}{2n}) - f_{true}(\frac{i}{2n})}{f_{true}(\frac{i}{2n})}$$

To compare ground-truth SFS to the estimated SFS by each allele frequency bin, we made error matrices E of dimension $(2n + 1)$ by $(2n + 1)$. Each element E_{ij} of the error matrix E $(i, j = 0, 1, ..., 2n)$ is the fraction of the sites where the observed counts (the nonreference allele counts at each site computed from sequencing data) are j and the ground-true counts (the nonreference allele counts from genotype data) are i. Hence, diagonal elements E_{ii} of E represent the fraction of correctly estimated sites (true positives) for each allele frequency bin $i/(2n)$.

3.2.6 Genome-Wide Selection Scans

To simulate a genome-wide selection scan, we generated 10Mb genomic regions in which a new beneficial mutation arose in the middle of the region and identified a candidate region of positive selection by an outlier detection approach scan (Andolfatto, 2007; Begun *et al.*, 2007; Andersen *et al.*, 2012; Axelsson *et al.*, 2013):

1. Estimated the SFS by using the call-based or the direct estimation approach in sliding windows of size 100Kb with an increment of 20Kb.

2. Computed Tajimas D associated with each window based on the estimated SFS.

3. Converted Tajimas D to empirical P values based on their ranks.

4. Identified outlier windows if the empirical P value associated with a given window is less than 1%. The cutoff of 1% was chosen based on visual identification of an outlier mode presumed to represent selected loci (figure 3.15).

3.2.7 Estimating Parameters in an Exponential Population Growth Model

For demographic inference, we used the python module dadi (Gutenkunst *et al.*, 2009). Dadi finds MLEs of parameters for a user-specified demographic model based on the observed SFS. We simulated a 10Mb genomic region under the exponential population growth model and then estimated the present population size (N) and time when the growth had started (T, measured in units of 2 N generations). We found the MLEs first by a grid search to find a peak of likelihood surface and then by BFGS steps to localize the peak.

3.3 Results

3.3.1 Evaluating Accuracy of the Inferred SFS under the Standard Model

We first evaluated the performance of the two SFS estimation approaches (the call-based and direct estimation approach) as a function of sequencing coverage. For this comparison, we simulated 100 replicates of sequencing data for 10 diploid individuals each from genomic regions of length 100Kb under the standard model. The accuracy of the inferred SFS was evaluated by two metrics: (1) the shape of the inferred SFS in comparison to the ground-truth SFS (figure 3.1A and B) and (2) the distance between the inferred SFS from the ground-truth SFS as measured by the KullbackLeibler divergence metric (KL divergence, see Materials and Methods) (figure 3.1C).

We found that the direct estimation approach (represented as Direct) outperformed the call-based approach (represented as Single-GATK, Multi-GATK, Single-SAMtools,and Multi-SAMtools) across all coverage ranges (figure 3.1). The inferred SFS by the direct estimation approach was most similar to the ground-truth SFS. In contrast, the estimated SFS by the call-based approach became less accurate as coverage decreased and most of the deviation came from the sites with low allele frequency, such as singletons and doubletons (figure 3.1A and B). For higher coverage data (10X per individual), the estimated SFS by the call-based methods approaches the ground-truth SFS, but the difference does not become negligible until 20X or higher (data not shown).

We also found that, depending on the genotyping pipeline (single-sample or multisample calling), the call-based approach resulted in different levels of performance in estimating the SFS. Interestingly, bias at the sites with rare variants went in opposite directions single-sample calling led to overestimation of rare variants, whereas multisample calling led to underestimation of rare polymorphisms (figure 3.1A and B). At coverage 2X, on average, singleton calls by single-sample calling were increased by more than 100% and doubleton calls were increased by

90%, thus leading to a skew in the SFS toward rare variants. In comparison, singleton calls by multisample calling were decreased by 60% and doubleton calls were decreased by 10%. This led to a distortion of the observed SFS, so that singletons were observed less often than doubletons, which is unexpected under the standard model. Overall though, we observed that the call-based approach with multisample calling (represented as Multi-GATK and Multi-SAMtools) performed better than the call-based approach with single-sample calling (represented as Single-GATK, Single-SAMtools) as reflected by the smaller KL divergence for multisample calling (figure 3.1C).

The opposite performance of the single-sample and multisample caller (i.e., the multisample caller leading to underestimation of rare variants, whereas single-sample caller leading to overestimation of rare variants) is likely because a small number of erroneous reads strongly affects a single-sample caller, whereas a small number of correct alternate reads tends to be ignored in multisample caller. For example, at a site for an individual, suppose that we observe three aligned reads with two reference bases (R) and one nonreference base (V). If the base quality is reasonable, a single sample caller will often weigh the nonreference base as a real variant and produce a heterozygote call (G = R/V) even though a site is truly fixed for a reference allele. In contrast, if all other individuals are fixed for the reference, the multisample caller will more often consider the nonreference base as a sequencing error and produce a homozygote call (G = R/R) even though a site is a truly singleton site and reads come from a heterozygous individual.

Finally, controlling for the genotype calling pipeline, the KL divergence was smaller for SAMtools than GATK (figure 3.1C). Consistent with this, we observed that SAMtools led to less overestimation (with single-sample calling) or less underestimation (with multisample calling) problems at sites with low frequency (figure 3.1A and B). That said, SAMtools appears to be systematically underestimating minor allele frequencies, which causes underestimation for low-frequency nonreference alleles and overestimation for high-frequency nonreference alleles. Around frequency 1/2, SAMtools either underestimates or overestimates nonreference allele frequencies (depending on which allele is minor) leading to the lowest accuracy around frequency 1/2 (figure 3.8). The different performance between GATK and SAMtools might be due to different models for calculating genotype likelihoods (step 1 in table 3.1) and different priors for inferring genotypes (step 3 in table 3.1).

3.3.2 Impact of Filtering

When analyzing sequencing data, researchers often use strict filters to account for uncertainty associated with genotype calls. A common practice is to use genotype calls that exceed some threshold for genotype quality (GQ) or depth of coverage (DP) and treat less confident genotype calls as missing data. However, these filters can adversely affect SFS estimation based on genotype calls (Johnson and Slatkin, 2008; Kim *et al.*, 2011). Therefore, we explored whether it is better to estimate the SFS with filtering or without filtering. As a filter, we used a combination of GQ of 0 or 20, and DP of 0 or half of mean coverage (i.e., 1 for 2X, 2 for 5X, 5 for 10X, and 10 for 20X). Figure 3.2 shows that filtering based on GQ or DP does not alleviate the bias associated with called-based approaches.

3.3.3 Impact on θ Estimators and on Neutrality Tests under the Standard Model

Next, we investigated the impact of bias in inferred SFS on θ estimators and a neutrality test. With the call-based approach, both θ estimators and the neutrality test were biased. The bias direction depended on the genotype calling pipeline (figure 3.3, call-based): with the single-sample calling pipeline, $\hat{\theta}_s$ and $\hat{\theta}_\pi$ were overestimated and Tajima's D was negatively skewed because of an excess of low frequency variants in the inferred SFS, whereas with the multisample calling pipeline, $\hat{\theta}_s$ and $\hat{\theta}_\pi$ were underestimated and Tajima's D was skewed toward positive values due to a deficit of low frequency variants in the inferred SFS. Comparing $\hat{\theta}_s$ and $\hat{\theta}_\pi$, the bias was bigger in $\hat{\theta}_s$ than in $\hat{\theta}_\pi$ for a sample size of 10. This is because adding a new artificial singleton by sequencing errors adds a new segregating site but adds only 2/10 to the average pairwise differences. In contrast, for the direct estimation approach, both $\hat{\theta}_s$ and $\hat{\theta}_\pi$ were unbiased (mean $\hat{\theta}_s$ and $\hat{\theta}_\pi$ were close to true value of 0.001) and consequently Tajima's D was unbiased (mean D was close to zero as expected under the standard model (figure 3.3, Direct).

Motivated by the fact that sequencing errors typically appear as artificial singletons and result in a false excess of observed singletons, Achaz (2008) proposed to ignore singletons when computing θ estimators to reduce bias while retaining a powerful enough test to detect deviations from the standard model. We explored if using Achaz's correction followed by the call-based approach can reduce bias in θ estimators and in the neutrality test (figure 3.3, call based + correction). In our simulated sequencing data, however, his assumptions about sequencing errors occurring as only singletons were violated. We observed sequencing errors affected not only singletons but also other allele frequency bins (figure 3.8) and sequencing errors led to either an excess of singletons (with the single-sample calling pipeline) or a deficit of singletons (with the multisample calling pipeline). Nevertheless, Achaz's correction followed by the call-based approach could reduce bias in θ estimators and Tajima's D across ranges of coverage.

3.3.4 SFS and Parameter Estimation under the Exponential Population Growth

To explore robustness of SFS estimation to departures from the standard model, we evaluated the performance based on the simulated sequencing data under an exponential population expansion model with a growth rate of 0.01% (figure 3.4). As expected, we observed that the ground-truth SFS under the exponential population growth model showed an excess of rare polymorphisms compared with that under the constant population size model (figure 3.13) and resulted in a negative Tajima's D (figure 3.4D).

We observed similar bias patterns as in figure 3.1: The direct estimation outperformed that the call-based approach. The estimated SFS by the direct estimation approach was most similar to the ground-truth SFS across the range of coverages simulated, whereas the estimated SFS by the two-step estimation approach had bias in that rare variants were overestimated with the single-sample calling pipeline and underestimated with the multisample calling pipeline at low coverage (figure 3.4A and B). Furthermore, bias in the estimated SFS subsequently influenced neutrality tests: Tajima's D with the multisample calling pipeline was more negative (figure 3.4D).

Interestingly, under the population growth model, the single-sample calling pipeline performed better than the multisample calling pipeline as shown by the KL divergence (figure 3.4C). In particular, at coverage 2X, the estimated SFS with the multisample calling pipeline in GATK was extremely distorted in that singleton calls were less than doubleton calls (figure 3.4A), which in turn led to a positive Tajima's D showing an evidence of population contraction (figure 3.4D). The poor performance of the multisample calling pipeline is because the Bayesian inference for SNP discovery and genotype calling in GATK and SAMtools is based on priors that are derived under a constant size model.

Next, we investigated how the bias in the estimated SFS affects demographic inference based on the inferred SFS. By using *dadi*, we estimated parameters for the exponential population growth model, such as a present population size (N) and a time when population growth has started (T), based on the inferred SFS from sequencing data (figure 3.5). The MLE of the growth rate with the direct estimation approach was almost unbiased across all ranges of coverage (close to the true growth rate 0.01%), whereas the growth rate was overestimated with the call-based approach with the single-sample calling pipeline and underestimated with the call-based approach with the multisample calling pipeline. This bias became more serious as coverage decreases: In particular, at coverage 2X, the growth rate estimate from GATK multisample calling becomes negative (-1%) indicating the inappropriate inference of population contraction rather than growth.

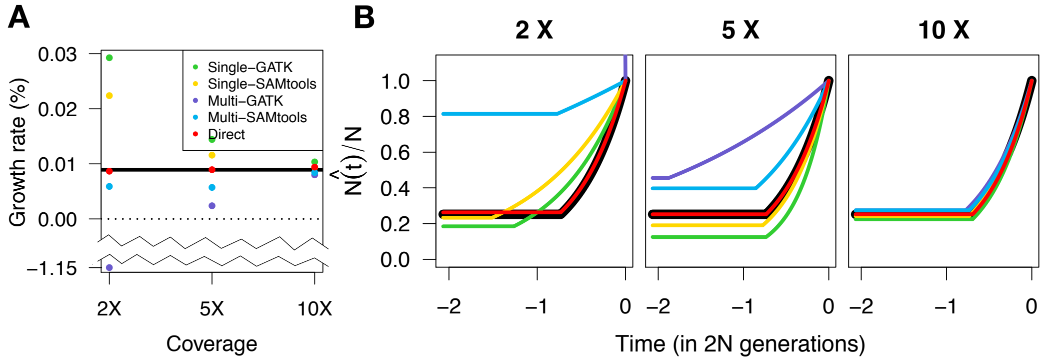


Figure 3.5: Estimation of a population growth rate by using dadi as a function of coverage. (A) Inferred growth rates for each method and the true growth rate (shown in black, 0.01%). (B) Inferred population size trajectory over time compared with the simulated trajectory (shown in black).

3.3.5 Impact of Changes in Parameters

To assess the robustness of our results, we explored how changes in nucleotide diversity (θ), sequencing error rates (ϵ), and underlying coalescent models affect the SFS estimation. To allow a straightforward comparison, we used the same parameters as in figure 3.1 apart from varying one parameter of interest at a time.

First, we examined the case where expected nucleotide diversity is five times smaller than the sequencing error rate ($\theta = 2 \times 10^{-4}, \epsilon = 10^{-3}$) and five times larger than the error rate ($\theta = 5 \times 10^{-3}, \epsilon = 10^{-3}$). Figure 3.9 and 3.10 show that the SFS reconstruction methods behave almost identically as in figure 3.1 we observe that the SFS estimated by the direct estimation method is close to the true SFS even at 2X, whereas the SFS by the call-based approach is biased in that the single-sample caller overestimates rare variants and the multisample caller underestimates rare variants. However, when diversity gets smaller than the error rate, we observe that the KL divergence is larger for the single-sample caller compared with the multisample caller (figure 3.6A). When diversity becomes larger than the error rate, the KL divergence for both single-sample and multisample caller becomes larger (figure 3.6A).

Next, we explored the effect of sequencing error rates on the SFS reconstruction with a fixed diversity of 10^{-3} under the standard model. We observed similar bias patterns to previous cases (figure 3.12), but when the error rate reaches 10^{-1}, we need coverage higher than 20X for the estimated SFS by the call-based approach to be correct.

Finally, we examined how underlying coalescent models affects the SFS reconstruction based on sequencing data. We examined the case where the SFS is skewed to rare variants (population growth model) and the SFS is skewed to medium frequencies (population decline model) (figure 3.13). In both cases, we observed that the bias pattern in the inferred SFS was similar to that for the constant population size model (figure 3.4 for the population growth model, figure 3.14 for the population decline model). We also observed that the violation to the constant size model led to a larger KL divergence for the multisample caller than the single-sample caller (figure 3.6).

3.3.6 Genome-Wide Selection Scans

We next explored how error in the SFS affects the performance of genome-wide selection scans by an outlier detection approach. For this evaluation, we simulated sequencing data of length 10 Mb where a new beneficial mutation arose around 5 Mb, increased in frequency, and became fixed at the time of sampling. Figure 3.7B shows that at coverage 2X, Tajimas D with the direct estimation approach was almost unbiased in both neutral and selected regions, whereas Tajimas D was skewed positive with the call-based approach with multisample calling and skewed negative with the call- based approach with the single sample calling. However, after converting Tajimas D to rank-based statistics, such as empirical P values, the difference between the direct estimation and call-based approach became negligible enough to select the same set of windows as a candidate region of a positive selection even at low coverage (figure 3.7A). This indicates that rank-based statistics are less sensitive to bias in the inferred SFS, and if a positive selection is strong enough to be distinguishable from the neutral background, one can identify regions of positive selection with relative robustness to the SFS estimation approach used. However, over 100 replicates, the direct method had a higher power and smaller false-positive rates than the call-based approaches, and all call-based approaches performed with similar power and false-positive rates (figure 3.7C).

3.4 Discussion

With the rapid development of sequencing technologies, the obstacle in population genetic studies is in our ability to interpret such data with precision. The results shown here demonstrate that, depending on the pipeline used to analyze sequencing data, one can reach starkly different conclusions with the same data set. Simple allele counting after inferring individual genotypes from aligned sequencing data (call-based approach) leads to bias in the estimated SFS toward the sites with rare variants, and this bias is in opposite directions depending on the pipeline to infer genotypes: Multisample calling leads to underestimation of rare variants, whereas single-sample calling leads to overestimation of rare variants. Next, the bias in the inferred SFS subsequently results in bias in θ estimators, neutrality test, and demographic inference. In contrast, we have shown that the SFS directly estimated from aligned sequencing data (direct estimation approach) was almost unbiased across ranges of coverage. Finally, genome- wide selection scans based on rank-based statistics are less sensitive to bias in the inferred SFS enough to capture the correct regions of positive selection even at low coverage. Given that many current studies using low to medium coverage sequencing data often use inferred genotypes to precede population genetic inference, our studies highlight that care is vital to avoid any potential bias problems and incorrect conclusions.

We reason that the increased performance of the direct estimation approach over the call-based approach is that it gains information from other individuals across all sites, whereas the call-based approach with multisample calling gains information from other individuals only at a given site and that with single-sample calling considers read data only for a given individual at a given position. Moreover, because the direct estimation approach can easily handle missing data, more information can be utilized to infer the SFS. To estimate the SFS from genotype calls by allele counting, we only used fully observable sites. The fraction of fully observable sites rapidly decreases as coverage decreases. We observed that for a sample of 10 individuals, only 20% of sites are fully observable at coverage 2X, 90% of sites at coverage 5X, and 99.9% of sites at coverage 10X. Handling missing data in SFS- based approaches has been a problem before short-read sequencing data and approaches to ameliorate the problem include subsampling the data down to a sample size for which most sites are observed (e.g., (Nelson *et al.*, 2012). An advantage of the direct estimation approach is that it can easily handle missing data during SFS estimation: It assigns a noninformative genotype likelihood for missing genotypes and maximized the likelihood of the SFS. In this way, it can utilize full information available in data, though it comes at a greater computational cost associated with the EM algorithm.

It is worth noting that there exist other frequently used tools for SNP discovery and genotype calling other than GATK and SAMtools. Among them, Stacks (Catchen *et al.*, 2013) is a popular pipeline commonly used. Stacks is similar to the single sample calling in that it only considers read data for a given individual at a given site: It models read data for a single individual at a specific site with a multinomial distribution with a sequencing error rate for each site estimated by maximum likelihood (Hohenlohe *et al.*, 2010). Then, it uses a likelihood ratio test (LRT) to assess the support for the most likely genotype at a 5% significance level. If the LRT is not significant, then the model assigns a homozygote genotype for the most commonly observed nucleotide. Another tool, Beagle (Browning and Browning, 2009; Browning and Yu, 2009), takes advantage of the pattern of LD at nearby sites to infer genotypes, and as a result, genotype calling accuracy is significantly improved and missing genotypes can be imputed. However, Beagle requires a modest sample size (e.g., on the scale of 50 individuals or higher) for LD information and imputation, and this can be challenging for studies with nonmodel organisms.

We should emphasize that our simulation studies are based on multiple assumptions that can be often violated in reality. In our simulation of sequencing data, we assumed that reads had been aligned to the reference without errors. In practice, however, this assumption is often violated in a region with repeats, insertions, deletions, and copy number variants. Hence, it might be important to catalog such regions to avoid potential bias due to alignment errors. Furthermore, we assumed that the number of reads at each site for a given individual is distributed according to a Poisson distribution. It is well known that the distribution of the number of reads follows an overdispersed Poisson distribution. Therefore, even though we concluded that the bias is almost negligible at mean coverage greater than 20X from our simulation studies, in reality, we might still observe nonnegligible bias at such coverage.

One may argue that future studies will have increased coverage and many of these problems will disappear. However, with limited budgets, we expect a category of experimental work will continue in which it is most advantageous to maximize the number of individuals by using low coverage. The insights gained here suggest how careful analysis of low-coverage data can provide useful population genetic inferences and that unquestioning use of basic analysis pipelines will be problematic.

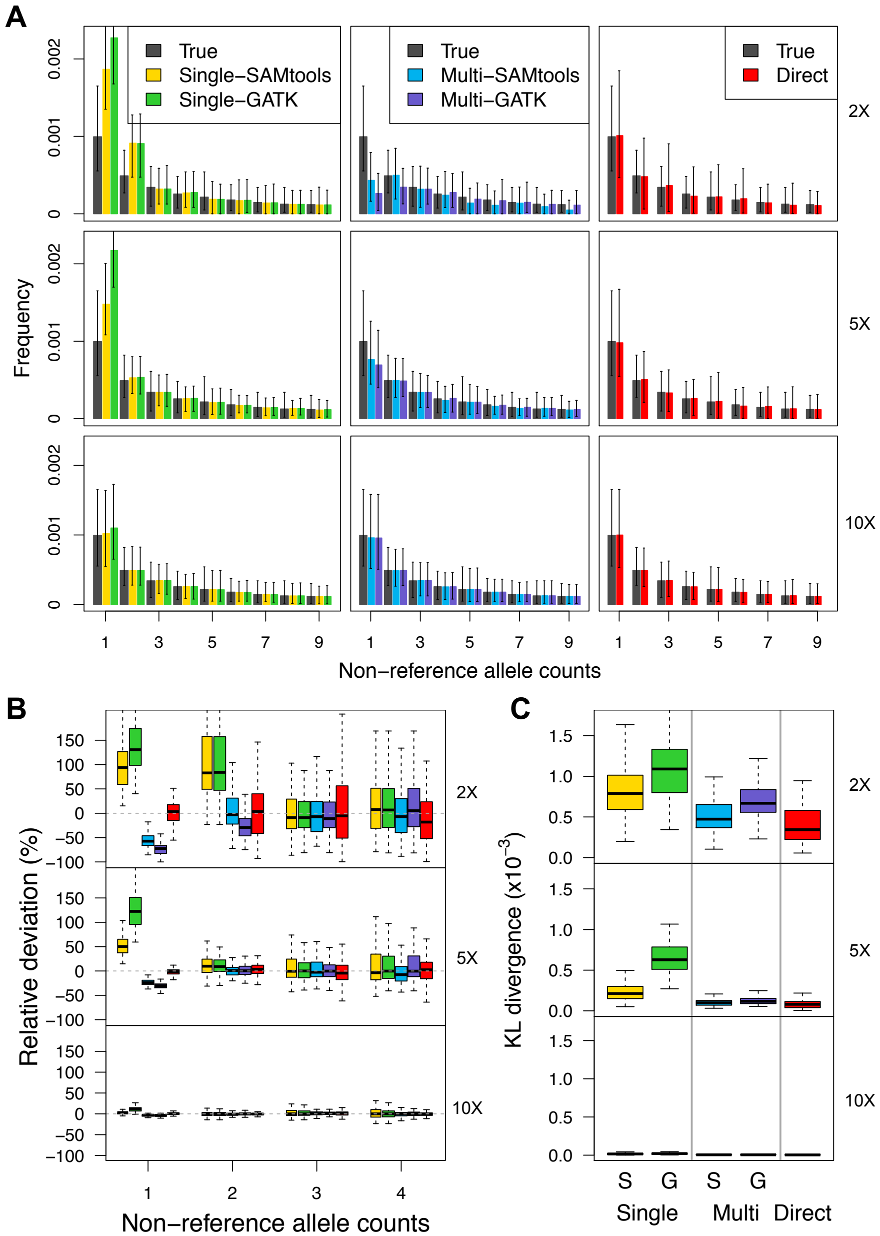


Figure 3.1: Evaluation of accuracy of inferred SFS by the call-based and direct estimation approach based on 100 replicates of genomic regions of length 100 kb. (A) Shapes of the inferred SFS (shown in colors in legend) compared with the ground-truth SFS (shown in gray) for coverage 2X (top), 5X (middle), and 10X (bottom). (B) Relative deviation of a fraction of sites with the nonreference allele counts of 1-4. (C) Distance between the inferred and ground-truth SFS as measured by KL divergence.

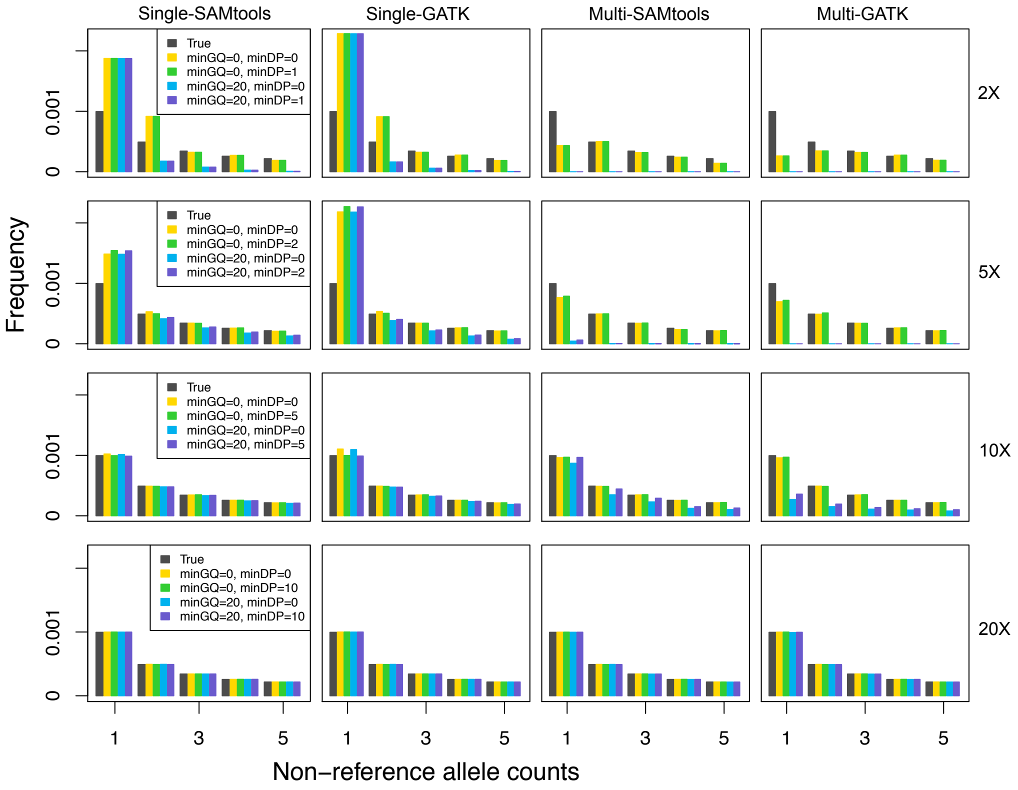


Figure 3.2: The effect of filtering of on the SFS construction for each call-based approach (panel columns) and coverages of 2X, 5X, 10X, and 20X (panel rows).

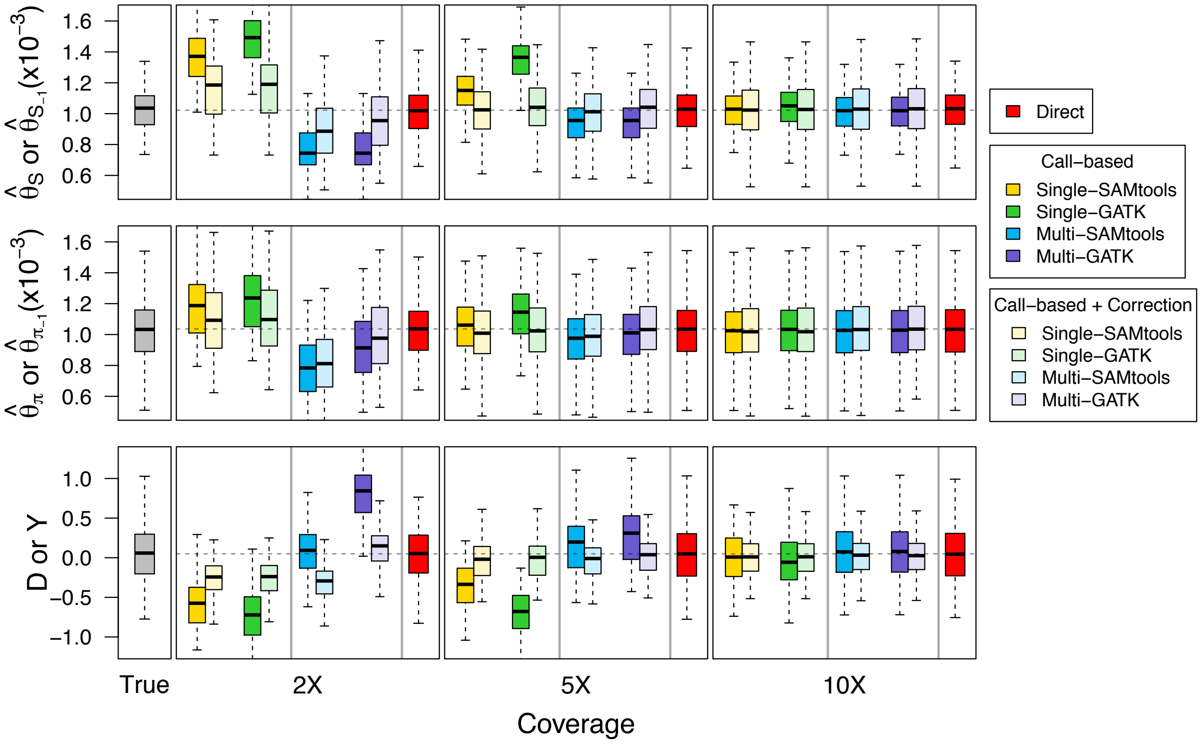


Figure 3.3: Bias in θ estimators (top, middle) and neutrality test statistics (bottom) by the call-based approach alone, the call-based approach plus Achazs correction, and the direct estimation approach, as a function of mean coverage.

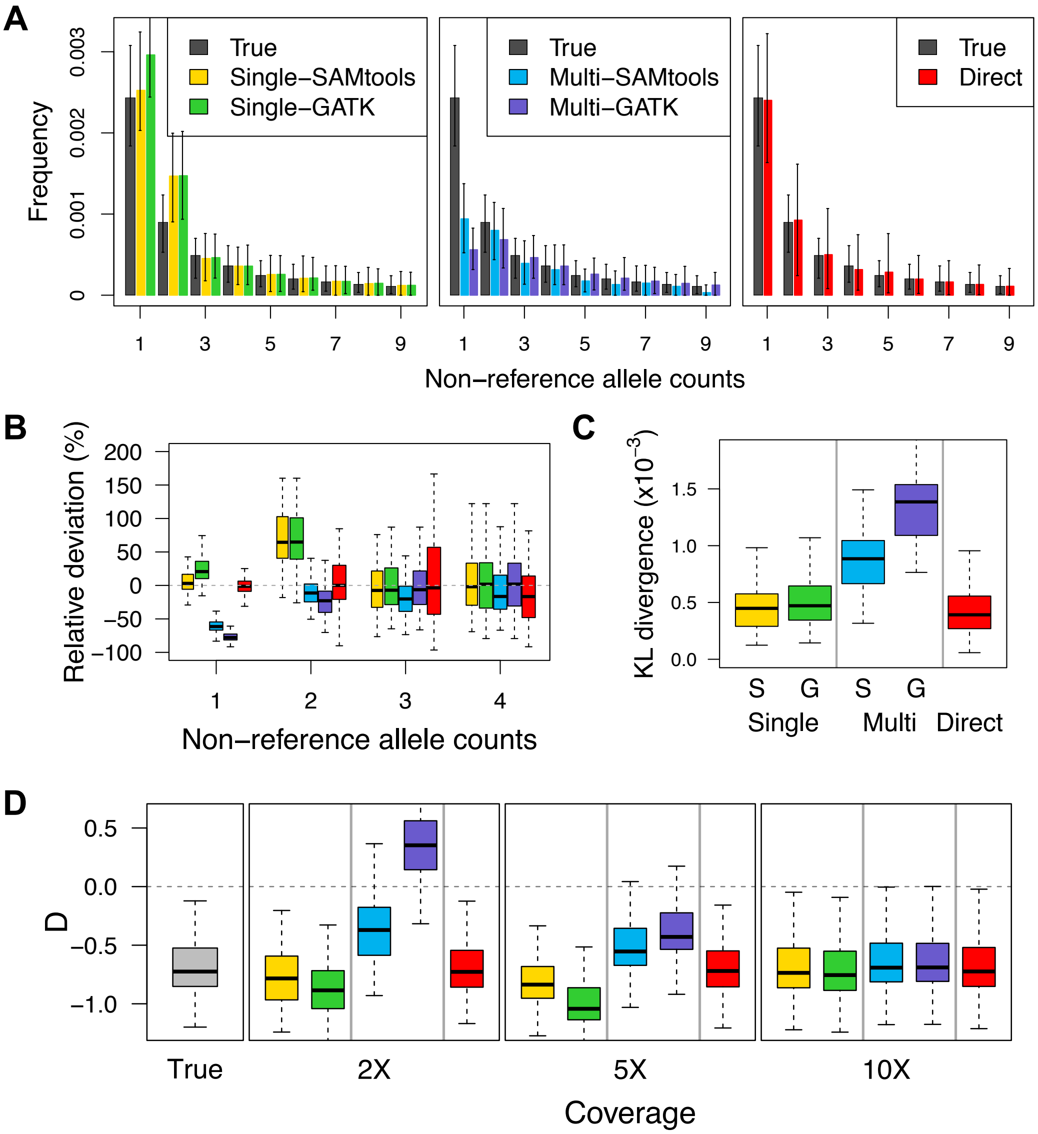


Figure 3.4: Accuracy of the inferred SFS (AC) for coverage 2X and bias in the neutrality test (D) under the exponential population growth model for coverage 2X, 5X, and 10X.

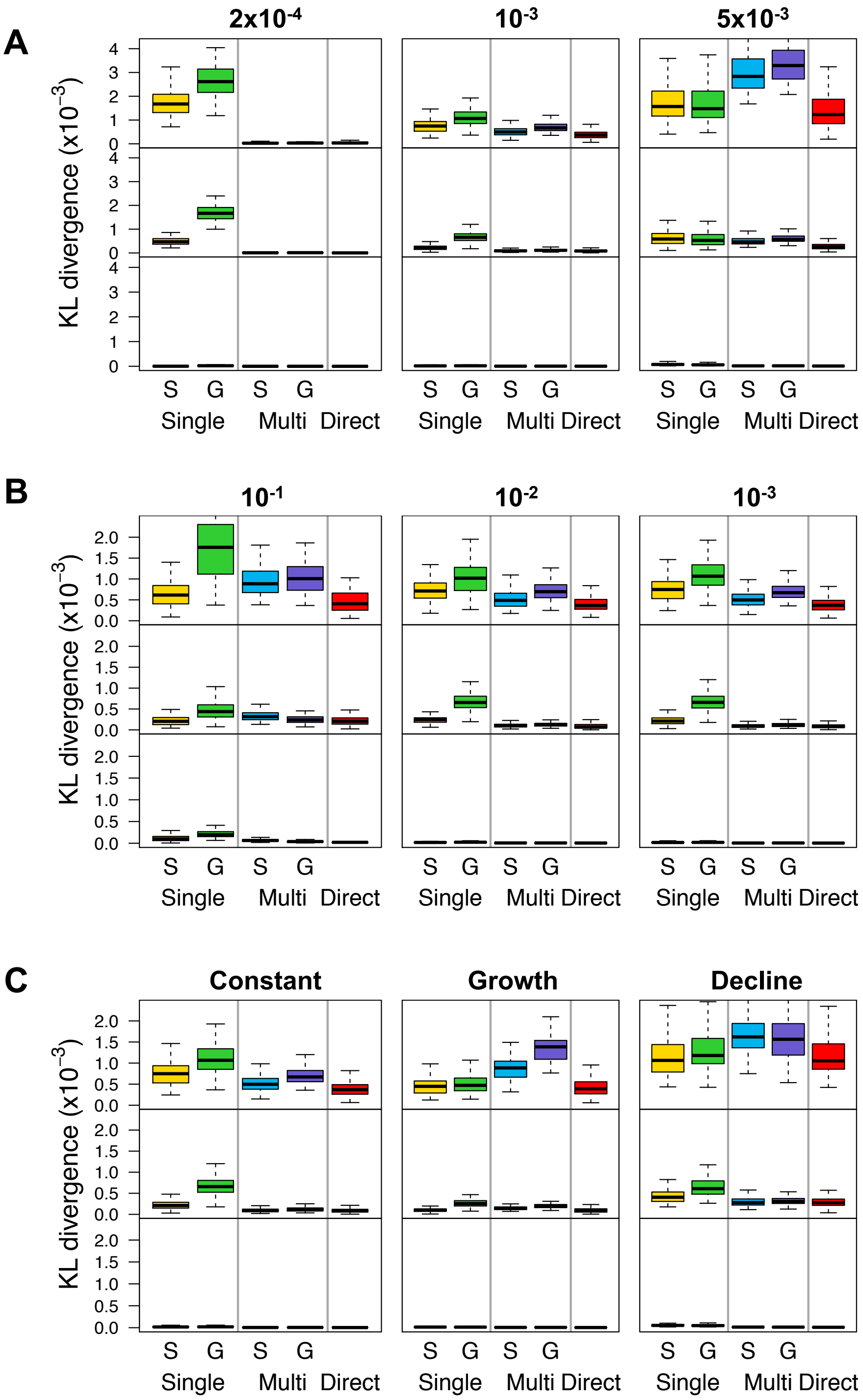


Figure 3.6: The effect of changes in parameters on the SFS estimation. The distance between the inferred and ground-truth SFS is measured by KL divergence. We modified the following parameter with others fixed as in figure 3.1. (A) Nucleotide diversity. (B) Sequencing error rate. (C) Underlying coalescent model.

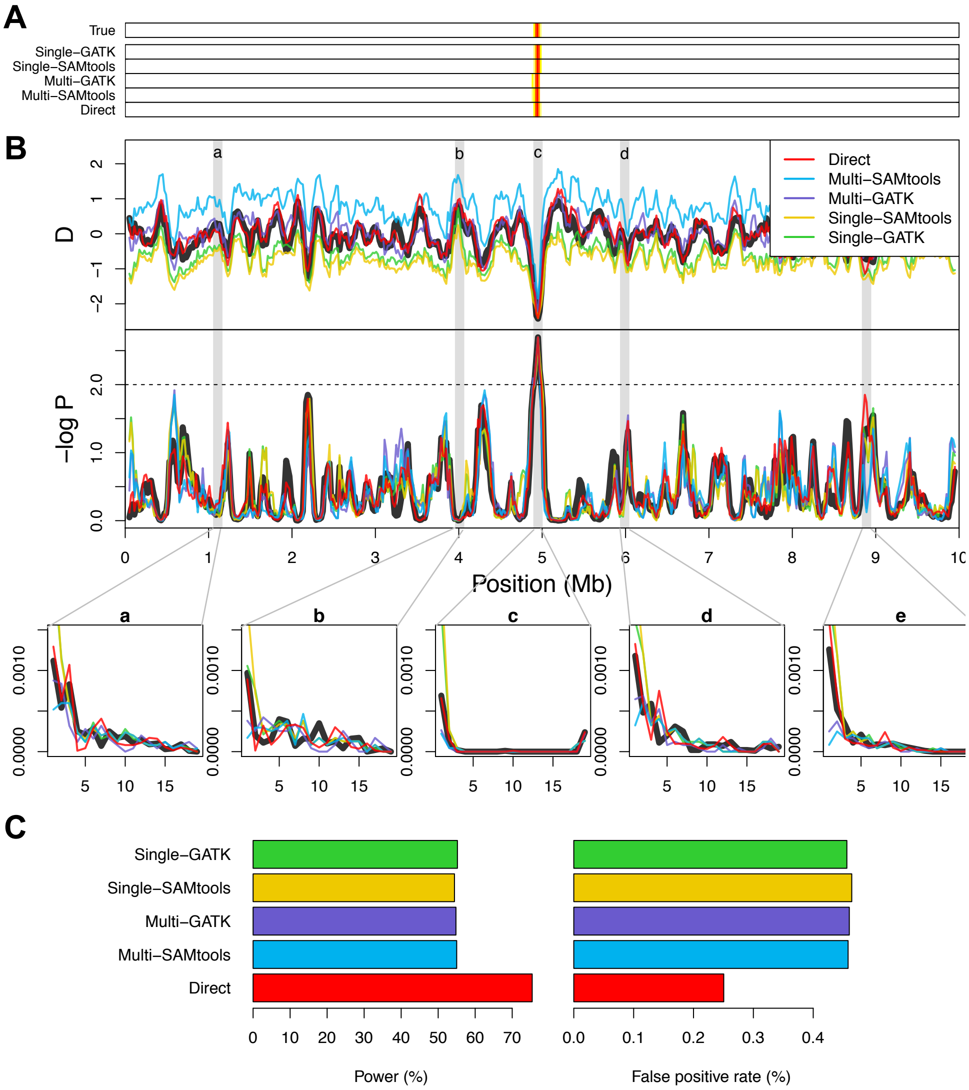


Figure 3.7: Genome-wide selection scans by an outlier detection approach based on 10Mb genomic region simulated with mean coverage 2X. (A) Classification of neutral vs. positively selected windows by empirical P values. The windows associated with empirical P values less than top 1% are shown with heat-map colors (the smaller the P value is, the more red the color is). These colored windows are candidate regions of having undergone positive selection. (B) Tajimas D (top), empirical P values (middle), and inferred SFS (bottom five panels) for 10 Mb genomic region. Bottom five panels corresponding to gray bars above (labeled ae) are chosen to contrast SFS patterns in neutral regions (panels a, b, d, and e) and those in positively selected regions (panel c). (C) Power and false position rates in percent for each of the five approaches.

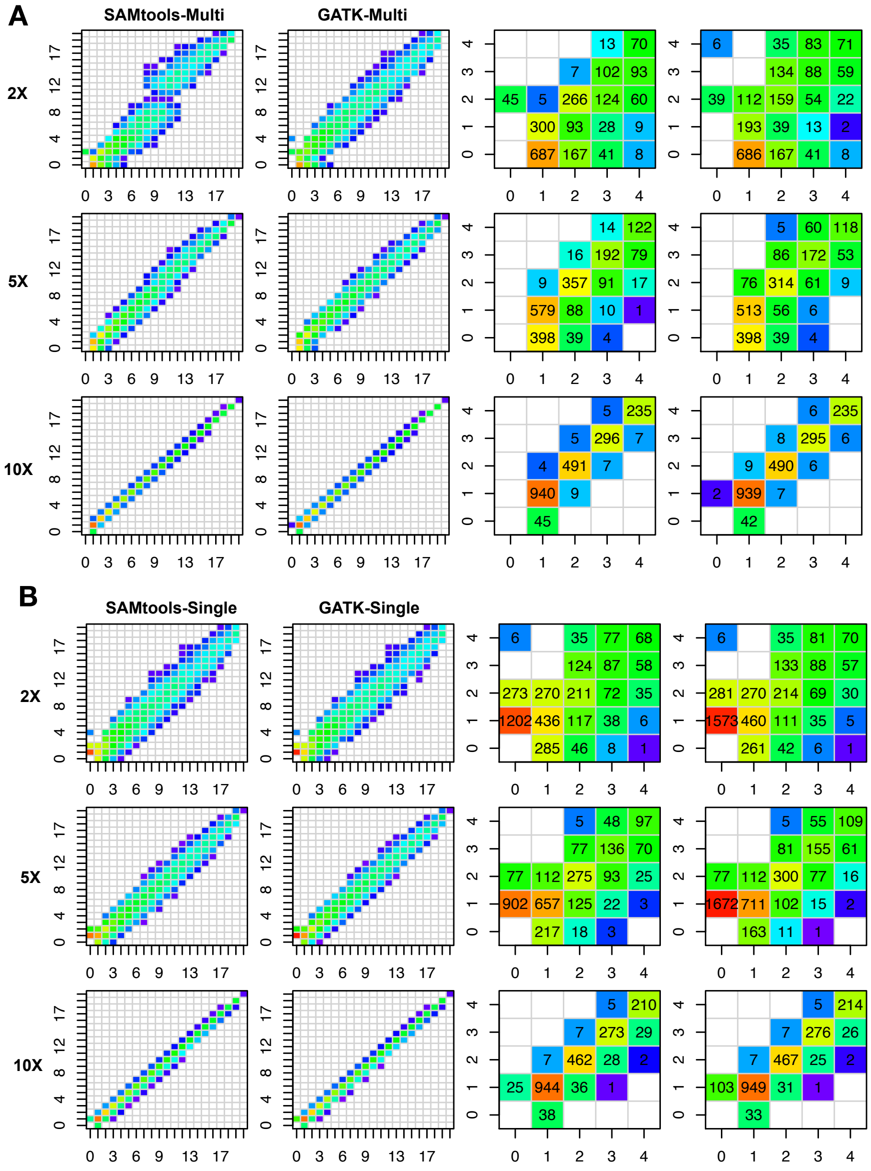


Figure 3.8: Error matrices where true allele counts are shown on x-axis and observed allele counts based on genotype calls are shown on y-axis. The SFS is inferred by either GATK or SAMtools with the multisample calling pipeline (A) or the single-sample calling pipeline (B). The first two columns show the error matrix with all allele counts and the next two columns show the error matrix with first five allele counts.

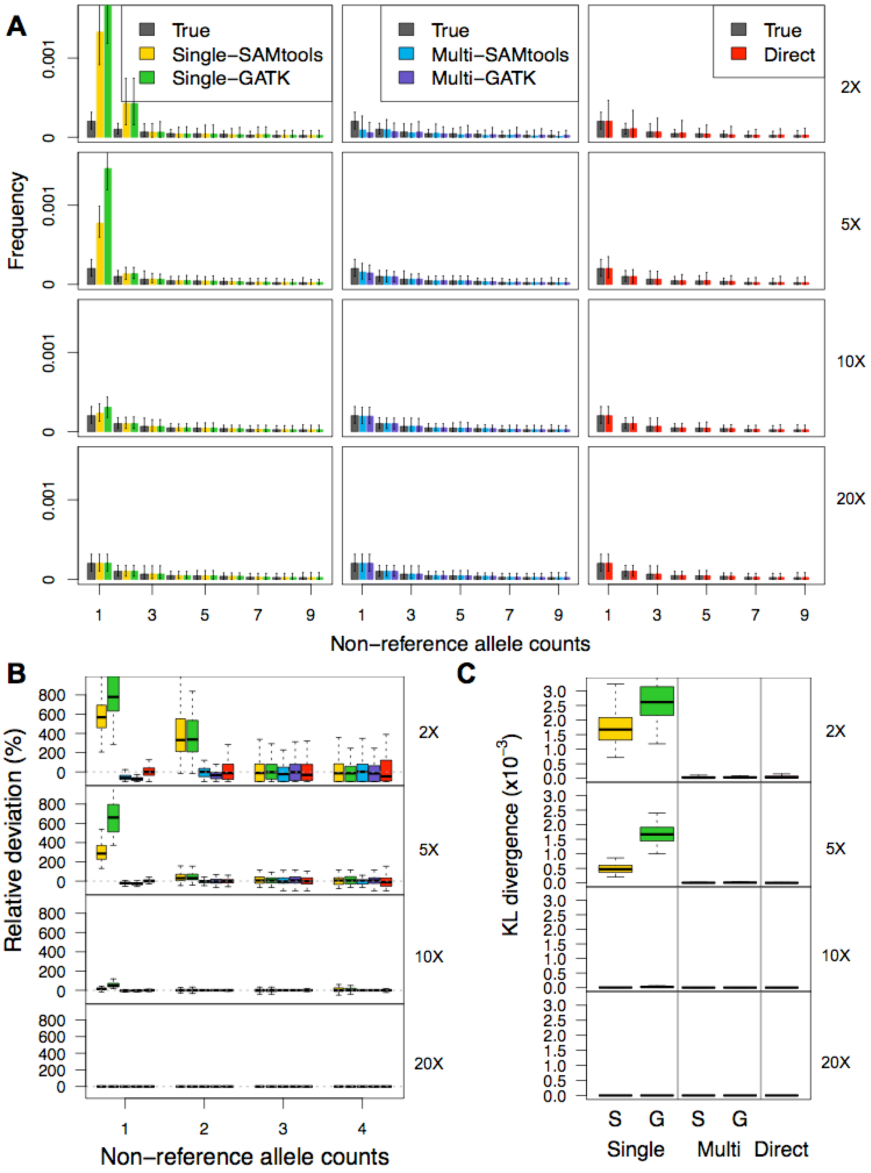


Figure 3.9: Evaluation of accuracy of inferred SFS by the call-based and direct estimation approach based on 100 replicates of genomic regions of length 100Kb when expected nucleotide diversity is 2×10^{-4} and sequencing error rate is 10^{-3}. A. Shapes of the inferred SFS (shown in colors in legend) compared to the ground-truth SFS (shown in grey) for coverage 2X, 5X, 10X, 20X, B. relative deviation of a fraction of sites with the non-reference allele counts of 1-4, C. a measure of a distance between the inferred and ground-truth SFS (KL divergence).

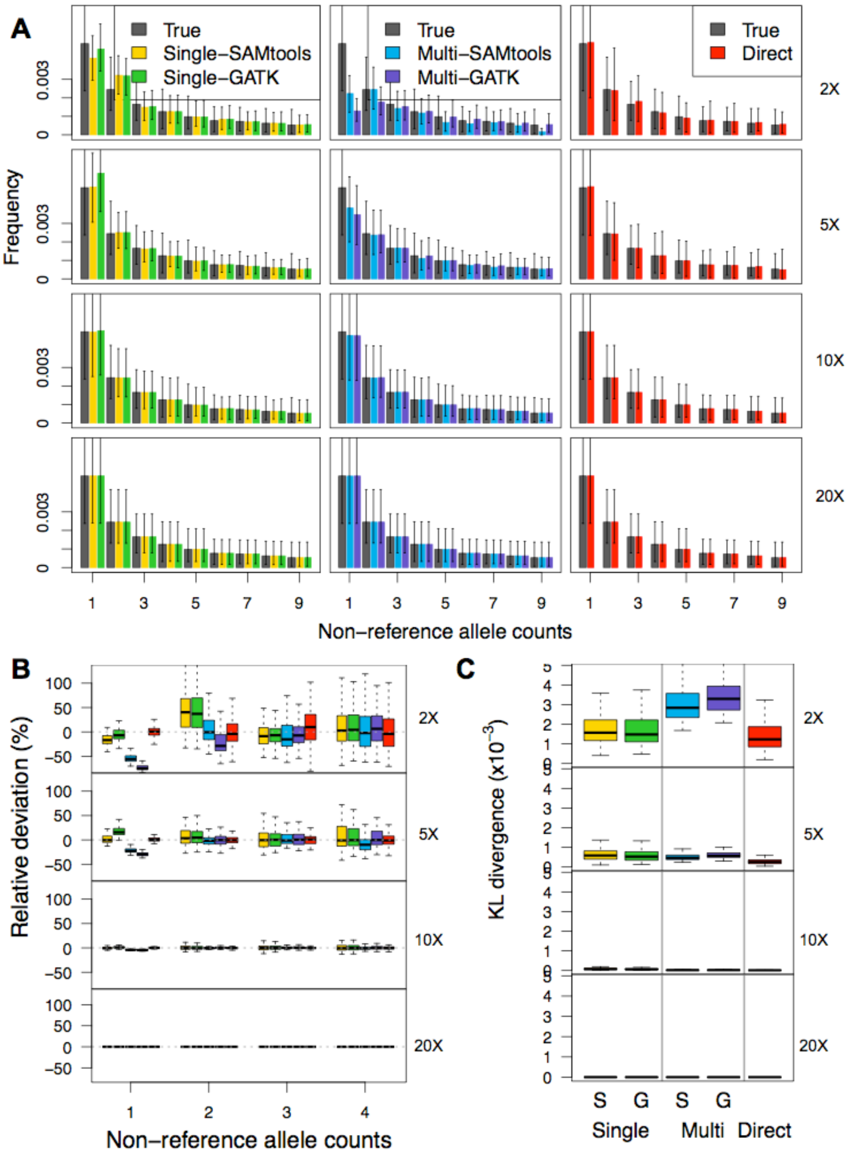


Figure 3.10: Evaluation of accuracy of inferred SFS by the call-based and direct estimation approach based on 100 replicates of genomic regions of length 100Kb when expected nucleotide diversity is 5×10^{-3} and sequencing error rate is 10^{-3}. A. Shapes of the inferred SFS (shown in colors in legend) compared to the ground-truth SFS (shown in grey) for coverage 2X, 5X, 10X, 20X, B. relative deviation of a fraction of sites with the non-reference allele counts of 1-4, C. a measure of a distance between the inferred and ground-truth SFS (KL divergence).

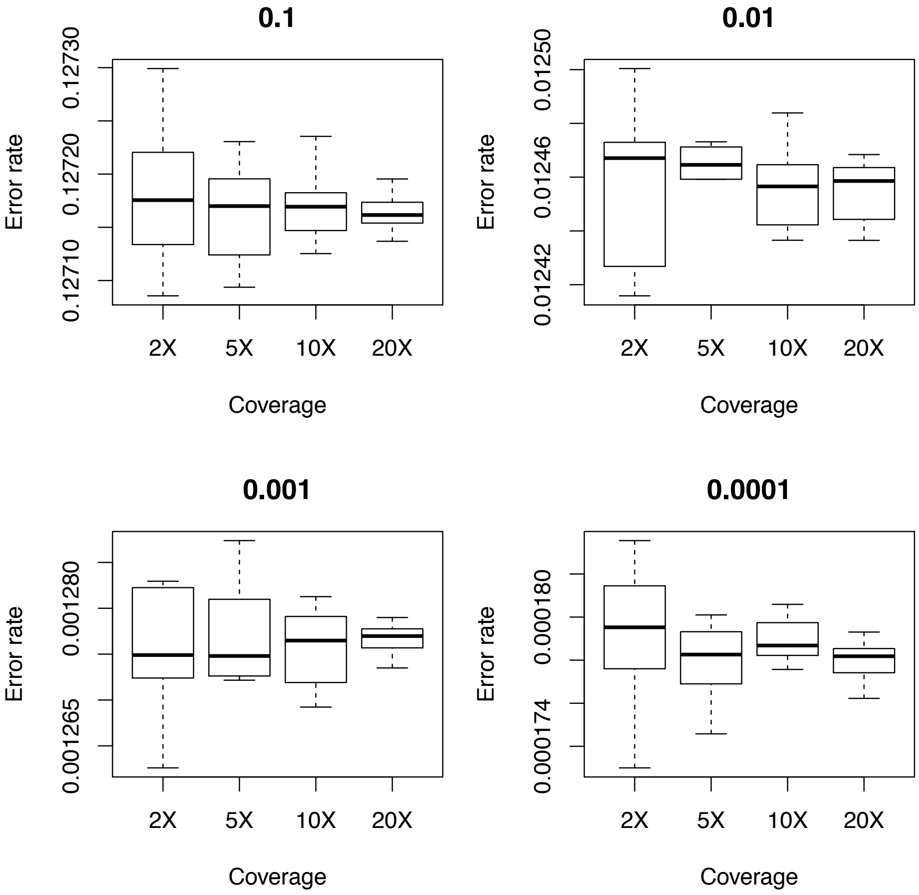


Figure 3.11: The distribution of observed error rates from sequencing experiment simulations given the sequencing error rate we used for simulations (shown in the title).

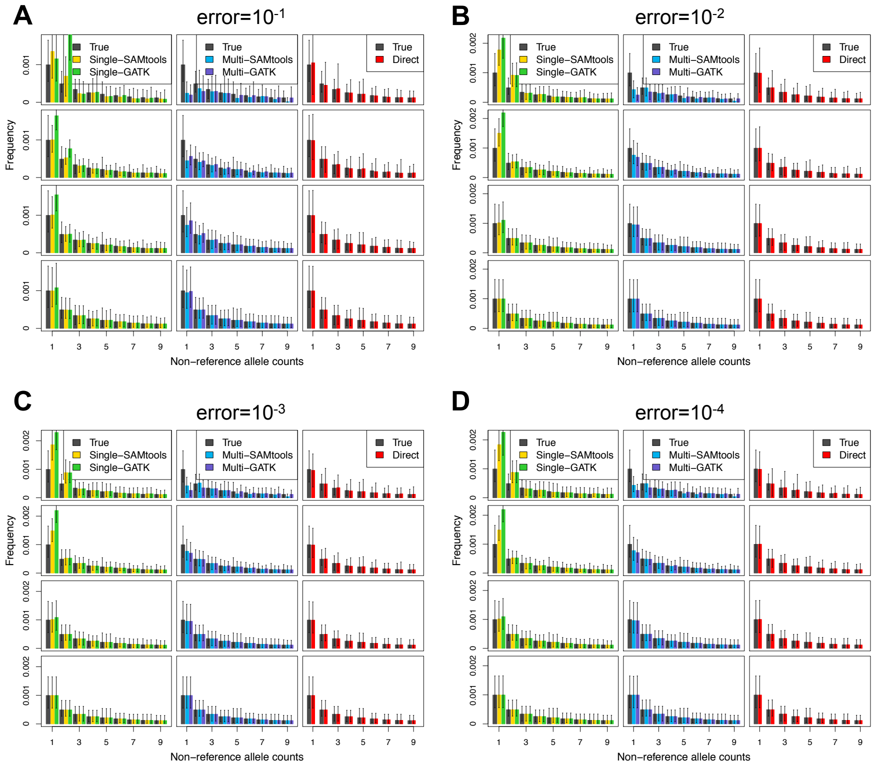


Figure 3.12: Shapes of the inferred SFS (shown in colors in legend) compared to the ground-truth SFS (shown in grey) for coverage 2X, 5X, 10X, 20X.

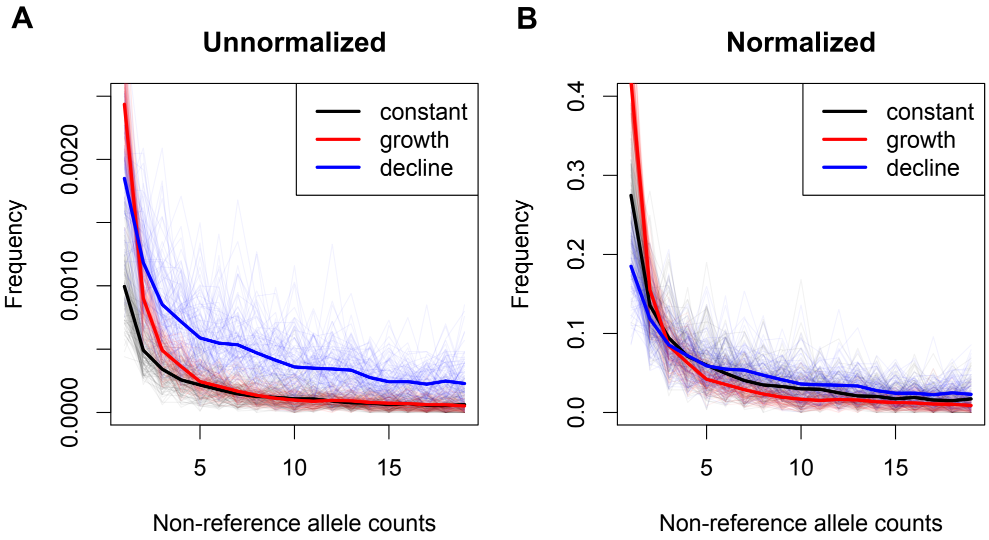


Figure 3.13: The SFS from simulated genotype data of length 100kb. Data are simulated with a constant population size (blue), an exponential population growth with a rate of 0.01% (red), and an exponential population decline with a rate of -0.01% (blue). For each scenario, the SFS from 100 replicates are shown with thin lines and the mean SFS over 100 replicates is shown with a bold line.

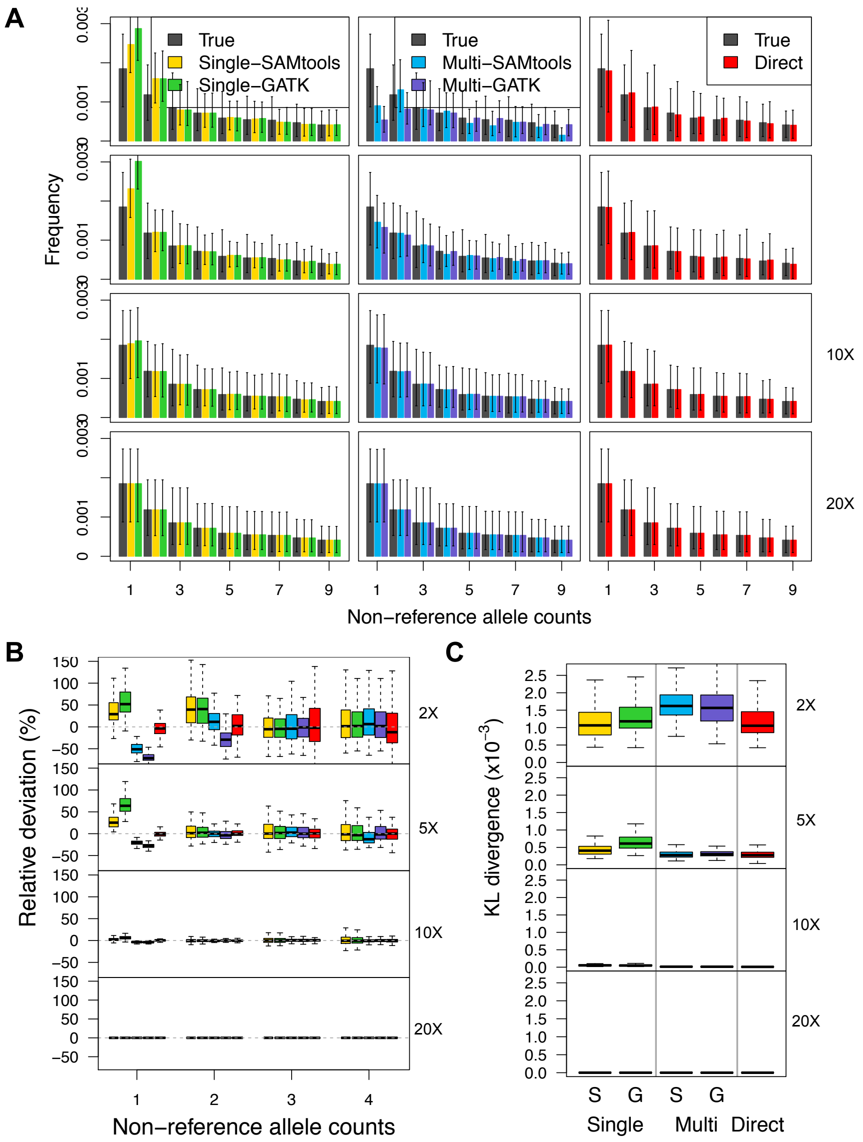


Figure 3.14: Evaluation of accuracy of inferred SFS by the call-based and direct estimation approach based on 100 replicates of genomic regions of length 100Kb under an exponential population decline model (rate=-0.01%). A. Shapes of the inferred SFS (shown in colors in legend) compared to the ground-truth SFS (shown in grey) for coverage 2X, 5X, 10X, 20X, B. relative deviation of a fraction of sites with the non- reference allele counts of 1-4, C. a measure of a distance between the inferred and ground-truth SFS (KL divergence).

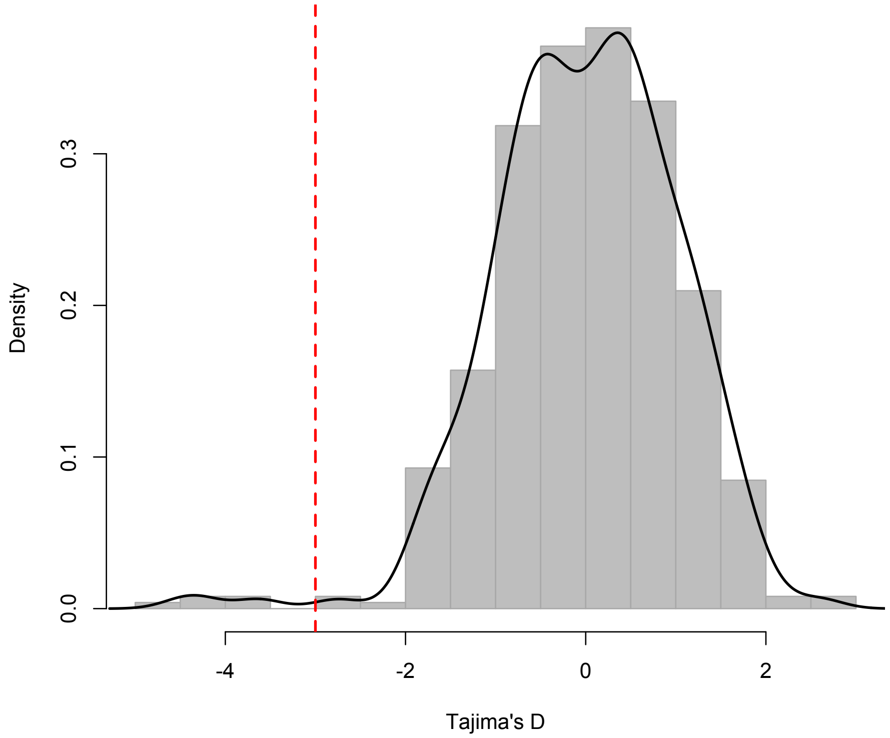


Figure 3.15: The empirical distribution of Tajimas D associated with sliding windows of length 100kb in a 10Mb genomic region. The dotted red line indicates top 1% (ranked in an increasing order).

CHAPTER 4

Fast and Accurate Site Frequency Spectrum Estimation from Low Coverage Sequence Data

4.1 Introduction

A site frequency spectrum (SFS) describes the distribution of allele frequencies across sites in the genome of a particular species. The SFS is of primary interest in population genetics, as it is a complete summary of sequence variation at unlinked sites and its shape reflects underlying population genetic processes, such as growth, bottlenecks and selection. Moreover, a number of population genetic inferences can proceed directly from the SFS. For example, demographic history (eg. evidence for population expansions, bottlenecks, or migrations) can be directly inferred from the SFS (using, for example, dadi (Gutenkunst *et al.*, 2009) or (Excoffier *et al.*, 2013). The SFS can also be compressed down to univariate summary statistic that form the basis of popular neutrality tests (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Achaz, 2008, 2009) that underlie many empirical genome-wide selection scans (Andolfatto, 2007; Begun *et al.*, 2007, e.g.). Hence, inferring the precise SFS from genetic data is crucial in many population genetic analyses.

With the recent rapid progress in sequencing techniques, obtaining large-scale genomic data from thousands to tens of thousands of individuals is practical (e.g. 1000 Genomes Project Consortium, 2010, 2012; Nelson *et al.*, 2012; Fu *et al.*, 2013) and this increased sample size enables us to conduct more accurate population genetic inference. However, current massively parallel short-read sequence technologies also pose many inherent challenges - for example, reads have high error rates, read mapping is sometimes uncertain, and coverage is variable and in many cases low or completely absent. These challenges make accurate individual-level genotype calls difficult and make some downstream analysis based on the inferred genotypes problematic.

In a previous study (Han *et al.*, 2014), we showed that the SFS computed from genotype calls (a *call-based* estimation approach) is biased at low to medium coverage ($\leq$ 10X), whereas the SFS directly inferred from aligned short-read sequencing data (a *direct* estimation approach) is unbiased even at low coverage. The direct estimation approach infers the maximum likelihood estimate (MLE) of the SFS by an EM algorithm (Li, 2011) or a Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Nielsen *et al.*, 2012) assuming independence across all individuals and sites. Both of these algorithms are implemented in the ANGSD software package (Nielsen *et al.*, 2012).

Both of these algorithms require computation of *site likelihood vectors* for all sites. These vectors contain the likelihood that an allele at a polymorphic site has each possible allele frequency conditional on observed sequence reads. Based on the precomputed site likelihood vectors, the MLE of the SFS is obtained by optimization, using either the EM (Li, 2011) or the BFGS algorithm (Nielsen *et al.*, 2012). The bottleneck in obtaining the MLE of the SFS is computing site likelihood vectors, rather than optimization. In fact, the maximization of the likelihood either by the EM or the BFGS algorithm takes only a small fraction of time compared to the computation of the site likelihood vectors. This is because computation of the site likelihood vector at each site requires a summation over all possible genotype combinations for n individuals and naive computation of this sum has a runtime complexity of $O(3^n)$. To overcome this computational burden, Li (Li, 2011) proposed a dynamic programming (DP) algorithm to effectively compute the site likelihood vector for each site in $O(n^2)$ and Nielsen and co-workers (Nielsen *et al.*, 2012) implemented this algorithm in the ANGSD software. However, this algorithm is still not practical to use if there are large numbers of individuals, because it is quadratic in the number of genomes (see Figure 4.6B for runtime). Moreover, this algorithm is numerically unstable for a large sample (Li, 2011). To solve this problem of computational inefficiency and numerical instability, we compute site likelihood vectors in a more efficient way that still retains the accuracy of the original DP algorithm. Our new method uses a combination of rescaling and sensible approximation to compute the site likelihood vector.

4.2 Approach

To establish notation and background, we first review the existing DP algorithm implemented in the ANGSD software (Nielsen *et al.*, 2012) and then introduce our approach.

4.2.1 Dynamic Programming Algorithm used by ANGSD

Let D denote the short-read sequencing data and X represent a total count of the derived allele for a sample of n diploid individuals at a particular site. The corresponding site likelihood vector,= $\mathbf{h} = (h_0, h_1, \ldots, h_{2n})$, is a $(2n+1)$-dimensional vector in which each element $h_x(= P\{D|X = x\})$ is the site likelihood function for the derived allele frequency of $x/(2n)$ in the sample:

$$h_x = \frac{1}{\binom{2n}{x}} \sum_{g_1=0}^{2} \cdots \sum_{g_n=0}^{2} I\left(\sum_{k=1}^{n} g_k = x\right) \prod_{k=1}^{n} \binom{2}{g_k} L_{g_k}^{k} \tag{4.1}$$

where $I()$ is an indicator function and $L_{g_k}^{k} = P(D_k|G_k = g_k)$ is a genotype likelihood of the individual k for genotype g_k.

To calculate the site likelihood vector $\mathbf{h}$ efficiently, define a $(2j+1)$-dimensional *raw* site likelihood vector for j individuals, given by $\mathbf{z^j} = (z_0^j, z_1^j, \ldots, z_{2j}^j)$, in which each element is defined as

$$z_x^j = \sum_{g_1=0}^{2} \cdots \sum_{g_j=0}^{2} I\left(\sum_{k=1}^{j} g_k = x\right) \prod_{k=1}^{j} \binom{2}{g_k} L_{g_k}^{k} \tag{4.2}$$

where $j = 1, \ldots, n$ and $x = 1, \ldots, 2j$. Note that this expression does not include a rescaling factor $1/\binom{2n}{x}$.

The vector $\mathbf{z^j}$ can be iteratively updated from the vector $\mathbf{z^{j-1}}$ (raw site likelihood vector for $j-1$ individuals) by the following recurrence relation:

$$z_x^j = L_0^j z_x^{j-1} + 2L_1^j z_{x-1}^{j-1} + L_2^j z_{x-2}^{j-1}. \tag{4.3}$$

In a final step, each element of the vector $\mathbf{z^n}$ is *rescaled* by a corresponding factor $1/\binom{2n}{x}$ to obtain the vector $\mathbf{h}$ (i.e. $h_x = z_x^n/\binom{2n}{x}$), and then because likelihoods need only be defined proportional to a constant, the resulting vector $\mathbf{h}$ is *standardized* such that the maximum element of the vector becomes one.

To illustrate the procedure, we show how the raw site likelihood vector is recursively updated from $\mathbf{z^1}$ to $\mathbf{z^n}$ by the DP algorithm in the ANGSD software. Each row in a lower triangular matrix in Figure 4.1A (top) represents the raw site likelihood vector for j individuals of length $2j+1$. Figure 4.1A also shows how the raw site likelihood vector $\mathbf{z^n}$ (middle) is converted to the final site likelihood vector $\mathbf{h}$ after rescaling by $1/\binom{2n}{x}$ and standardization (bottom).

4.2.2 Rescaled Dynamic Programming Algorithm

In our preliminary work, we observed that the value at the mode of $\mathbf{z^n}$ can be relatively large. In this example with 50 diploid individuals, the mode of $\mathbf{z^n}$ is 1531. With 500 diploid individuals, the mode of $\mathbf{z^n}$ can be about 8×10^{12} (data not shown). This implies that the DP algorithm can have a overflow problem for large samples because the mode of $\mathbf{z^n}$ increases exponentially as a sample size increases. Furthermore, the value at edges of $\mathbf{z^n}$ is very small. In this example with 50 individuals, the value of the site likelihood function for the allele count of 100 is 10^{-200}. With 500 diploid individuals, the value of the site likelihood function for the allele count of 1000 is smaller than 10^{-300} (data not shown). This implies that the DP algorithm can have an underflow problem for large samples because the values at the edge of $\mathbf{z^n}$ keeps decreasing exponentially

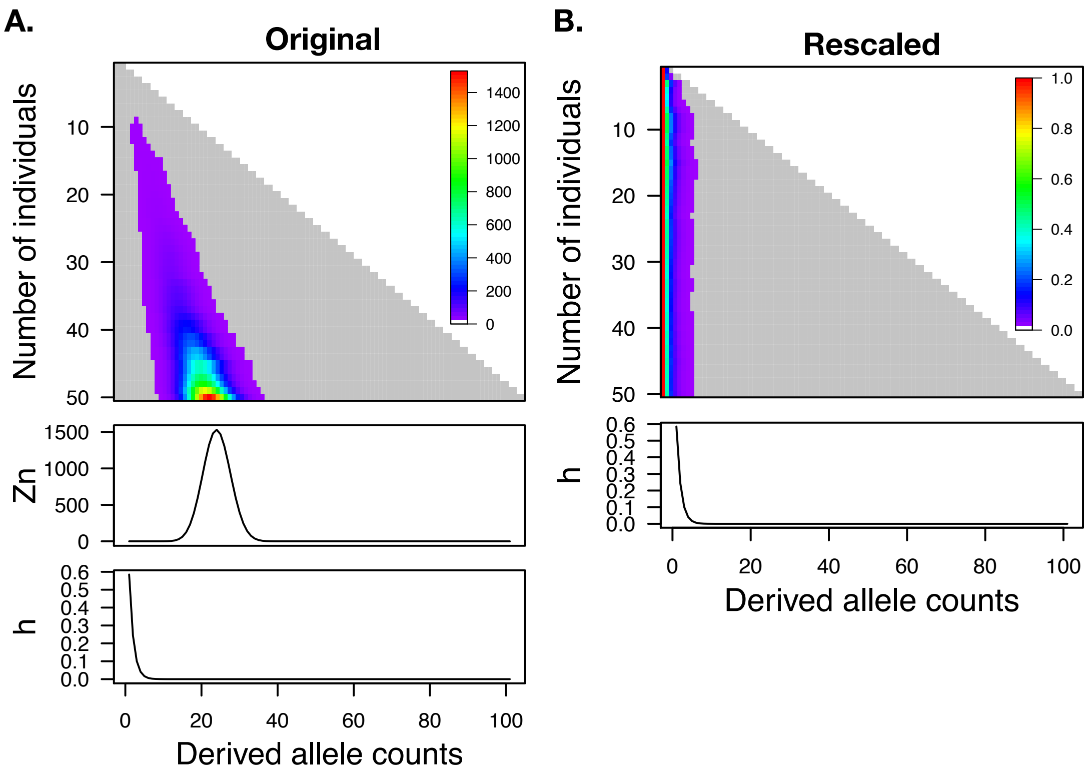


Figure 4.1: Updating the site likelihood vector for 50 diploid individuals at a particular site fixed for an ancestral allele by the DP (A) or the rescaled DP (B) algorithm. The sequencing data was simulated at coverage 3X with error rate of 0.001. Genotype likelihoods were calculated using a GATK model. **A**. Top row shows how the raw site likelihood vector is recurrently updated by the original DP algorithm. Each row in a low triangular matrix represents the raw site likelihood vector $\mathbf{z^j}$ for j individuals. Middle and bottom rows show how the raw site likelihood vector $\mathbf{z^n}$ (the last row of the lower triangular matrix) is converted to the final site likelihood vector $\mathbf{h}$ by rescaling and standardization. **B**. Top row shows how the rescaled site likelihood vector is recurrently updated by the rescaled DP algorithm. Each row in the lower triangular matrix represents the rescaled site likelihood vector $\mathbf{h^j}$ for j individuals. Bottom row shows the final site likelihood vector $\mathbf{h}$. Note that the final site likelihood vector $\mathbf{h}$ has a peak at the derived allele count of zero. The gray area represents the range of values used for the original and the rescaled DP algorithm. In this example, it requires to compute 2600 elements (because $3 + 5 + \cdots + 101 = \sum_{i=1}^{50} 2i + 1$) to update the site likelihood vector.

as a sample size increases. Therefore, the DP algorithm can be numerically unstable because $\mathbf{z^n}$ has both overflow and underflow problems for large samples.

To overcome the numeric instability problem of the DP algorithm, we modified the DP algorithm such that rescaling and standardization take place at each step of updating the site likelihood vector. For this modified algorithm, we define a $(2j+1)$-dimensional *rescaled* site likelihood vector for j individuals, $\mathbf{h^j} = (h_0^j, h_1^j, \ldots, h_{2j}^j)$, of which each element is defined as

$$h_x^j = \frac{1}{\binom{2j}{x}} \sum_{g_1=0}^{2} \cdots \sum_{g_j=0}^{2} I\left(\sum_{k=1}^{j} g_k = x\right) \prod_{k=1}^{j} \binom{2}{g_k} L_{g_k}^k \tag{4.4}$$

where $j = 1, \ldots, n$ and $x = 1, \ldots, 2j$. We can derive a recurrence relation to iteratively update the vector $\mathbf{h^j}$ from the vector $\mathbf{h^{j-1}}$ (rescaled site likelihood vector for $j-1$ individuals) as follows:

$$\begin{aligned} h_x^j &= \frac{\binom{2(j-1)}{x} L_0^j h_x^{j-1} + 2\binom{2(j-1)}{x-1} L_1^j h_{x-1}^{j-1} + \binom{2(j-1)}{x-2} L_2^j h_{x-2}^{j-1}}{\binom{2j}{x}} \\ &= \frac{1}{2j(2j-1)}\{(2j-x)(2j-x-1)L_0^j h_x^{j-1} + 2x(2j-x)L_1^j h_{x-1}^{j-1} + x(x-1)L_2^j h_{x-2}^{j-1}\}. \end{aligned} \tag{4.5}$$

Because the constant $\frac{1}{2j(2j-1)}$ in both numerator and denominator in Equation (4.5) is cancelled out during standardization, we can use the following recurrence equation to update the rescaled site likelihood vector:

$$h_x^j = (2j-x)(2j-x-1)L_0^j h_x^{j-1} + 2x(2j-x)L_1^j h_{x-1}^{j-1} + x(x-1)L_2^j h_{x-2}^{j-1}. \tag{4.6}$$

Figure 4.1B shows how the site likelihood vector is recurrently updated from $\mathbf{h^1}$ to $\mathbf{h^n}$ by the rescaled DP algorithm in a lower triangular matrix (Figure 4.1B, top) and the final site likelihood vector $\mathbf{h}$ (Figure 4.1B, bottom). Now, all values in the intermediate site likelihood vectors $\mathbf{h^j}$ range between zero and one, suggesting there will be no potential overflow problem. Importantly, we observed that the most of the cells in the intermediate site likelihood vector have a value close to zero (shown in gray). This implies that computing all values of the site likelihood vector is inefficient and we can accurately approximate this vector by only computing the first few elements and setting the rest of the elements to zero. This motivated the development of the adaptive K-restricted algorithm.

4.2.3 Adaptive K-restricted Algorithm

For a site that is fixed for the ancestral allele, we observe that all non-negligible values of the rescaled site likelihood vectors, $\mathbf{h^1}$ to $\mathbf{h^n}$, are consistently concentrated on the first few cells, and the final site likelihood vector $\mathbf{h}$ has a peak at the allele frequency of zero (Figure 4.2A, middle). For a site that is polymorphic, we observe that the mode typically stays at zero when we add an individual whose best-guess genotype is 0/0 (i.e. the genotype likelihood vector of that individual has the highest value at genotype 0/0), whereas the mode of the site likelihood vector typically moves to the right when we add an individual whose best-guess genotype is 0/1 or 1/1 (Figure 4.2B, middle). If we add an individual whose best-guess genotype is 0/1, the mode moves one bin to the right and the best-guess allele count increases by one. By the same token, if we add an individual whose best-guess genotype is 1/1, the mode typically moves two bins to the right and the best-guess allele count increases by two.

Based on these observations, we propose a new algorithm, called the adaptive K-restricted algorithm. This algorithm first proposes left and right boundaries within which we update values of the site likelihood vector and outside of which we set the values of the site likelihood vector to zero. We can update left and right boundaries using the best-guess genotype. For the individuals whose best-guess genotype is 0/0 we do not change the boundaries. For individuals whose best-guess genotype is 0/1 we move both boundaries one bin to the right, and for the individuals whose best-guess genotype is 1/1 we move both boundaries two bins to the right. Next, we check whether the boundary values are greater than a very small value ϵ (for example, we set $\epsilon = 10^{-9}$) and if so, we expand the appropriate boundary accordingly until the new boundary value is less than or equal to ϵ. For example, we check the value at the left boundary and if the

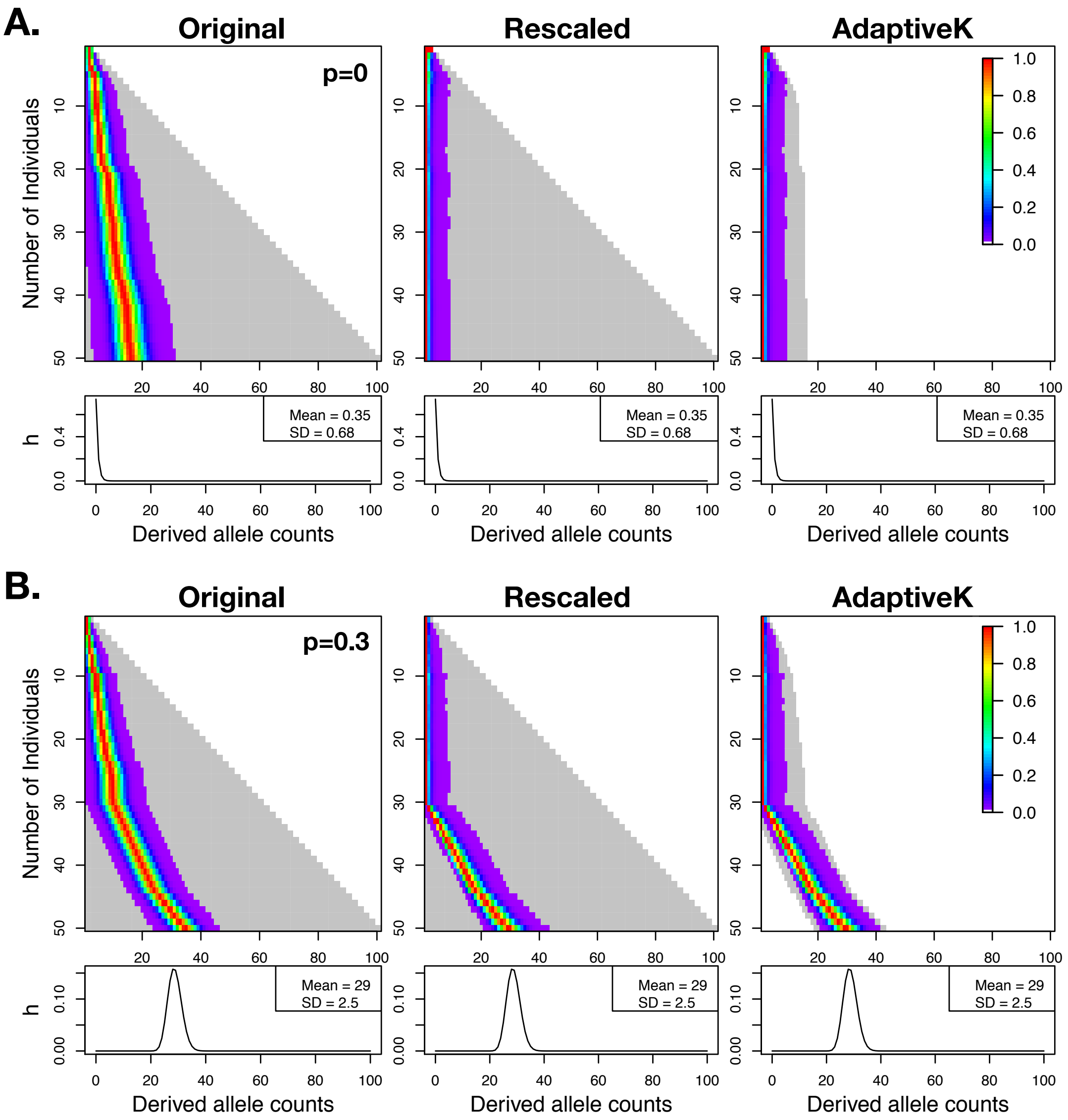


Figure 4.2: Updating the site likelihood vector for 50 diploid individuals by the original DP (referred to as Original), the rescaled DP (referred to as Rescaled), and the adaptive K-restricted (referred to as AdaptiveK) algorithm. The sequencing data was simulated at coverage 3X with error rate of 0.001. **A**. A random site fixed for the ancestral allele is chosen. **B**. A random site with the true derived allele frequency of 0.3 is chosen. Each row in the lower triangular matrix represents the intermediate site likelihood vector for j individuals. For a plotting purpose, each row is standardized such that the maximum elements is assigned to one. The gray area represents the range of values used for all three algorithms.

value is greater than ϵ, we move the left boundary one bin to the left and check the value at the updated left boundary again. By the same token, we check the value at the right boundary and if the value is greater than ϵ, we move the right boundary one bin to the right. By doing this, at each step of calculating the intermediate site likelihood vector, we only compute k elements of the vector, where k is the number of elements between the left and right boundary. Note that k is dynamically changing at each updating step, but k is always much smaller than $2n + 1$. Therefore, we can update the site likelihood vector in a linear fashion (computing at most K values at each updating step, where K is the maximum value of all k's) rather than updating it in a triangular fashion (computing $3 + 5 + \cdots + (2n + 1)$ values). This makes computation time close to $O(Kn)$ rather than original $O(n^2)$.

Figure 4.2 shows with an example how that the adaptive K-restricted algorithm captures the important regions of the intermediate site likelihood vector. Hence, the adaptive K-restricted algorithm is faster than the original algorithm, as reflected by the reduced computation area (shown in gray in Figure 4.2). Moreover, we retain the accuracy of the final site likelihood vector $\mathbf{h}$ with the adaptive K-restricted algorithm and it is as stable as the rescaled DP algorithm. The shape of the distribution $\mathbf{h}$ is identical in all three cases (Original, Rescaled, AdaptiveK), reflected by the same mean and variance of $\mathbf{h}$ in all three cases (Figure 4.2).

4.3 Methods

4.3.1 Generating Simulated Sequences

To compare the three algorithms (original DP, rescaled DP, and adaptive K-restricted algorithm) for computing site likelihood vectors from NGS data, we generated aligned short-read sequencing data by changing sequencing coverage (3X, 5X, and 10X) and sample size (50, 100, 300, 500, and 1000 diploid individuals). For this purpose, we first conducted population genetic simulations to produce haplotype data of a given sample size assuming the standard model (with an effective population size of 10,000 diploid individuals, a mutation rate per-base per-generation of 2.5×10^{-8} and a recombination rate of 10^{-8}), and then overlaid sequencing errors (with error rate of 0.001) to generate paired-end short-read sequencing data given sequencing coverage. For detailed descriptions of the coalescent and sequencing simulations, refer Material and Methods section in Han *et al.* (2014).

4.3.2 Sequencing Data from the 1000 Genomes Project

To demonstrate the adaptive K-restricted algorithm's utility with real data, we downloaded the VCF file and the BAM files from the 1000 Genomes Project FTP site in order to estimate the SFS. We used the genotype calls of 365 European and 228 sub-Saharan African individuals from the VCF file, which contains the genotype calls for 1,092 individuals sampled from 14 populations drawn from Europe, East Asia, sub-Saharan Africa and the Americas (1000 Genomes Project Consortium, 2010, 2012). For the BAM files, we only used low-coverage Illumina sequencing data (coverage 2X to 4X) (1000 Genomes Project Consortium, 2010, 2012) for these same individuals. Due to file size constraints, we downloaded only a subsection of the genome (region of 10Mb-20Mb in chromosome 10) by using SAMtools (version 0.1.18) (Li *et al.*, 2009a).

4.3.3 Estimating the SFS

To infer the SFS from simulated aligned short-read sequencing data, we used the direct estimation approach using the freely available program ANGSD (version 0.588) with the EM algorithm option to obtain the MLE of the SFS (Nelson *et al.*, 2012). We refer to results of this procedure as Original. Then, we modified the source code of ANGSD to implement the rescaled DP (referred to as Rescaled) and the adaptive K-restricted algorithm (referred to as AdaptiveK). All code is written in C++.

For the 1000 Genomes project data, we evaluated two approaches to infer the SFS: the call-based and direct estimation approaches. For the call-based estimation approach, we used genotype calls in the VCF file and then reconstructed the SFS by allele counting using vcfTools (version 0.1.10) (Danecek *et al.*, 2011). For the direct estimation approach, we directly estimated the SFS from the BAM files with the adaptive K-restricted algorithm.

To evaluate the accuracy of the SFS estimated from simulated short-read sequencing data, we computed the relative deviation of the inferred SFS (computed from sequencing data) compared to the ground-truth SFS (computed from the known values for the genotype data) in each derived allele frequency bin $i/(2n)$:

$$\text{Relative deviation } (\frac{i}{2n}) = \frac{f_{seq}(\frac{i}{2n}) - f_{true}(\frac{i}{2n})}{f_{true}(\frac{i}{2n})}$$

where $f_{seq}(\frac{i}{2n})$ represents a fraction of sites with a derived allele frequency $i/(2n)$ in the inferred SFS and $f_{true}(\frac{i}{2n})$ represents a fraction of sites with a derived allele frequency $i/(2n)$ in the ground-truth SFS.

4.4 Results

We evaluated whether the adaptive K-restricted algorithm is robust to different sequencing coverage and the variance in the site likelihood vector. This evaluation is important because one of the characteristics of the next-generation sequencing data is variable coverage across sites, which affects the variance of the genotype likelihood vector at the individual-level and the variance of the site likelihood vector at the sample-level.

4.4.1 Performance for changing sequencing coverage

First, we investigated the impact of different sequencing coverage on the performance of the adaptive K-restricted algorithm in computing the site likelihood vector. For this purpose, we simulated sequencing data for 50 diploid individuals under the standard model at coverage 3X, 5X, and 10X. Figure 4.3 shows how the site likelihood vector is updated at a random site with true allele frequency of 0.3 as a function of sequencing coverage. We observed that the site likelihood vector $\mathbf{h}$ is more diffuse as coverage decreases, whereas it is more peaked around the true allele frequency of 0.3 as coverage increases (Figure 4.3B, the variance of the site likelihood vector $\mathbf{h}$ is 6.2 with 3X, 2.7 with 5X, and 0.1 with 10X). This is because the genotype likelihood vectors tend to be more spread out at low coverage, whereas they tend to be more peaked at the unknown individual genotype at high coverage. This implies that the choice of K for the adaptive K-restricted algorithm should depend on coverage - the higher coverage, the smaller K. We observed that this is automatically done with the adaptive K-restricted algorithm. Furthermore, the resulting site likelihood vector computed by the adaptive K-restricted algorithm has a comparable accuracy to the site likelihood vector computed by the original DP algorithm across all coverage (Figure 4.3). The shape of the distribution $\mathbf{h}$ is the same, with the same mean and standard deviation of $\mathbf{h}$, for all three algorithms across coverage.

4.4.2 Performance for variation in the site likelihood vector

Next, we evaluated whether the adaptive K-restricted algorithm can capture the site-to-site variation in the site likelihood vector. For this purpose, we used low-coverage sequencing data for 50 diploid GBR individuals in the 1000 Genome Project, and then compared the site likelihood vector computed by the three algorithms (Original, Rescaled, AdaptiveK) at multiple random sites with the same best-guess allele frequency in the sample. Compared to the simulated sequencing data matched at average coverage (5X), we observed that low-coverage sequencing data in the 1000 Genomes Project tend to have bigger site-to-site variation of the site likelihood vector. Figure 4.4 shows how the site likelihood vector $\mathbf{h}$ is updated at two random sites with the best-guess allele frequency of 0.5 in the sample. We observed that the first site (position 10,085,321 in chromosome 10) has a smaller variance in the site likelihood vector than the second site (position 10,012,499 in chromosome 10). The variance of the site likelihood at the first

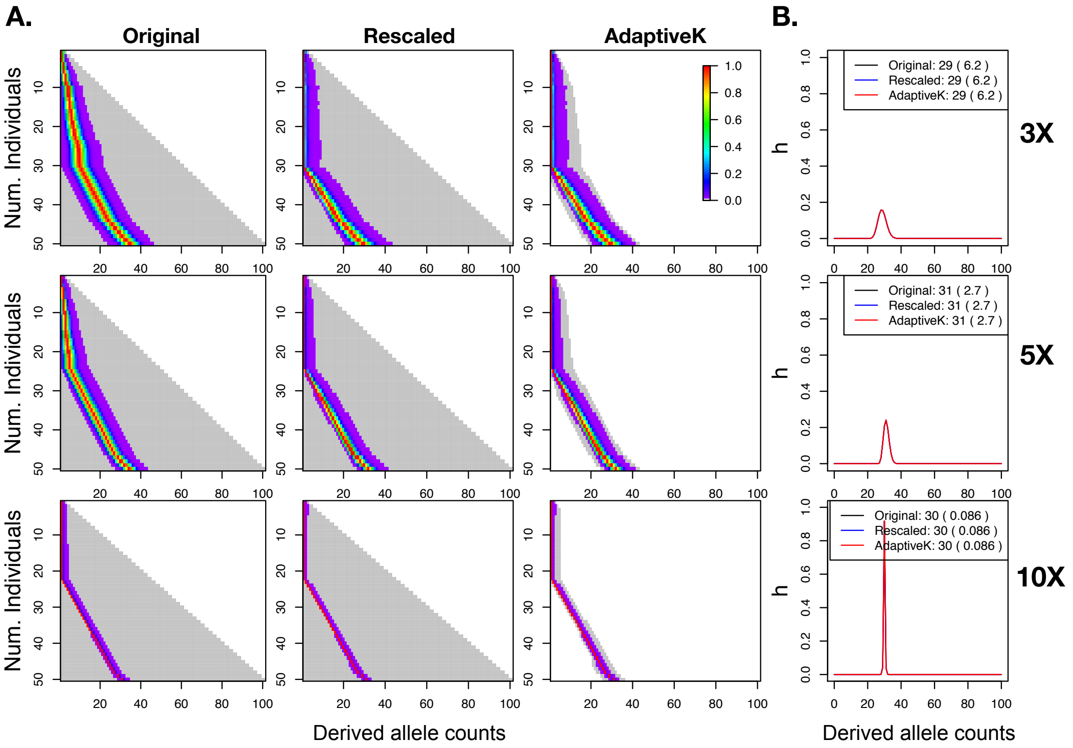


Figure 4.3: Performance of the three algorithms (Original, Rescaled, and AdaptiveK) for updating the site likelihood vector for 50 individuals as a function of sequencing coverage. The sequencing data was simulated at coverage 3X (top), 5X (middle), and 10X (bottom) with error rate of 0.001, and a site with the true allele frequency of 0.3 is randomly picked. **A**. It shows how the site likelihood vector is updated for 50 individuals. Each row of the lower triangular matrix represents the site likelihood vector for j individuals. The gray area represents the range of values used for each algorithm. **B**. The final site likelihood vector **h** by all three algorithms is shown in black (Original), blue (Rescaled), and red (AdaptiveK). Note that all three distributions almost completely overlap, and the mean and variance are the same for all three distributions.

site is 1.2, whereas that at the second site is 3.5 (Figure 4.4). This implies that the adaptive K-restricted algorithm should be capable of changing K according to the observed variance of the site likelihood vector at different sites - the larger the variance of the site likelihood vector, the larger the value of K. Moreover, the resulting site likelihood vector computed by the adaptive K-restricted algorithm has a comparable accuracy to the site likelihood vector computed by the original DP algorithm - same shape, and the mean and standard deviation with the three algorithms (Figure 4.4)

4.4.3 Evaluating the accuracy of the inferred SFS

We evaluate the accuracy of the inferred SFS by the adaptive K-restricted algorithm (Adaptive K) compared with the inferred SFS using the original DP (Original). For this comparison, we simulated 100 replicates of sequencing data for 100, 300, and 500 diploid individuals each from genomic regions of length 100Kb under the standard model. The accuracy of the inferred SFS was evaluated by two metrics: 1) the shape of the inferred SFS in comparison to the ground-truth SFS (Figure 4.5A) and 2) the relative deviation of the inferred SFS compared to the ground-truth SFS at each allele frequency bin (Figure 4.5B).

We found that the adaptive K-restricted algorithm behaves equivalently to the original DP algorithm. We observed the identical shape of the inferred SFS (Figure 4.5) with both algorithms. Moreover, consistent with the previous study (Han et al. 2014), both algorithms led to unbiased estimates of the SFS even at low coverage (such as 3X) regardless of sample sizes. The shape of the inferred SFS was similar to the ground-truth SFS (Figure 4.5A) and the relative deviation of the inferred SFS was close to zero in all allele frequency bins (Figure 4.5B) across all sequencing coverage.

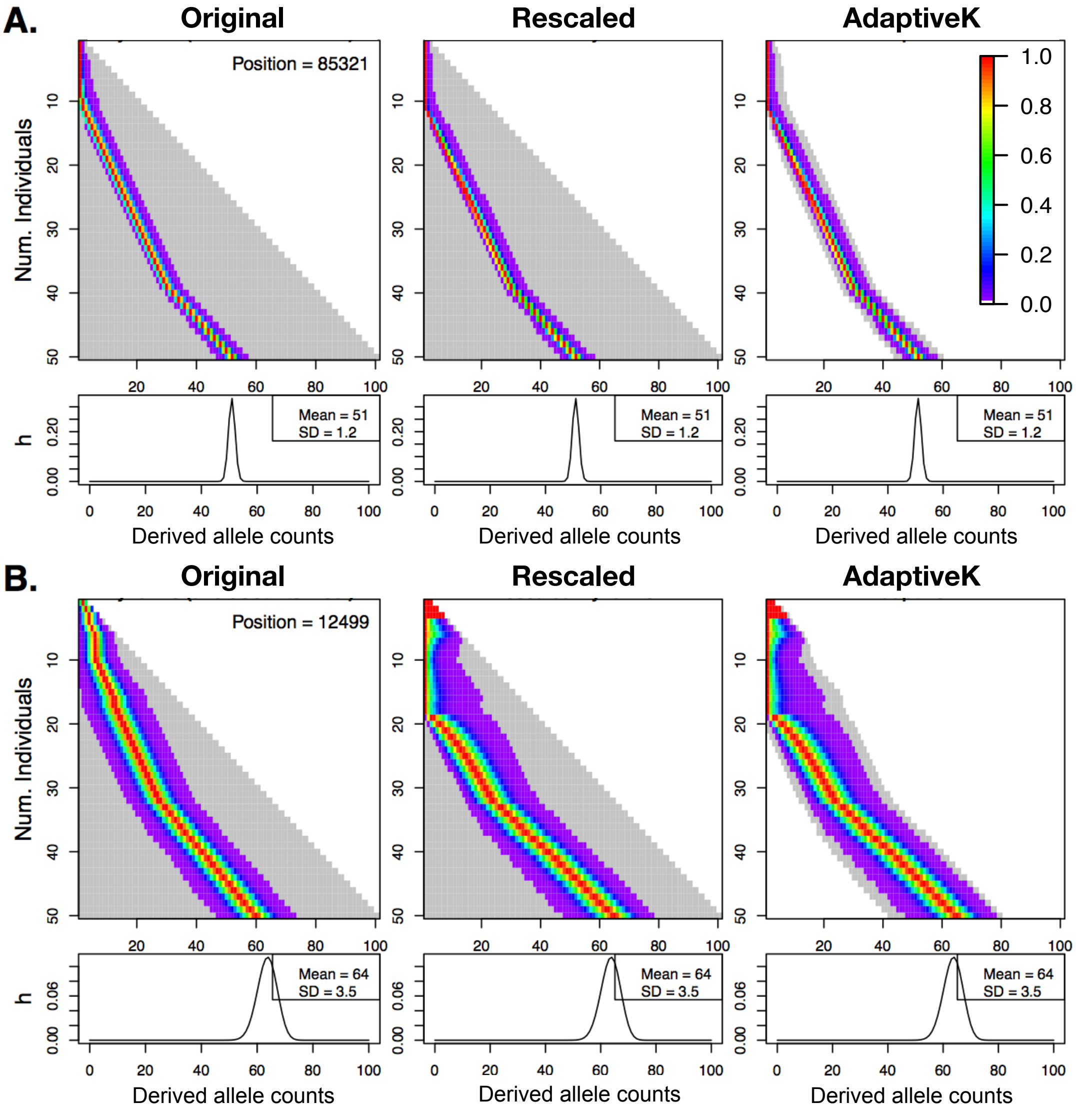


Figure 4.4: Performance of the three algorithms (Original, Rescaled, and AdaptiveK) for updating the site likelihood vector for 50 GBR individuals at two random sites with the different variance of the site likelihood vector. Each row represents the site likelihood vector for j individuals. The gray area represents the range of values used for each algorithm. **A.** Position 10,085,321 in chromosome 10. **B.** Position 10,012,499 in chromosome 10.

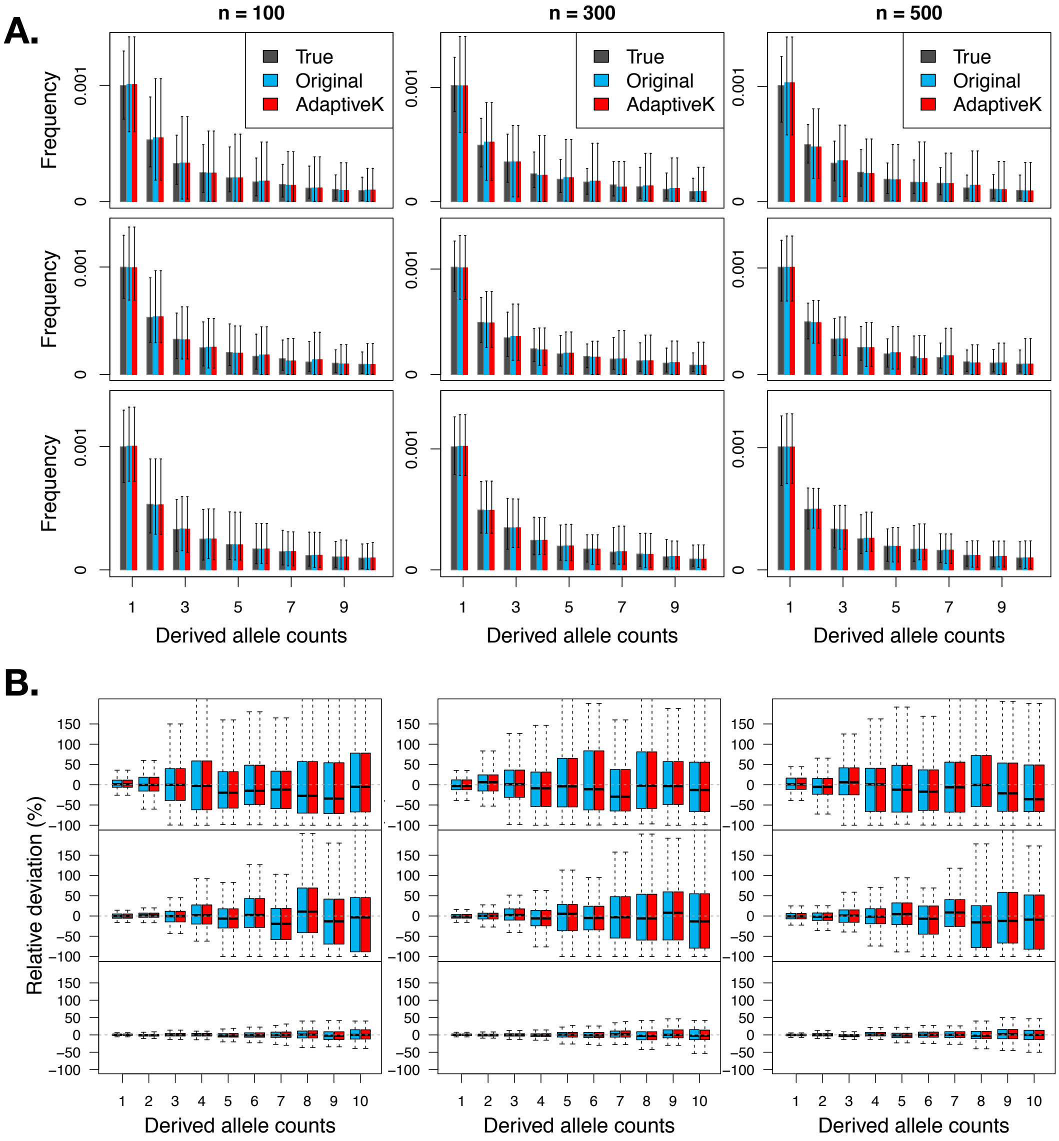


Figure 4.5: The accuracy of the inferred SFS as a function of sequencing coverage for different sample sizes. The sequencing data were simulated at coverage 3X (top), 5X (middle), and 10X (bottom) with the error rate of 0.001 and the sample size of 100, 300, and 500. **A.** Shapes of the inferred SFS (shown in colors in legend) compared with the ground-truth SFS (shown in gray). **B.** Relative deviation of a fraction of sites with the derived allele counts of 110.

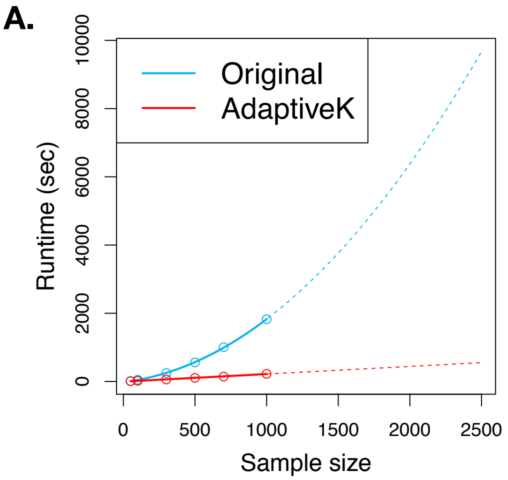
4.4.4 Runtime comparisons

We next evaluated the runtime for computing site likelihood vectors by the adaptive K-restricted algorithm (AdaptiveK) compared to the runtime by the original DP algorithm (Original). We observed a reduction of runtime with the adaptive K-restricted algorithm compared to the original DP algorithm for all sample sizes we tested (Figure 4.6). For example, on average, with 500 individuals we observed 5.3-fold speed-up, and with 1000 individuals we observed 8.4-fold speed up. Moreover, consistent with our expectation, the runtime of the original DP algorithm increases quadratically with sample size, whereas the runtime of the adaptive K-restricted algorithm has a linear fit (Figure 4.6). These results imply that as a sample size increases, we will observe even more dramatic differences in the runtimes of the two algorithms. For example, when we extrapolated runtime from the results with $n \leq 1000$, we expect 17.5-fold speed-up with 2,500 individuals and 63-fold speed-up with 10,000 individuals. The speed improvement with the adaptive K-restricted algorithm will greatly facilitate direct inference of the SFS even when the number of individuals is large.

We also note that the adaptive K-restricted algorithm will have less memory usage than the original DP algorithm, which requires memory on the order of n to store the site likelihood vector. With the adaptive K-restricted algorithm, the memory needed is to store the K elements of the vector, and we expect K to stay nearly constant or to scale upwards slowly in proportion to n.

4.4.5 Application to the low-coverage 1000 genomes project sequencing data

Finally, we compared the SFS inferred by the call-based approach to the SFS inferred by the direct estimation approach using our adaptive K-restrictive algorithm. We used 365 European (EUR) individuals and 228 sub-Saharan African (AFR) individuals to infer the SFS. For the call-based estimation approach, we used the genotype calls stored in the VCF file and then estimated



B.

Sample size	Fold Difference	Original (10Mb)	AdaptiveK (10Mb)
100	2.7	1.3 hr	30 min
500	5.3	15.8 hr	3 hr
1,000	8.4	2.1 days	6 hr
2,500	17.5	11.2 days	15.4 hr
10,000	63.0	162.3 days	2.6 days

Figure 4.6: Runtime comparisons for updating the site likelihood vector by two different algorithms (Original and Adaptive K). The sequencing data were simulated at coverage 3X, 5X, and 10X with the error rate of 0.001 and the sample size of 50, 100, 300, 500, 750, and 1000. Results for $n \geq 1000$ are extrapolated from results with $n \leq 1000$.

the SFS by allele counting. Note that the VCF file is generated by an LD-aware multisample genotype calling pipeline (1000 Genomes Project Consortium, 2012). For the direct estimation approach, we inferred the SFS directly from low coverage short-read sequencing data (stored in the BAM files, coverage 2-4X) using our adaptive K-restricted algorithm.

First, we constructed the SFS for 365 EUR individuals with either the call-based approach or the direct estimation method. We observed a striking lack of singletons in the call-based SFS compared with the directly estimated SFS (Figure 4.7A,B). The proportion of singletons in the inferred SFS by the VCF file is 95% less than that in the inferred SFS by the BAM files (Figure 4.7B). This is consistent with our previous study (Han *et al.*, 2014) that shows multisample callers lead to underestimation of rare variants, because a small number of correct alternate reads tend to be ignored. Consistent with this, we observed more positive Tajima's D for the call-based SFS compared with the directly estimated SFS (Figure 4.7C). Moreover, we observed an excess of sites fixed for an ancestral allele in the called-based SFS, implying

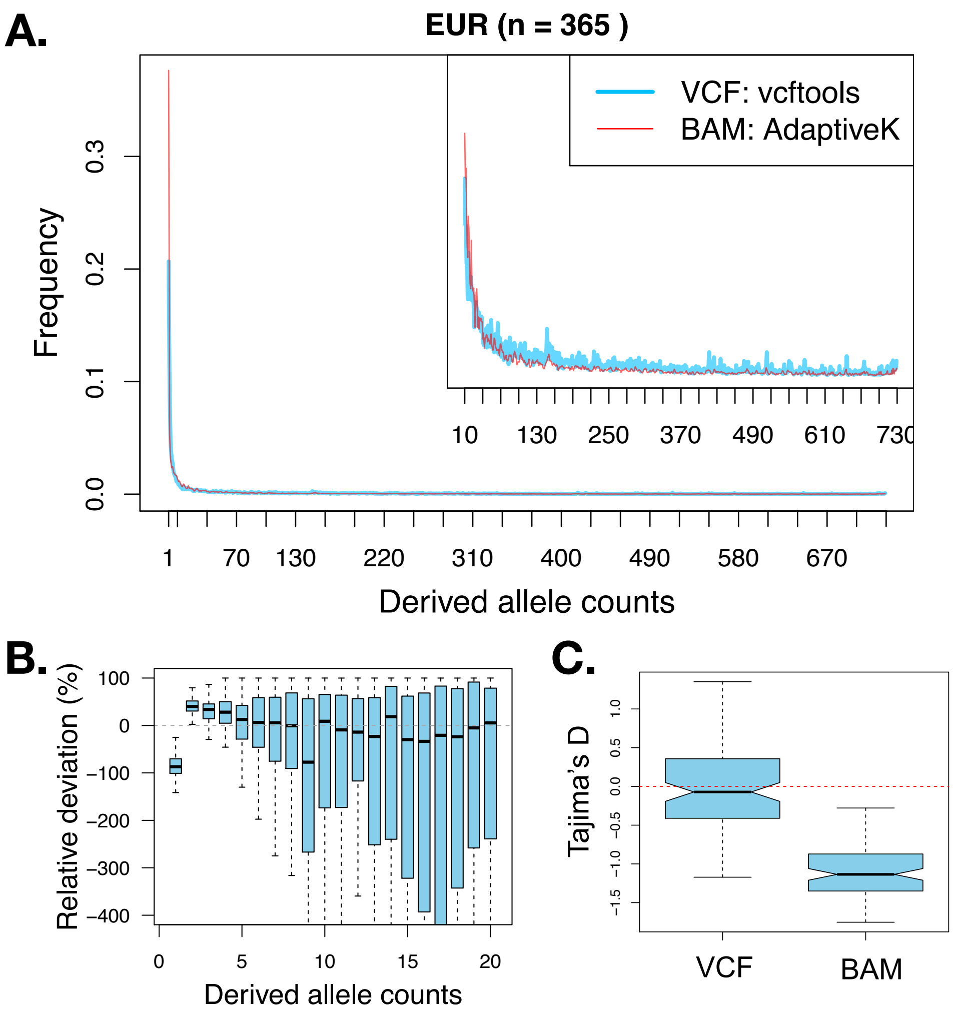


Figure 4.7: Comparison of the called-based SFS (referred to as VCF: vcftools) and the directly estimated SFS (referred to as BAM: AdaptiveK). The SFS was constructed for 365 EUR individuals in the 1000 Genomes Project. **A.** Shapes of the inferred SFS (shown in colors in legend). As the VCF file only contains sites that are inferred to be polymorphic, we only considered polymorphic sites for the SFS inferred from the BAM files and rescaled it so that all elements sum to one. **B.** Relative deviation of a fraction of sites with the derived allele count of 120. We computed the relative deviation of the SFS inferred from the BAM files compared to the SFS computed from the VCF file in each derived allele frequency bin $i/(2n)$. **C.** Tajima's D.

that there might be more polymorphic sites in the genetic region we analyzed than the reported polymorphic sites in the VCF file provided from the 1000 Genome Project (data not shown).

Next, we inferred the SFS for 228 AFR individuals and a combined sample of 593 EUR and AFR individuals with either the call-based approach or the direct estimation approach. We observed a similar pattern as with the European population, implying that our results apply to all samples in the 1000 Genomes Project (data not shown).

4.5 Discussion

A large sample size enables us to infer more precise summary statistics and parameters in many population genetic analyses. However, at the same time, we confront computational challenges with large samples and in many cases, we have to deal with these challenges to make the method practical with large sample sizes. We showed that although the direct estimation approach for computing the SFS can provide the unbiased SFS even at low coverage, it does not scale up to large sample sizes because the computation time for running this method is quadratic in a number of diploid individuals. To overcome this problem, we developed a new algorithm, called the adaptive K-restricted algorithm and showed that the computation time for running this algorithm is linear in the number of genomes. This algorithm exploits the observation that for most sites the site likelihood vector's non-negligible values are all concentrated on a few elements around the element corresponding to the best-guess allele count. Therefore, we approximate this vector by curtailing computation to only the K components of the DP update vectors. More importantly, this algorithm can adaptively choose K for each site. We showed that the choice of K is robust to sequencing coverage and the variation of the site likelihood vector. We also showed that the EM combined this new algorithm has comparable accuracy but is 8-fold faster than the original DP combined with the EM algorithm when analyzing the data from 1000 individuals. Our new algorithm's improvement in speed makes it possible to directly estimate the SFS from very large samples of low coverage short-read sequencing data.

Our adaptive K-restricted algorithm could be applied to other DP algorithm whose runtime is quadratic in a sample size. For example, Yi and coworkers (Yi *et al.*, 2010) proposed an empirical Bayes approach to estimate a posterior probability of a minor allele frequency (MAF). They used a DP algorithm to effectively compute summation over all possible genotype configurations for n diploid individuals, and therefore this algorithm has a runtime complexity of $O(n^2)$ similar to the DP algorithm introduced here. Furthermore, similar to the distribution of the site likelihood vector, the distribution of the posterior probabilities of the MAF is unimodal and most of the probabilities are close to zero. Therefore, we can apply our adaptive K-restricted algorithm for this DP algorithm to reduce runtime complexitiy to $O(Kn)$ rather than original $O(n^2)$.

Our adaptive K-restricted algorithm can also be directly applied to speed up estimation of the 2-dimensional SFS. Li (Li, 2011) derived the EM algorithm to get the MLE of the 2-dimensional SFS as an extension to the 1-dimensional SFS estimation, and this requires precomputation of the site likelihood vectors for all sites for each population independently. This implies that we can make this method faster with the adaptive K-restricted algorithm compared to the original DP algorithm. The computation time for running the original algorithm is $O(n_1^2 + n_2^2)$, whereas the runtime of the adaptive K-restricted algorithm becomes $O(K_1 n_1 + K_2 n_2)$, where n_1 and n_2 represent a sample size for each population.

One might argue that uncertainty associated with genotype calls can be overcome by simply increasing sequencing coverage and there is therefore little need for algorithms that handle low coverage data. However, cost constraints require difficult choices between increasing sample size and increasing coverage. There are certain cases where one should prefer a large sample of low-coverage sequencing data over a smaller sample size with high coverage. For example, in genome-wide association studies, one can obtain more power by sequencing a large number of individuals at low coverage (Kim *et al.*, 2010; Pasaniuc *et al.*, 2012). As another example, identification of rare variants always requires large sample sizes, and moderately rare loci will be detectable even with low coverage data. Finally, even though sequencing cost keeps dropping, cost constraints will not disappear because users will continue to work with limited budgets or push the limits with applications involving very large numbers of individuals; thus we expect low-coverage sequencing will remain an attractive approach for many investigators.

4.6 Appendix

Below is pseudo-codes for the adaptive K-restricted algorithm.

```
Input: Genotype likelihood vectors for n individuals, L1,...,Ln
Output: Site likelihood vector, h
epsilon=$10^{-9}$
h=array(0, length=2n+1)
/* first individual */
h[0]=L1[0];   h[1]=L1[1];   h[2]=L1[2];
left=0;   right=2
for j in 2:n {
        nChr=2*j
        L0=Lj[0];    L1=Lj[1];    L2=Lj[2];
        if(L0>=L1 and L0>=L2)
                bestGuessGT=0
        else if(L1>L2)
                bestGuessGT=1
        else
                bestGuessGT=2
        /* Set a left boundary and update accordingly */
        left= left+ bestGuessGT
        checkVal=(nChr-left)*(nChr-left-1)*L0*h[left]+2*left*(nChr-left)*L1*h[left-1]
        +left*(left-1)*L2*h[left-2]
        while(check>epsilon) {
            left=left-1
            checkVal=(nChr-left)*(nChr-left-1)*L0*h[left]+2*left*(nChr-left)*L1*h[left-1]
            +left*(left-1)*L2*h[left-2]
        }
        /* Set a right boundary and update accordingly */
        right = right+ bestGuessGT
        checkVal=(nChr-right)*(nChr-right-1)*Lj[0]*h[right]
        +2*right*(nChr-right)*Lj[1]*h[right-1]+right*(right-1)*Lj[2]h[right-2]
        while(check>epsilon) {
            right=right+1
            checkVal=(nChr-right)*(nChr-right-1)*Lj[0]*h[right]
            +2*right*(nChr-right)*Lj[1]*h[right-1]+right*(right-1)*Lj[2]h[right-2]
    }
    /* Update the site likelihood vector by the adaptive K-restricted algorithm */
    for x in right:left
        h[x]=(nChr-x)*(nChr-x-1)*Lj[0]*h[x]+2*x*(nChr-x)*Lj[1]*h[x-1]+x*(x-1)*Lj[2]*h[x-2]
    /* Normalization */
    mymax=max(h)
    for k in left:right
        h[x] = h[x]/mymax
}
```

CHAPTER 5

Identification of Genetic Regions of Local Adaptation During Early Dog Domestication

Work in this chapter is part of a larger research effort of the Novembre group (Adam H. Freedman, Eunjung Han, Diego Ortega-Del Vecchyo, John Novembre) and the Wayne group (Rena M. Schweizer, Pedro M. Silva, Marco Galaverni, Robert K. Wayne) for which I am included as a co-author. In particular, section 5.2.1 Samples and Sequencing was the work of Rena M. Schweizer, Adam H. Freedman, Holly Beale, Elaine Ostrander, Kevin M. Squire, Vasisht Tadigotla, Clarence Lee, Timothy Harkins, Stanley F. Nelson, Robert K. Wayne, and John Novembre (Supplementary material S1, S2, and S3 of the original manuscript). Section 5.2.2 Quality Filtering was the work of Adam H. Freedman, Pedro Silva, Marco Galaverni, Eunjung Han, Robert K. Wayne, and John Novembre (Supplementary material S4 of the original manuscript). Section 5.2.3 Detection of Selective Sweeps was the work of mine, section 5.2.4 Validation of Selective Sweep Regions was the work of Rena M. Schweizer and me, and section 5.2.5 Gene Ontology/Enrichment Analyses was the work of Pedro Silva (Supplementary material S13 of the original manuscript; I am the lead author for this document). These sections are provided here to allow better understanding of the selection scan method presented in section 5.2.3 and the interpretation of the results pertaining to selection scans in section Results.

5.1 Introduction

Domestic dogs are the most phenotypically diverse mammal. Dogs exhibit the extreme variation in size and skeletal and cranial proportions (Wayne R et al. 2012). Moreover, dogs have extreme behavioral and physiological attributes, such as herding, attentiveness, hunting, docility and the ability to form social bonds with humans (American Kennel Club 1992; Wilcox and Walkowicz 1995). Given the unique phenotypic and behavioral traits of dogs, comparative genomics analyses of dogs and wolves holds great promise for identifying genetic loci involved in complex phenotypes.

However, there are several complications that make detection of regions under selection in domestic dogs challenging. First, typical modes of selection during domestication is selection from standing variation (soft sweeps). Some preexisting variants (that are neutral and subject to random drift) in the founder population of its wild progenitor suddenly become beneficial due to humans' desire to acquire certain phenotypes, and therefore rapidly increase in frequency and become fixed in the derived population. However, soft sweeps are harder to detect compared to hard sweeps, because a signature of selection (eg. reduction in genetic variability) is more subtle. Next, demographic history is another complication. Most domesticated species experience bottlenecks during early domestication and breed formation. Often these events reduce overall genetic variation, making it harder to detect genetic signatures of selective sweeps.

Innan and Kim (2008) quantified the power of statistical tests for detecting the signature of soft sweeps with polymorphism data by using coalescent simulations, and concluded that comparing the patterns of polymorphisms in both parental and derived population can improve the power substantially rather than comparing the patterns in the focal region with those in different regions in the derived population. Consistent with this observation, they found that F_{ST} and $\Delta\pi = \pi_{derived}/\pi_{parental}$ are most powerful summary statistics for detecting soft sweeps. Moreover, they found that the overall pattern is similar under different demographic parameters.

Here, we analyze 10 million variant sites from the whole-genome data of six canids to advance our understanding of dog domestication. These data include the first whole-genome sequences of three individual wolves (Canis lupus), the Australian Dingo, a Basenji, and a golden jackal (Canis aureus). With these data, we investigated the specific loci involved in selection during early dog domestication by using combinations of summary statistics that were shown to be powerful to detect soft sweeps.

5.2 Materials and Methods

5.2.1 Samples, Sequencing, and Genotyping pipeline

The three wolves sampled were from Croatia, Israel, and China, and were chosen to represent the broad regions of Eurasia where domestication is hypothesized to have taken place (Europe, the Middle East, and East/Southeast Asia) (Larson *et al.*, 2012). The two dogs, a Basenji and a Dingo, represent basal canine lineages and were sampled to maximize phylogenetic divergence and geographic diversity, as the Basenji breed originated in Africa and Dingoes are free-living dogs of Australia that arrived there at least 3500 years ago (Savolainen *et al.*, 2004). Sequencing the golden jackal allowed us to identify the ancestral state of variants so we could ascertain those changes that occurred uniquely on the dog lineage.

For each of the six samples (Table 5.1), we produced high-quality whole-genome sequencing data. For all individuals besides the Chinese wolf, we used a combination of SOLiD (single end and long mate pair) and Illumina HiSeq paired end (PE) libraries, while for the Chinese wolf we only used Illumina PE data, as they were provided subsequent to our sequencing efforts for the other lineages. For most downstream analyses, we also utilized sequence information from the Boxer reference genome (CanFam 3.0). Cumulative coverage was 72X for the wolves (24X average per individual), 38X coverage for the two dogs (19X average per individual), and 24X for the golden jackal, for a total of 335Gb of uniquely aligned sequence from 11.2

Table 5.1: Sample origins, and sequencing efforts by sample, platform, and library

					SOLiD	
Sample	**Sample ID**	**Sample Origin**	**Sex**	**SOLiD LMP**[a]	**fragment**	**HiSeq**[b]
Basenji	RKW 13764	Bethesda, MD, USA	M	1	—	1
Dingo	RKW13760	Bargo Dingo Sanctuary, Australia	M	1	2[c]	1
Israeli wolf	RKW13759	Neve Ativ, Golan Heights, Israel	F	1	1[d]	1
Chinese wolf	RKW13451	San Diego Zoo, CA, USA	F	—	—	3
Croatian wolf	RKW 3919	Perković, Croatia	F	1	1[d]	1
Golden jackal	RKW 1332	Tel Aviv, Israel	F	2	1.75[d]	1

[a] Number of slides, long mate pair, 1.5kb insert, 50bp per end

[b] Number of lanes, paired end 400bp insert, 100bp per end

[c] Number of slides, 75bp

[d] 50bp

billion sequence reads. In contrast, surveys of wolf genetic diversity to date have been limited to shotgun sequencing with incomplete genomic coverage (Lindblad-Toh *et al.*, 2005), small numbers of sequence loci (Gray *et al.*, 2009), or limited pooled sequencing (6X average from a pool of 12 wolves, 30X average from a pool of 60 dogs) (Axelsson *et al.*, 2013).

We implemented a sequencing alignment and genotyping pipeline customized for combining SOLiD and Illumina HiSeq short read data, using aligners tailored to the specific platforms, then post-processing alignments using the Picard (http://picard.sourceforge.net) and Genome Analysis Toolkit (GATK) toolsets (DePristo *et al.*, 2011). This pipeline converted short read raw data to .bam format alignment files (Li *et al.*, 2009a), and from bam files to genotype files in vcf format (http://www.1000genomes.org/node/101). Our analyses draw on 10,265,254 high quality variants detected by our genotyping pipeline, of which 6,970,672 were at genomic positions with no missing data for any lineage. We estimate error rates to be low based on comparison to genotype calls from genotyping arrays (e.g. heterozygote discordance rates of 0.01-0.04%).

5.2.2 Quality Filtering

In line with previous studies utilizing next-generation sequencing data, we developed a series of conservative data quality filters, implemented post-genotyping. Filters served two purposes. First, we sought to minimize the effects of sequencing and alignment errors that might bias downstream analyses (Jordan and Goldman, 2012; Nielsen *et al.*, 2011). Second, we sought to exclude regions of the genome that, irrespective of such errors, might show accelerated rates of evolution for reasons other than positive selection on the dog lineage, and might falsely appear as outliers in our selection scans; such regions might also be prone to misalignment of short reads. We established sets of criteria with which to filter at both the level of genomic position and individual lineages. Genome feature filters were applied to genomic positions based upon intrinsic features of the reference (Canfam3) and polymorphism across samples (i.e. tri-allelic and CpG sites), while sample feature filters were applied to individual lineage genotypes based upon features of the data underlying the corresponding genotype call. We annotated our VCF files according to whether genomic positions and samples passed the respective filtering criteria.

Genome feature filters Genomic positions in a VCF file were flagged as not passing the genome feature filter according to the following four criteria.

1. Repeat Regions: We identified all genomic positions falling within repeat regions of the reference genome identified with RepeatMasker (Smit *et al.*, 2010) and Tandem Repeat Finder (TRF) (Benson, 1999). We annotated our VCF file according to the class of repeat detected, collapsing the output repeat classes into a reduced set of 14 classes: SINE, LINE, LTR, DNA, RNA, rRNA, scRNA, snRNA, srpRNA, tRNA, Satellite, Simple repeat, Low complexity sequence, and Unknown. Because ancient repeats can make up a substantial portion of genomes, and because these regions will have diverged enough to allow accurate read mapping with short read alignment algorithms, we sought to retain these, and only mask out younger repeats prone to sequence misalignment. We considered that erroneous

mapping of short reads to these regions should lead to increased frequency of heterozygous genotype calls, and we conservatively chose 25% divergence as our minimum repeat divergence threshold, as repeats in this interval show no increase in heterozygosity with decreasing repeat age.

2. CpGs: Mutation rates at CpG sites are substantially higher than non-CpG sites (Hodgkinson and Eyre-Walker, 2011), so that regions enriched for CpGs may display elevated diversity and/or divergence leading to outliers in window-based analyses, independent from any demographic or selective forces germane to our investigation of domestication. If in any of our six lineages, a nucleotide that otherwise passed filter fell within a CpG dinucleotide, because at least some proportion of our data fell into that hyper-mutable site category, we flagged the genomic position.

3. Copy Number Variants (CNVs): When true CNVs are not included in a reference genome assembly, or when samples mapped to the reference contain novel CNVs, misalignment of paralogous reads is more probably, and can lead to false positive SNVs that can bias estimated levels of polymorphism and divergence. To minimize the effects of such misalignment, we constructed a set of CNV regions to exclude from downstream analyses, by combining a set of previously discovered CNVs reported in a diverse panel of dog breeds (**?**), and those we discovered directly from the short read data generated for our six canid lineages.

4. Triallelic sites: Preliminary comparisons of genotypes from sequencing with those from the Illumina CanineHD BeadChip indicated triallelic sites were more prone to genotyping errors, and so these sites, while making up a relatively small fraction of the genome, were excluded.

Sample feature filters Samples were flagged as not passing the sample feature filter according to the following four criteria.

1. Proximity to Indel: Short reads generated by next-generation sequencing platforms are prone to misalignment near indels, and attempts at local realignment around indels may not fully rectify this problem. As a result, these indel-proximate misaligned regions may be enriched for false positive SNVs. To account for this potential source of bias, for each sample we excluded any genotype containing an alternative allele relative to Canfam3 that was within 5bp (either up or downstream) of another SNV containing genotype within the same sample.

2. Genotype Quality: Genotype quality (GQ) metrics output by the GATK (DePristo *et al.*, 2011) Unified Genotyper (UG) represent phred-scaled probabilities that the called genotype does not match the true underlying genotype, i.e $-10 \times log_{10}P(error)$. We chose a hard minimum GQ threshold of 20 ($P(error) = 0.01$) based upon two considerations. First, we sought to minimize genotyping errors as measured by discordance with an independent, high quality genotype data set from the Illumina SNP chip. Second, we sought to balance the competing goals of retaining maximum genomic coverage while being able to correctly identify specific mutations of functional significance, particularly those fixed between dogs and wild canid species. Hard genotype quality thresholds may lead to undercalling of heterozygotes in samples with low or moderate coverage, but works well with those at $>$ 20X coverage (Nielsen *et al.*, 2011). All but one of our canid lineages were sequenced at $>$ 20X. Two additional lines of evidence support our use of a hard GQ threshold. First, the majority of all emitted genotypes have GQ $>$20 (Basenji 83.1%, Dingo 93.5%, Israeli wolf 95.6%, Croatian wolf 93.2%, Chinese wolf 98.9%, golden jackal 93.7%). Second, for our lowest coverage sample, the basenji, filtering on GQ appears to exclude more low quality homozygous genotypes, as the proportion of heterozygous calls shows an increasing trend with GQ above GQ=20.

3. Excess Depth of Coverage: Extremely high depth of coverage relative to the genome-wide average likely indicates misalignment of reads generated from paralogous positions

in the genome, particularly those containing CNVs. Indeed, excess depth of coverage is a typical metric used to define CNV regions, but CNV filtering alone will fail to detect finer-resolution CNV signatures. Thus, we conservatively filtered all sites where depth of coverage exceeded twice the mean depth of coverage recorded for each lineage. GATK UG filters out reads that fail to meet certain criteria (see above). As a result, post-GATK filtering, depth of coverage may fall below our 2X threshold, even when the GATK filtering of hundreds of reads would indicate a region that may intrinsically be prone to read misalignment. Thus, our filtering on depth of coverage is based upon the number of reads overlapping a genomic position prior to imposition of the UG's internal filters.

4. Clustered SNVs: Within any sample, we excluded all SNV-containing genotypes falling within 5 bp of another SNV-containing genotype. In identifying clustered SNVs, to be conservative we required that proximate SNVs only have a minimum genotype quality of 10, rather than the 20 employed in our downstream evolutionary analyses.

5.2.3 Detection of Selective Sweeps

Step 1: Computation of Summary Statistics for Selection Scans Based on the results of Innan and Kim (Innan and Kim, 2004, 2008), we chose three summary statistics, F_{ST}, $\Delta\pi$ and ΔTD, to detect candidate regions of selective sweeps during early dog domestication. These three summary statistics are all standard statistics to summarize patterns of genetic variation and are shown to be most powerful to detect soft sweeps among nine summary statistics they tested in their simulation studies.

F_{ST} (Weir and Cockerham, 1984) measures differences in allele frequency between populations. For a region under positive selection, we expect unusually high F_{ST}. $\Delta\pi$ (Innan and Kim, 2004, 2008) is a ratio of nucleotide diversity between two populations, defined as $\Delta\pi = \pi_{wolf}/\pi_{dog}$. The large value of $\Delta\pi$ can capture a local reduction in diversity around the selected locus in dogs, which is a characteristic signal of selective sweeps. ΔTD is a difference in Tajima's D, defined as $\Delta TD = TD_{wolf} - TD_{dog}$. Tajimas D (Tajima, 1989) is a site frequency spectrum (SFS)-based neutrality test statistic that contrasts the θ estimator based on the number of segregating sites ($\hat{\theta}_S$) to the θ estimator based on average pairwise differences ($\hat{\theta}_\pi$). For regions under selective sweeps in dogs, more positive values of ΔTD are expected.

Since all three statistics capture different footprints on genetic variation generated by positive selection, we consider individual statistic separately (F_{ST}, $\Delta\pi$ and ΔTD) and all three statistics jointly (joint empirical percentiles of F_{ST}, $\Delta\pi$ and ΔTD).

We used a sliding window approach in which we divided the reference genome into overlapping windows of size 100Kb with 10Kb increments. For each 100Kb-window, we computed summary statistics using only sites that pass the genome feature (GF) filter and where genotypes are observed and pass sample feature (SF) in both dogs and all three wolves. We considered the boxer reference haplotype when we compute statistics within the dog sample or between the dog and wolf sample. To facilitate these calculations, we wrote a C++ program vcfSummary to compute the following summary statistics in each window from our VCF files:

- The number of fully observed sites: The number of sites passing the GF and SF filters in both dogs and all three wolves.
- The number of segregating sites per base pair within the dog or wolf sample (s_{dog}, s_{wolf})
- Average pairwise differences per base pair within the dog or wolf sample (π_{dog}, π_{wolf})
- Tajimas D within the dog or wolf sample (D_{dog}, D_{wolf})
- Fixation indices (F_{ST}): For each window, we took weighted averages over fully observed sites, because the weighted averages over loci have been shown to perform better than unweighted averages and they avoid any problems of zero denominators in computation of F_{ST} (Weir and Cockerham, 1984).

- Ratio of nucleotide diversities ($\Delta\pi$): $\Delta\pi = \pi_{wolf}/\pi_{dog}$ in each window. For the numerical stability, we used $log(\Delta\pi) = log(\pi_{wolf}) - log(\pi_{dog})$ in each window.
- Difference of Tajimas D (ΔTD): $\Delta TD = D_{wolf} - D_{dog}$ in each window.
- Recombination rates: We obtained the recombination map estimated from village dogs from A. Boyko (personal communication) and calculated the average recombination rate for each window.

Step 2: Filtering of windows with a low number of fully observed sites We obtained 220,020 sliding windows of size 100kb with 10kb increments genome-wide. We then discarded any windows in which the number of fully observed sites is less than 30,000, because it is more likely that those windows are within or close to repeat/CNV regions or regions of poor sequencing quality. Furthermore, as the variance of each summary statistic will depend on the number of fully observed sites, we can reduce false positive signals for detecting positive selection by discarding those windows with a low number of fully observed sites. We found that those windows with a low number of fully observed sites are clustered together in the genome rather than randomly scattered (figure 5.1, top). Thus, we clustered windows if they are within 200kb each other and also excluded the intervening regions in the subsequent analysis. Finally, we excluded any windows if they are within 1Mb from the telomere. After this, there remained 173,662 windows (78.93%) (figure 5.1, bottom). The empirical distribution of each summary statistic used for detecting outlier regions is shown in figure 5.2.

Step 3: Computing test statistics For each summary statistic (F_{ST}, $\Delta\pi$ and ΔTD), we computed empirical percentiles and empirical p-values for each window. They are rank-transformed statistics. They are computed by ranking each window by the summary statistic in question (minimum of the summary statistic is assigned to rank 1) and then transforming the ranks accordingly.

The empirical percentile for the window with rank x is defined as

$$\text{Empirical percentile } (x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n}$$

where n is the total number of windows, $I()$ is an indicator function and X_i is a rank of the i_{th} window. For example, the window with the maximum F_{ST} value (assigned to rank n) has an empirical percentile of 1.

The empirical p-value for the window with rank x is defined as

$$\text{Empirical p-value } (x) = \frac{\sum_{i=1}^{n} I(X_i \geq x)}{n}$$

where n is the total number of windows, $I()$ is an indicator function and X_i is a rank of the i_{th} window. For example, the window with the maximum F_{ST} value (assigned to rank n) has an empirical percentile of $1/n$.

We then calculated a *joint rank* of all three summary statistics by computing the product of the empirical percentiles obtained for the three summary statistics in each window ($\%F_{ST} * \%\Delta\pi * \%\Delta TD$) and ranking each window by the product. We transform the joint rank into either joint empirical percentiles or joint empirical p-values. In order to define outlier windows and outlier regions, we transformed the joint rank defined for each window into joint empirical percentiles. In order to draw Manhattan plots, we transformed the joint rank defined for each window into joint empirical p-values. The genome-wide distribution of joint empirical p-values is shown in figure 5.3.

Step 4: Defining outlier regions by clustering For each metric (F_{ST}, $\Delta\pi$, ΔTD and the joint empirical percentile), we defined the top 1% windows as outlier windows. Since the outlier windows are often clustered together in the genome, we joined outlier windows and intervening sequence to define outlier regions when windows were found within 200kb of each other. For each outlier region, we computed the maximum of the empirical percentiles and the number of outlier windows. We ordered the outlier regions by the ”maximum” of the empirical percentiles.

Top 100 outlier regions ordered by the maximum of the joint empirical percentile are shown in figure 5.8.

5.2.4 Validation of Selective Sweep Regions

Validating sweep signal at CCRN4L with external data To rule out the possibility that the observed sweep signal in our top outlier regions resulted from particular features unique to the genomes we sampled, we examined patterns of nucleotide diversity in two other data sets, (1) a panel of 12 dog breeds sequenced to 6-8x coverage on the Illumina sequencing platform, and the CanMap SNP data set (Vonholdt *et al.*, 2010). In both data sets, we observed a reduction in diversity in the dog lineage, consistent with a selective sweep at CCRN4L being a domestication-related signal generally found across breeds (figure 5.5 and 5.6).

Assessing data quality in outlier regions To ensure that our top outlier regions are not false positives due to a preponderance of low quality data, but instead due to observed patterns expected under selective sweeps, we examined the relationships of mean sequencing depth of coverage and proportions of fully observed sites in each window with each summary statistic (F_{ST}, $\Delta\pi$, ΔTD and joint empirical percentile). In almost all cases, the correlations were not significant, and in cases in which they were, the magnitudes were negligible (figure 5.7), suggesting that the sweep signals are not enriched for regions of low quality data.

5.2.5 Gene Ontology/Enrichment Analyses

The set of genes intersecting our outlier regions, defined by the joint empirical percentile, were tested for significant enrichment in Gene Ontology (GO) categories, Kegg/Reactome pathways (KGR) and Human Phenotype Ontologies (HPO) using the online tool g:Profiler (Reimand *et al.*, 2011) (http://biit.cs.ut.ee/gprofiler).

GO terms aim to summarize relevant information about gene function regarding their molecular activity (molecular function), the biological process they are involved in and the cellular compartments where they are active (http://www.geneontology.org). The KGR databases are curated reference databases for biological pathways (http://www.genome.jp/kegg/pathway.html, http://www.reactome.org). The Human Phenotype Ontology establishes standardized terms for phenotypic abnormalities encountered in human disease (http://www.human-phenotype-ontology.org). All the dog (Canis familiaris) genes annotated in Ensembl were used as background set, and the Benjamini-Hochberg false discovery rate (Benjamini and Hochberg, 1995) was applied to correct for multiple testing. We tested for enrichment for the list of all genes found within the top 10% of selection scan regions. We only report significantly enriched categories that included $\geq$5 genes and with multiple-testing corrected p-value less than or equal to 5%. Within enriched categories, those involved in skeletal and dental morphology were prevalent (Table 5.2)

5.3 Results

To find the top candidate regions that harbor potentially adaptive variation, we scanned the autosomal genome for signatures of positive selection on the dog lineage using three metrics (F_{ST}, $\Delta\pi$, and ΔTD, see Materials and Methods) that have been shown to have high power to detect regions under selection during domestication (Innan and Kim, 2004, 2008). We flagged extreme outliers in 100kb windows based on a joint percentile of these metrics, and then identified clusters of outliers to establish candidate selection regions (see Materials and Methods for details).

The top 100 outlier regions range in length from 10-530kb (figure 5.8). Forty-one of the top 100 regions did not contain any validated, annotated genes (figure 5.8). These regions may harbor important non-coding functional elements, but might also include unidentified coding regions. In support of this, we observe a 1.6-fold enrichment in CNEs in these regions relative to the genome-wide distribution (2.5% of outlier regions without genes vs. 1.6% genome-wide, Fisher's exact test, P-value=2.2×10^{-16}). Several of the genes in our top regions overlap with previous studies or with a re-analysis of previous SNP array data, in which we contrasted varia-

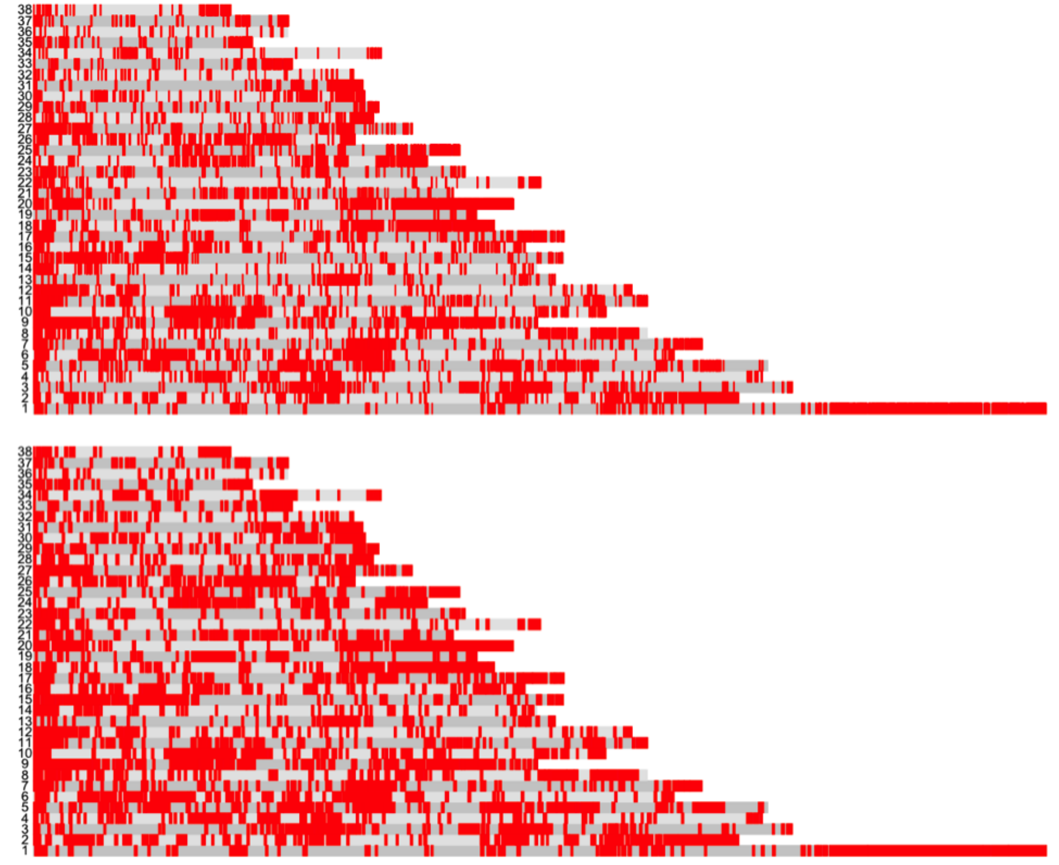


Figure 5.1: Genome-wide distribution of 100kb windows with the low number of fully observed sites.

tion between wolves and basal dogs (figure 5.8).

To assess whether particular functional groups of genes were enriched in our selection scan hits, we searched for Gene Ontology, Human Phenotype, and KEGG Pathway functional categories that are enriched across the top 10% of our selection scan hits. Functional enrichments were dominated by categories associated with skeletal and dental morphology such as abnormality of the joints of upper limbs, abnormality of the alveolar ridges, and abnormality of the 5th finger (Table 5.2). The last of these categories may underlie the development of the dewclaw in dogs, which is absent in wild canids. Additional enrichment categories include brain function (e.g. cerebellar malformation, cerebellar vermis hypoplasia, delayed closure of fontanelles)

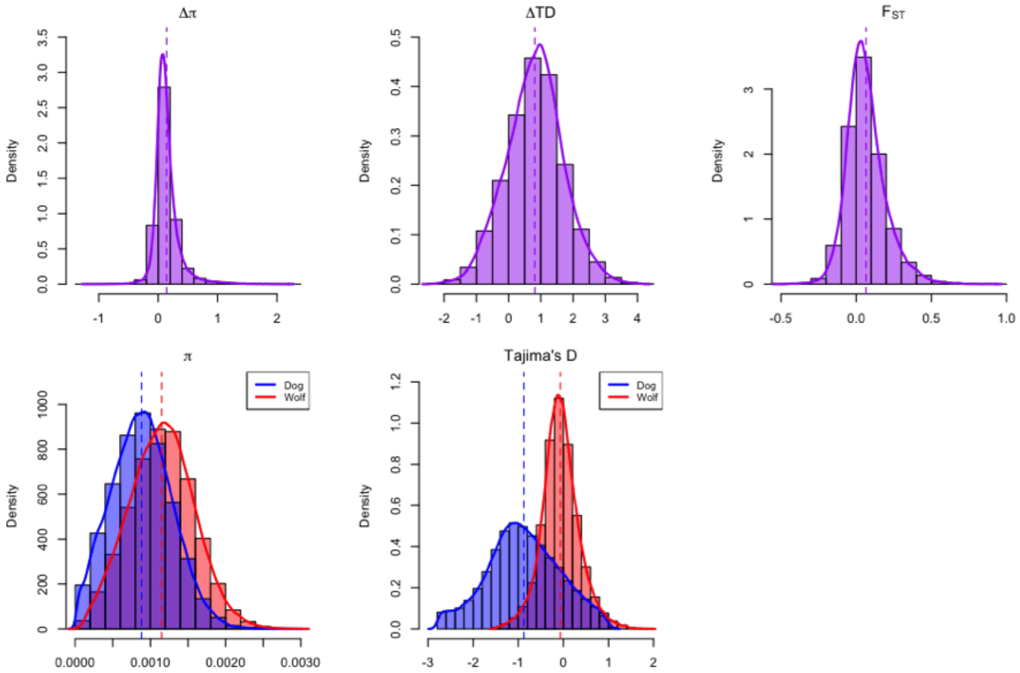


Figure 5.2: Empirical distribution of each summary statistic used for detecting outlier regions.

(Table 5.2).

Our top candidate region (based upon the joint percentile of selection scan statistics, see Materials and Methods) contains a portion of the ELF2 gene but is most strongly peaked on CCRN4L (figure 5.10A). CCRN4L (also known as Nocturnin) is expressed in a circadian fashion and studies in mice indicate CCRN4L activates PPAR-γ, a gene that promotes bone adipogensis as opposed to osteoblast formation and that is a known diabetes risk locus in humans (Kawai and Rosen, 2010). It also regulates the expression of additional genes involved in lipogenesis and fatty acid binding and knock-out mice are resistant to diet-induced obesity (Kawai and Rosen, 2010; Green *et al.*, 2007; Kawai *et al.*, 2010b,a). CCRN4L also suppresses IGF1, a well-known activator of bone growth (30) that underlies size variation amongst dog breeds (Sutter *et al.*, 2007; Hoopes *et al.*, 2012).

Four of the eight top candidate regions contained genes implicated with neurological func-

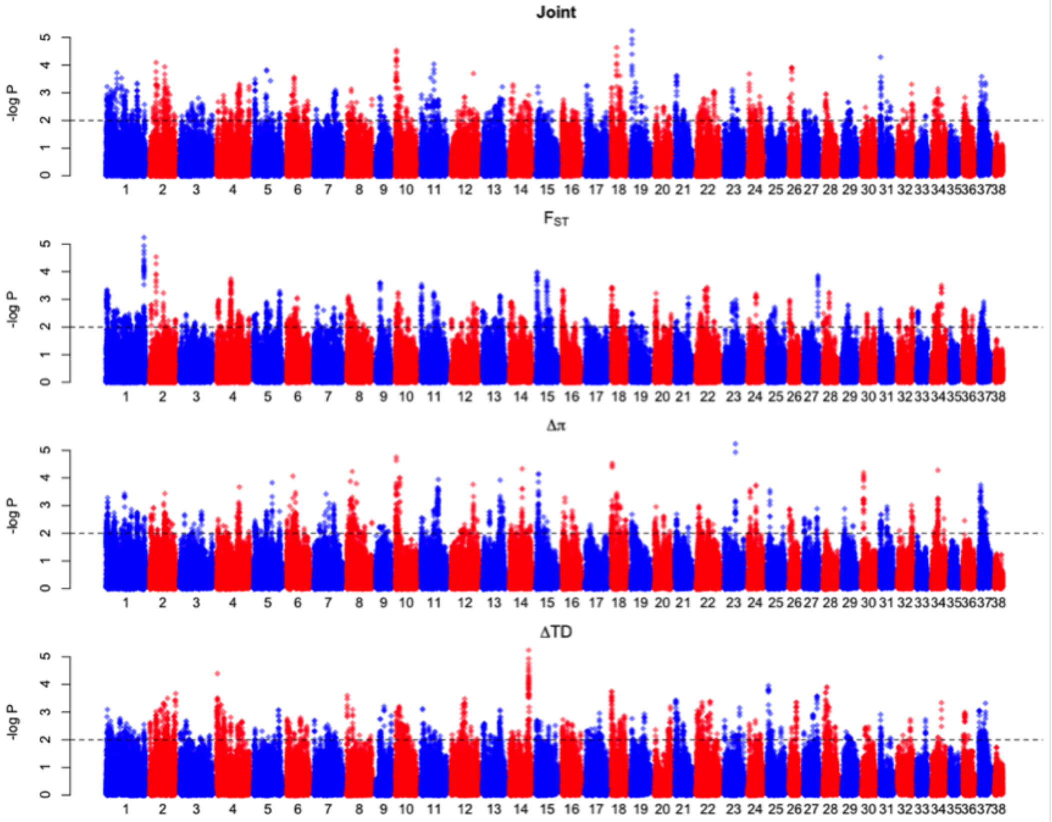


Figure 5.3: Genome-wide distribution of empirical p-values.

tions in other mammalian species: CADM2 (under the 4th hit) is a synaptic cell adhesion molecule whose flanking regions show reduced homozygosity in autism patients (Casey *et al.*, 2012); SH3GL2 (6th hit) affects synaptic vesicle formation (Schmidt *et al.*, 1999); PDE4D (7th hit) is a mammalian homolog of the dunce gene in Drosophila whose knockout in mice shows impaired learning (Rutten *et al.*, 2008); and CUX2 (8th hit) is a key marker of neuronal fate during mammalian cortex development (Franco *et al.*, 2012) whose knockout in mice shows deficits in working memory (Cubelos *et al.*, 2010). These genes are important in neural development and function, including cell type specification, synapse formation and function, and synaptic plasticity. This is consistent with the behavioral evolution required for dogs to cohabit with humans, although these genes may also have roles in other tissues.

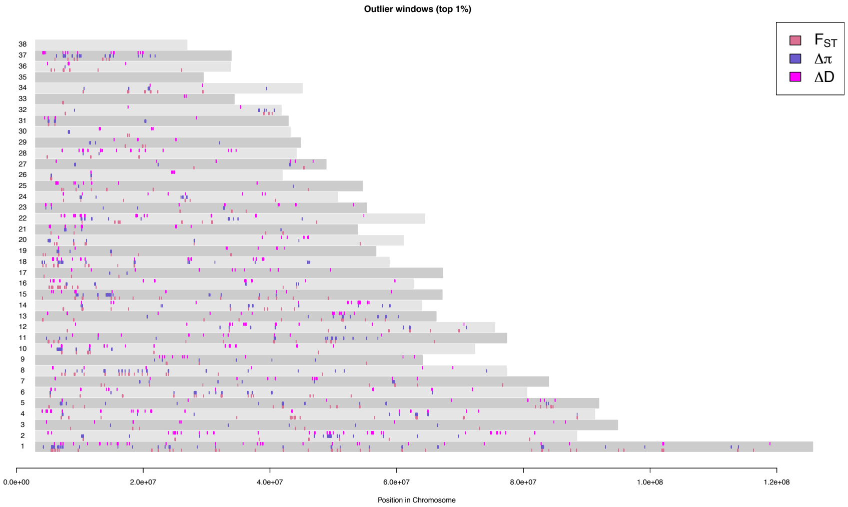


Figure 5.4: Overlap between top 1% outliers of three summary statistics

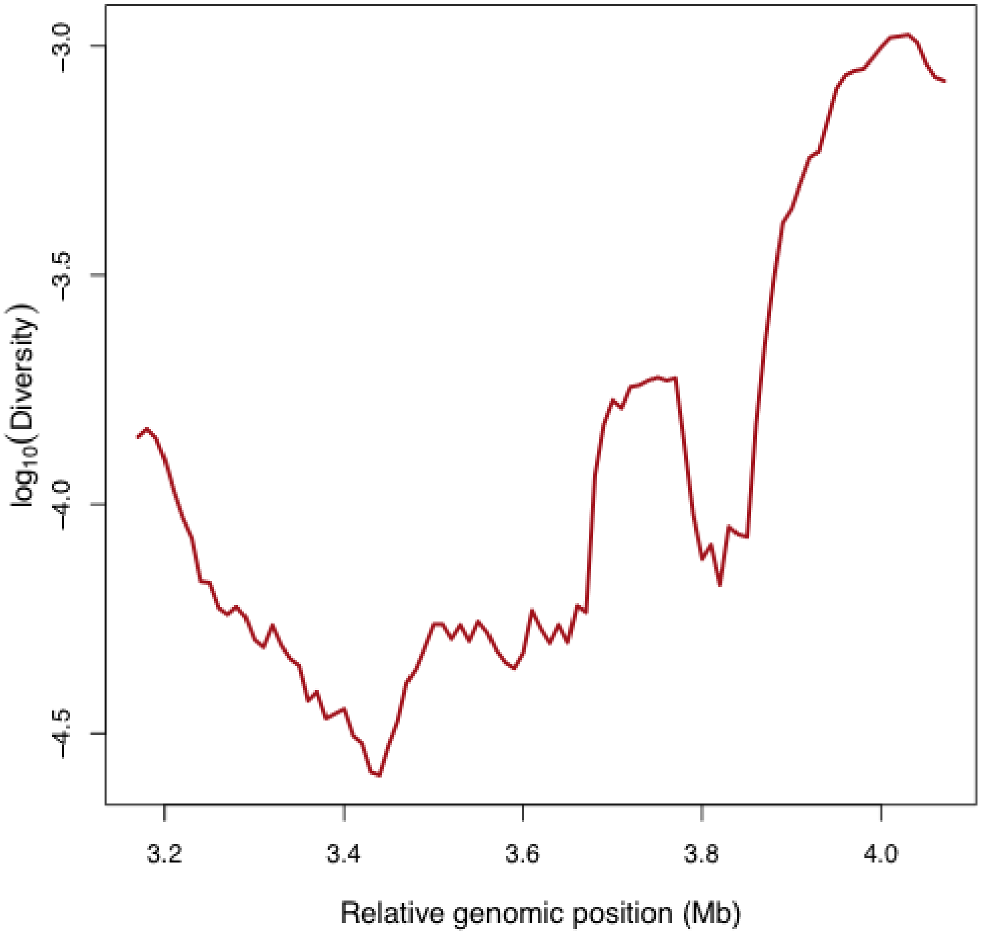


Figure 5.5: Pattern of nucleotide diversity among 12 dog breeds sequenced to low coverage on the Illumina platform (see Text S1), for region containing our top selection hit containing CCRN4L, showing similar reduction in nucleotide diversity consistent with a selective sweep. The region is depicted in coordinates for the updated dog reference genome, CanFam 3.1.

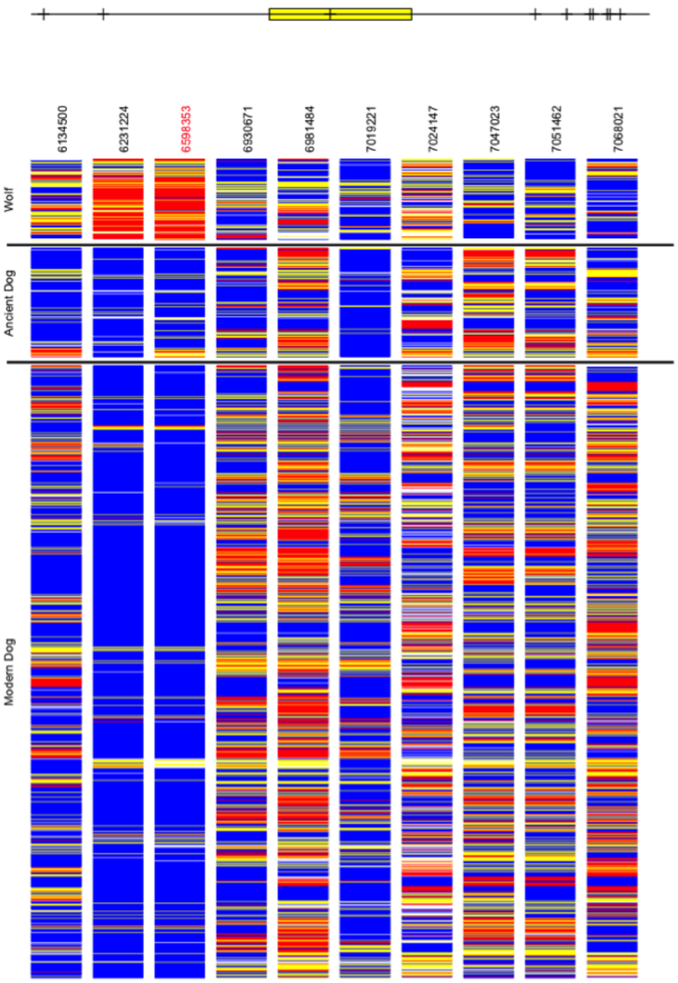


Figure 5.6: Genotype plot derived from CanMap SNP data, surrounding our top selection hit region containing CCRN4L. Vertical lines in graph (upper) indicate positions of SNPs relative to the outlier region in Fig. 4A (yellow box). Columns (lower) indicate diversity at those SNP positions; the SNP within the outlier region is indicated with red text. Diversity at this SNP shows a marked reduction in dogs relative to wolves, consistent with a selective sweep, and the pattern observed from dog and wolf whole genome sequencing (Fig. 4A). Blue, red, and yellow represent the reference (dog) allele homozygote (0/0), heterozygote (0/1), and alternative allele homozygote genotypes.

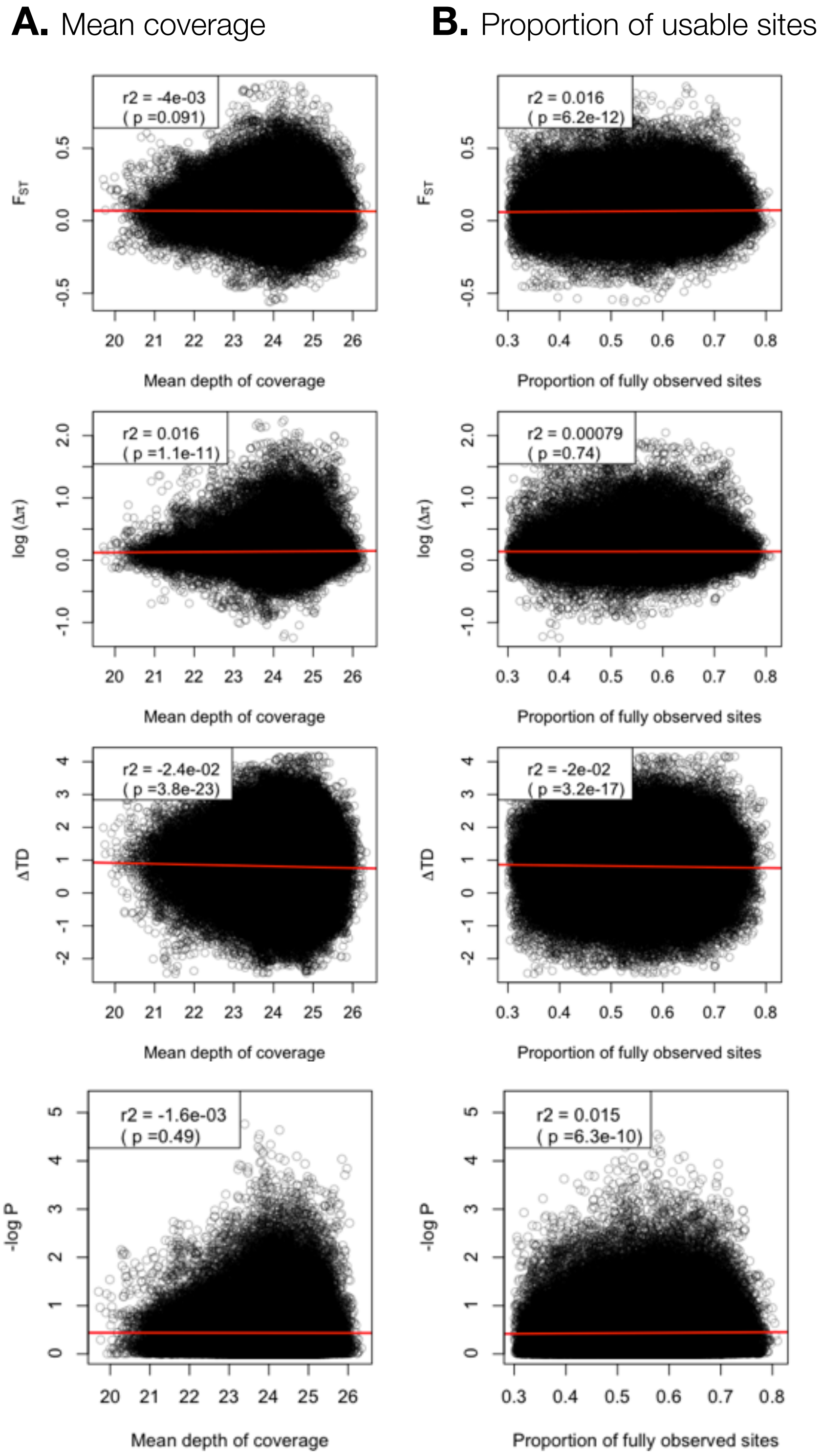
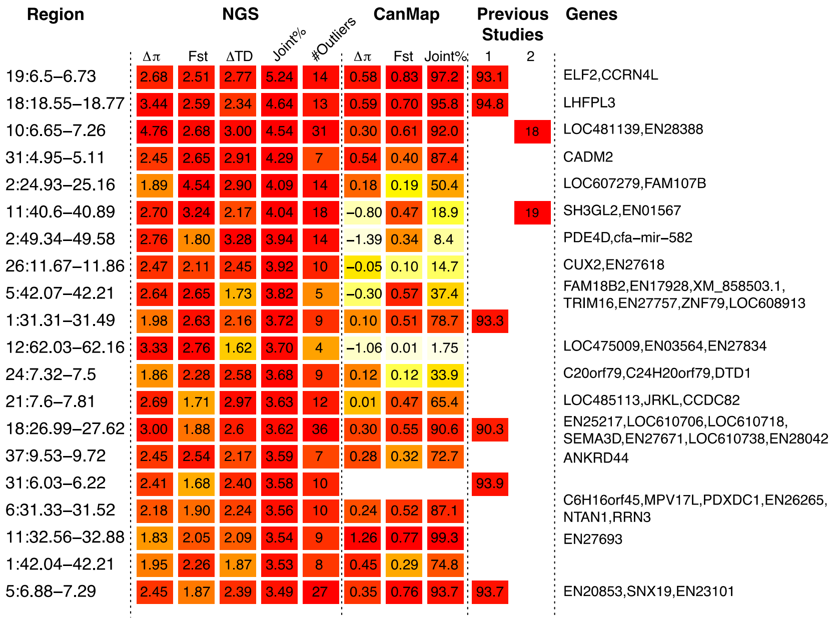


Figure 5.7: Correlation between selection scan metrics (F_{ST}, $\Delta\pi$, ΔTD and the joint empirical percentile) and mean depth of coverage in each window (A) and proportion of fully observable sites (B).



Region	NGS					CanMap			Previous Studies		Genes
	$\Delta\pi$	Fst	ΔTD	Joint%	#Outliers	$\Delta\pi$	Fst	Joint%	1	2	
19:6.5–6.73	2.68	2.51	2.77	5.24	14	0.58	0.83	97.2	93.1		ELF2,CCRN4L
18:18.55–18.77	3.44	2.59	2.34	4.64	13	0.59	0.70	95.8	94.8		LHFPL3
10:6.65–7.26	4.76	2.68	3.00	4.54	31	0.30	0.61	92.0		18	LOC481139,EN28388
31:4.95–5.11	2.45	2.65	2.91	4.29	7	0.54	0.40	87.4			CADM2
2:24.93–25.16	1.89	4.54	2.90	4.09	14	0.18	0.19	50.4			LOC607279,FAM107B
11:40.6–40.89	2.70	3.24	2.17	4.04	18	–0.80	0.47	18.9		19	SH3GL2,EN01567
2:49.34–49.58	2.76	1.80	3.28	3.94	14	–1.39	0.34	8.4			PDE4D,cfa–mir–582
26:11.67–11.86	2.47	2.11	2.45	3.92	10	–0.05	0.10	14.7			CUX2,EN27618
5:42.07–42.21	2.64	2.65	1.73	3.82	5	–0.30	0.57	37.4			FAM18B2,EN17928,XM_858503.1, TRIM16,EN27757,ZNF79,LOC608913
1:31.31–31.49	1.98	2.63	2.16	3.72	9	0.10	0.51	78.7	93.3		
12:62.03–62.16	3.33	2.76	1.62	3.70	4	–1.06	0.01	1.75			LOC475009,EN03564,EN27834
24:7.32–7.5	1.86	2.28	2.58	3.68	9	0.12	0.12	33.9			C20orf79,C24H20orf79,DTD1
21:7.6–7.81	2.69	1.71	2.97	3.63	12	0.01	0.47	65.4			LOC485113,JRKL,CCDC82
18:26.99–27.62	3.00	1.88	2.6	3.62	36	0.30	0.55	90.6	90.3		EN25217,LOC610706,LOC610718, SEMA3D,EN27671,LOC610738,EN28042
37:9.53–9.72	2.45	2.54	2.17	3.59	7	0.28	0.32	72.7			ANKRD44
31:6.03–6.22	2.41	1.68	2.40	3.58	10				93.9		
6:31.33–31.52	2.18	1.90	2.24	3.56	10	0.24	0.52	87.1			C6H16orf45,MPV17L,PDXDC1,EN26265, NTAN1,RRN3
11:32.56–32.88	1.83	2.05	2.09	3.54	9	1.26	0.77	99.3			EN27693
1:42.04–42.21	1.95	2.26	1.87	3.53	8	0.45	0.29	74.8			
5:6.88–7.29	2.45	1.87	2.39	3.49	27	0.35	0.76	93.7	93.7		EN20853,SNX19,EN23101

Figure 5.8: Top 20 outlier regions ranked by the joint empirical p-values. For each region (interval coordinates given in Mb), -log10 of the minimum value of empirical p-values of each test statistic for our next-generation sequencing data (NGS) is given within the rectangle, and rectangles are color coded by the percentile of the normalized test statistic (light yellow to red, 0-100th percentile. The ”number outliers” column gives the number of outliers within each region and is colored according to the percentile ranking. The value of each test statistic calculated using CanMap data is given within the rectangle, and rectangles are color coded by the empirical percentile (light yellow to red, 0-100th percentile, white is missing data). If applicable, the overlap of each region with any outliers from previous selection tests are given, i.e. joint percentile of F_{ST} and XP-EHH (Weir and Cockerham, 1984), F_{ST} (Boyko *et al.*, 2010; Vaysse *et al.*, 2011), or candidate domestication region number (Axelsson *et al.*, 2013). Finally, any genes that overlap the region are given, with a prefix of EN referring to Ensembl gene IDs, i.e. ENSCAFG.

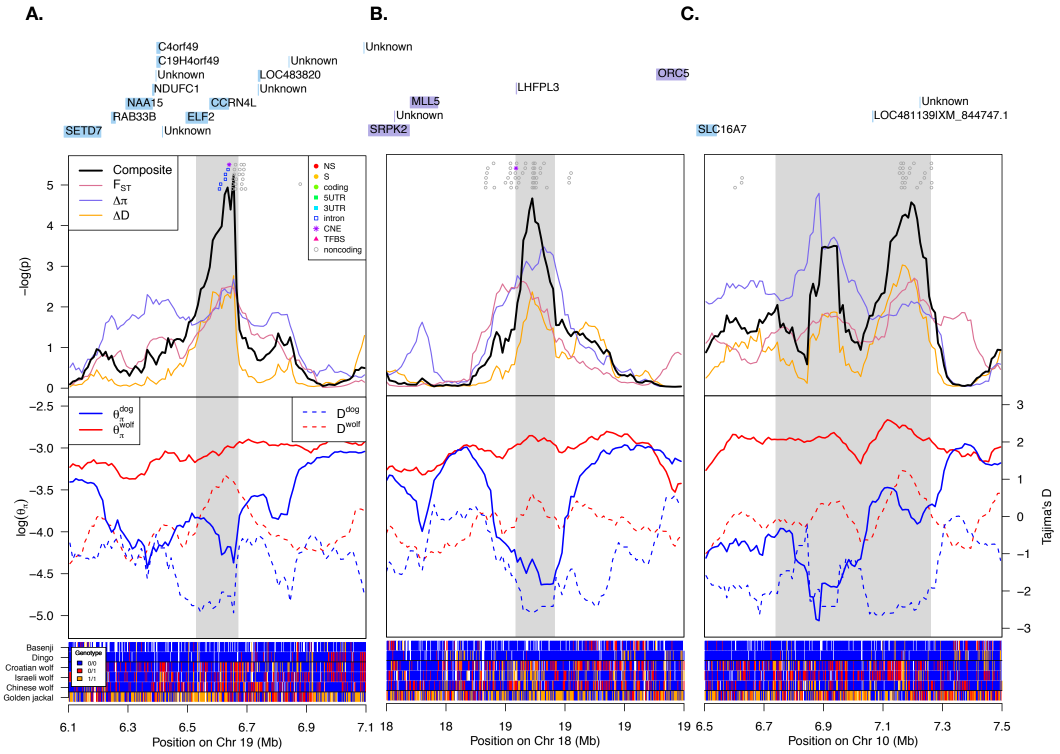


Figure 5.9: Top 3 outlier regions.

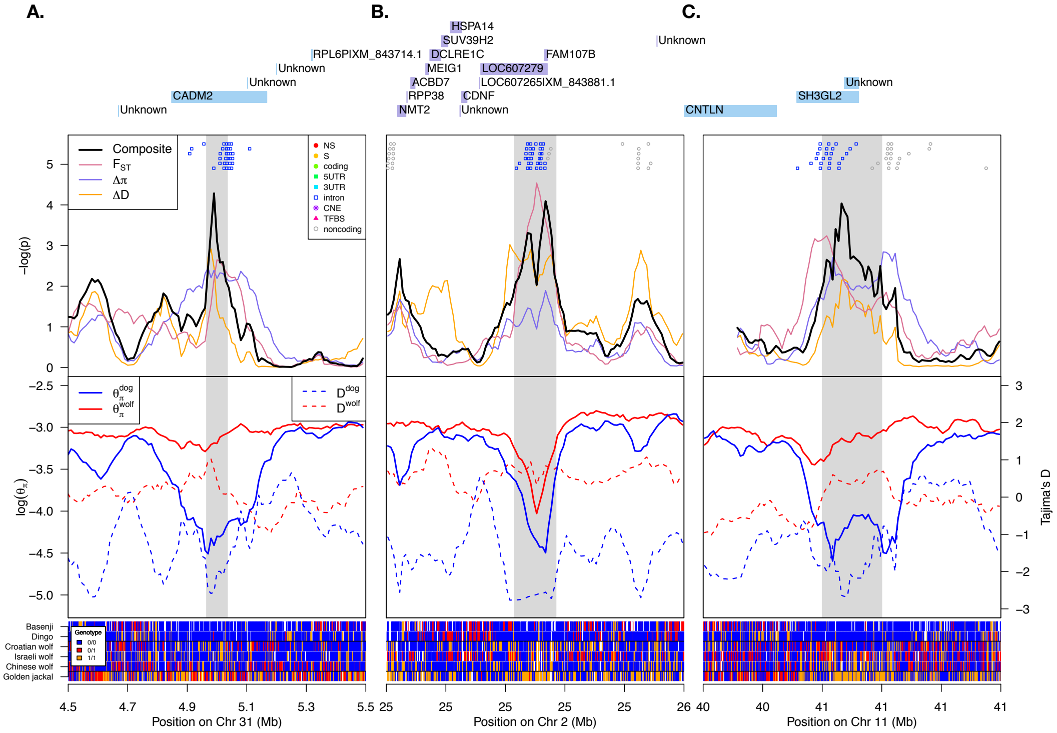


Figure 5.10: Top 4-6th outlier regions.

Table 5.2: gProfiler functional enrichment analysis for top 10% joint percentile (of selection scan statistics) outliers.

Term name	Category	Total # genes in category (T)	Total # genes in input list (Q)	# input genes in category (Q&T)	Fraction of input genes in term (Q&T/Q)	Fraction category genes detected in list (Q&T/T)	P^a
Abnormality of the joints of the upper limbs	HP:0009810	168	1704	29	0.160	0.173	0.00194
Hemangiomas	HP:0001028	32	1704	9	0.050	0.281	0.0034
Vascular neoplasia	HP:0100742	32	1704	9	0.050	0.281	0.0034
Narrow chest	HP:0000774	21	1704	7	0.039	0.333	0.00343
Abnormality of dental morphology	HP:0006482	38	1704	10	0.055	0.263	0.00347
Aplasia/Hypoplasia of the extremities	HP:0009815	159	1704	27	0.149	0.170	0.00361
End stage renal disease	HP:0003774	12	1704	5	0.028	0.417	0.0046
Abnormal number of teeth	HP:0006483	80	1704	16	0.088	0.200	0.0051
Abnormality of the radius	HP:0002818	54	1704	12	0.066	0.222	0.00631
Abnormality of the alveolar ridges	HP:0006477	13	1704	5	0.028	0.385	0.00686
Aplasia/Hypoplasia involving the skeleton	HP:0009115	259	1704	38	0.210	0.147	0.00802
Hypoplasia involving bones of the extremities	HP:0009826	91	1704	17	0.094	0.187	0.00808
Oculomotor apraxia	HP:0000657	19	1704	6	0.033	0.316	0.0091
Cone-shaped epiphyses of the phalanges of the hand	HP:0010230	14	1704	5	0.028	0.357	0.00982
Cone-shaped epiphyses	HP:0010579	14	1704	5	0.028	0.357	0.00982
Heterogeneous	HP:0001425	220	1704	33	0.182	0.150	0.00987
Hypoplastic ribs	HP:0000908	25	1704	7	0.039	0.280	0.00999
Aplasia/Hypoplasia involving bones of the lower limbs	HP:0006493	64	1704	13	0.072	0.203	0.01
Abnormality of the joints	HP:0001367	394	1704	53	0.293	0.135	0.0102
Cleft lip/palate	HP:0000202	116	1704	20	0.110	0.172	0.0104
Pes planus	HP:0001763	44	1704	10	0.055	0.227	0.0106
Abnormality of the 5th finger	HP:0004207	87	1704	16	0.088	0.184	0.0118
Short ribs	HP:0000773	20	1704	6	0.033	0.300	0.0119
Rhizomelic shortening	HP:0002968	20	1704	6	0.033	0.300	0.0119
Abnormality of the radial head	HP:0003995	20	1704	6	0.033	0.300	0.0119
Pectus carinatum	HP:0000768	32	1704	8	0.044	0.250	0.0123
Aplasia/Hypoplasia affecting bones of the axial skeleton	HP:0009122	190	1704	29	0.160	0.153	0.0124
Photophobia	HP:0000613	45	1704	10	0.055	0.222	0.0125
Hypodontia	HP:0000668	52	1704	11	0.061	0.212	0.0129
Abnormality of the teeth	HP:0000164	224	1704	33	0.182	0.147	0.013
Aplasia/Hypoplasia of the ribs	HP:0006712	33	1704	8	0.044	0.242	0.0148

Aplasia/Hypoplasia involving the metacarpal bones	HP:0005914	47	1704	10	0.055	0.213	0.0169
Atrial septal defect	HP:0001631	54	1704	11	0.061	0.204	0.0171
Cerebellar malformation	HP:0002438	54	1704	11	0.061	0.204	0.0171
Apraxia	HP:0002186	34	1704	8	0.044	0.235	0.0178
Abnormality of the cardiac septa	HP:0001671	99	1704	17	0.094	0.172	0.0186
Joint contractures involving the joints of the hand	HP:0009473	69	1704	13	0.072	0.188	0.0187
Abnormality of the cerebellar vermis	HP:0002334	41	1704	9	0.050	0.220	0.019
Osteopenia	HP:0000938	48	1704	10	0.055	0.208	0.0195
Aplasia/Hypoplasia involving bones of the thorax	HP:0006711	48	1704	10	0.055	0.208	0.0195
Reduced number of teeth	HP:0009804	70	1704	13	0.072	0.186	0.0209
Cerebellar vermis hypoplasia	HP:0001320	35	1704	8	0.044	0.229	0.0211
Abnormality of the atrial septum	HP:0001630	56	1704	11	0.061	0.196	0.0223
Abnormality of the ribs	HP:0000772	86	1704	15	0.083	0.174	0.0233
Abnormality of the iris	HP:0000525	71	1704	13	0.072	0.183	0.0234
Synostosis involving bones of the hand	HP:0004278	17	1704	5	0.028	0.294	0.0236
Abnormality of renal excretion	HP:0011036	17	1704	5	0.028	0.294	0.0236
Abnormality of the femoral head	HP:0003368	23	1704	6	0.033	0.261	0.024
Abnormality of the epiphysis of the femoral head	HP:0010574	23	1704	6	0.033	0.261	0.024
Abnormality of the sternum	HP:0000766	94	1704	16	0.088	0.170	0.0241
Abnormality of the forearm	HP:0002973	64	1704	12	0.066	0.188	0.0244
Abnormality of the metacarpal bones	HP:0001163	57	1704	11	0.061	0.193	0.0252
Abnormality of the pelvis	HP:0002644	159	1704	24	0.133	0.151	0.0258
Abnormality of the musculature of the upper limbs	HP:0001446	30	1704	7	0.039	0.233	0.0274
Abnormality of the cardiac atria	HP:0005120	58	1704	11	0.061	0.190	0.0284
Abnormality of the ilium	HP:0002867	37	1704	8	0.044	0.216	0.029
Abnormal iris pigmentation	HP:0008034	37	1704	8	0.044	0.216	0.029
Coloboma	HP:0000589	44	1704	9	0.050	0.205	0.0295
Abnormality involving the epiphyses of the limbs	HP:0006505	44	1704	9	0.050	0.205	0.0295
Sudden cardiac death	HP:0001645	18	1704	5	0.028	0.278	0.0301
Cardiac arrest	HP:0001695	18	1704	5	0.028	0.278	0.0301
Arteriosclerosis	HP:0002634	31	1704	7	0.039	0.226	0.0325
Thoracic hypoplasia	HP:0005257	31	1704	7	0.039	0.226	0.0325
Abnormality of the palate	HP:0000174	222	1704	31	0.171	0.140	0.033
Abnormality of the scapulae	HP:0000782	38	1704	8	0.044	0.211	0.0337
Abnormality of the lower limb	HP:0002814	438	1704	55	0.304	0.126	0.0351
Neoplasm of the skin	HP:0008069	75	1704	13	0.072	0.173	0.0355

Delayed closure of fontanelles	HP:0000270	25	1704	6	0.033	0.240	0.0356
Abnormality of femoral epiphyses	HP:0006499	25	1704	6	0.033	0.240	0.0356
Abnormality involving the epiphyses of the lower limbs	HP:0006500	25	1704	6	0.033	0.240	0.0356
Contractures of the joints of the upper limbs	HP:0100360	83	1704	14	0.077	0.169	0.0365
Patent ductus arteriosus	HP:0001643	53	1704	10	0.055	0.189	0.0373
Dislocated radial head	HP:0003083	19	1704	5	0.028	0.263	0.0376
Juvenile onset	HP:0003621	132	1704	20	0.110	0.152	0.0395
Short long bones	HP:0003026	69	1704	12	0.066	0.174	0.0418
Abnormality of the femoral neck and head region	HP:0003366	54	1704	10	0.055	0.185	0.0419
Cerebral edema	HP:0002181	26	1704	6	0.033	0.231	0.0425
Abnormality of the thorax	HP:0000765	262	1704	35	0.193	0.134	0.0433
Abnormality of the eyebrow	HP:0000534	85	1704	14	0.077	0.165	0.0438
Abnormality of the epiphyses of the phalanges of the hand	HP:0005920	20	1704	5	0.028	0.250	0.0461
Abnormality of the elbow	HP:0009811	78	1704	13	0.072	0.167	0.0472
Phosphoric diester hydrolase activity	GO:0008081	62	1709	19	0.016	0.306	0.05
Autosomal dominant inheritance	HP:0000006	751	1704	87	0.481	0.116	0.05
Chemokine signaling pathway	KEGG:04062	154	1709	28	0.054	0.182	0.05

[a] Corrected for multiple testing.

CHAPTER 6

Conclusions

6.1 Contribution

This dissertation addresses the challenges of analyzing NGS data for population genetics inferences and provides recommendations and guidelines to interpret such data with precision. In this dissertation, we emphasize the importance of the pipeline used to analyze NGS data to avoid any potential bias problems and incorrect conclusions. We showed that depending on the NGS data analysis pipeline, one can reach starkly different conclusions with the same data set. Simple allele counting after inferring individual genotypes from aligned sequencing data (call-based approach) leads to bias in the estimated SFS toward the sites with rare variants, whereas the SFS directly estimated from aligned sequencing data (direct estimation approach) was almost unbiased across ranges of coverage. Furthermore, we demonstrated that the bias in the inferred SFS subsequently results in bias in θ estimators, neutrality test, and demographic inference.

Next, we proposed the new adaptive K-restricted algorithm by which we can speed-up the original dynamic programming to compute site likelihood vectors. This algorithm exploits the observation that for most sites the site likelihood vectors probability mass is concentrated on a few cells around the best-guess allele counts and approximates the site likelihood vector by curtailing computation to only those K components of the dynamic programming update vectors. We showed that the adaptive K-restricted algorithm has comparable accuracy, but is faster than the original dynamic programming algorithm. Moreover, as a sample size increases, there will be even more dramatic differences in the computation time of the two algorithms. This speed improvement with the adaptive K-restricted algorithm will greatly facilitate direct inference of the SFS even when the number of individuals is large.

Finally, as an applied study, we analyzed high-coverage sequencing data of two dogs and three wolves to detect genetic signatures of adaptation during early dog domestication. Our top selection hit, a CCRN4L gene, showed the importance of dietary evolution in early dog domestication. This gene might be also involved in bone growth, because it affects cell fate via the growth regulator, IGF1. Furthermore, 4 of our top 8 selection regions each contain a gene known to impact memory and behavior in mice and humans. This confirms previous studies which concluded that domestication focused on genes involved in cognition. Finally, we found that gene categories involved in skeletal and dental morphology, and genes in these categories are enriched for our top selection hits, implying that genes in these categories may have contributed to early dogs having shortened, broader skulls, more extreme tooth crowding, smaller carnassials, and reduced body size. (I want to emphasize that Adam H. Freedman, Robert K. Wayne, John Novembre and I were all involved in the interpretation of the results pertaining to selection scans.)

6.2 Future Work

Estimation of 2-dimensional SFS As an extension to the EM algorithm for estimating the 1-dimensional SFS, we can also derive an EM algorithm for estimating the 2-dimensional SFS.

Suppose that we have sequencing data from two populations, denoted by D_1 and D_2, with a sample size of n_1 and n_2, respectively. The 2-dimensional SFS is defined as a $(2n_1 + 1) \times$

$(2n_2 + 1)$ matrix:

$$\xi = \begin{pmatrix} \xi_{0,0} & \xi_{0,1} \cdots & \xi_{0,2n_2} \\ \xi_{1,0} & \xi_{1,1} \cdots & \xi_{1,2n_2} \\ \vdots & \vdots & \vdots \\ \xi_{2n_1,0} & \xi_{2n_1,1} \cdots & \xi_{2n_1,2n_2} \end{pmatrix}$$

where $\xi_{i,j}$ represents a proportion of sites with the derived allele frequency of $i/(2n_1)$ in population 1 and the derived allele frequency of $j/(2n_1)$ in population 2.

By the EM algorithm, we can iteratively update the 2-dimensional SFS as follows:

$$\begin{aligned} \xi_{x,y}^{(t+1)} &= \frac{1}{l} \sum_{i=1}^{l} \frac{P(D_{1,i}, D_{2,i} | X_{1,i} = x, X_{2,i} = y) \xi_{x,y}^{(t)}}{\sum_{k=0}^{2n_1} \sum_{l=0}^{2n_2} P(D_{1,i}, D_{2,i} | X_{1,i} = k, X_{2,i} = l) \xi_{k,l}^{(t)}} \\ &= \frac{1}{l} \sum_{i=1}^{l} \frac{P(D_{1,i} | X_{1,i} = x) P(D_{2,i} | X_{2,i} = y) \xi_{x,y}^{(t)}}{\sum_{k=0}^{2n_1} \sum_{l=0}^{2n_2} P(D_{1,i} | X_{1,i} = k) P(D_{2,i} | X_{2,i} = l) \xi_{k,l}^{(t)}} \quad \text{by conditional independence} \\ &= \frac{1}{l} \sum_{i=1}^{l} \frac{h_{i,x}^1 h_{i,y}^2 \xi_{x,y}^{(t)}}{\sum_{k=0}^{2n_1} \sum_{l=0}^{2n_2} h_{i,k}^1 h_{i,l}^2 \xi_{k,l}^{(t)}} \end{aligned}$$

where $h_{i,x}^1$ denotes a site likelihood function for a derived allele frequency of $x/(2n_1)$ in population 1 and $h_{i,y}^2$ denotes a site likelihood function for derived allele frequency of $y/(2n_2)$ in population 2.

As this calculation includes a computation of site likelihood vectors for each population independently (i.e. $boldh_i^1$ and $\mathbf{h_i^2}$ separately), this implies that we can make the direct estimation method even faster and more efficient with the adaptive K-restricted algorithm - the computation time for running the original algorithm is $O(n_1^2 + n_2^2)$, whereas the runtime of the adaptive K-restricted algorithm becomes $O(K_1 n_1 + K_2 n_2)$.

However, we found that estimation of the 2-dimensional SFS is not as precise as estimation of the 1-dimensional SFS (Figure 6.1). This is because we have a lot more parameters to estimate for the 2-dimensional SFS compared to the 2-dimensional SFS. Hence, future works are required to improve the performance of the direct estimation method for estimating the 2-dimensional SFS.

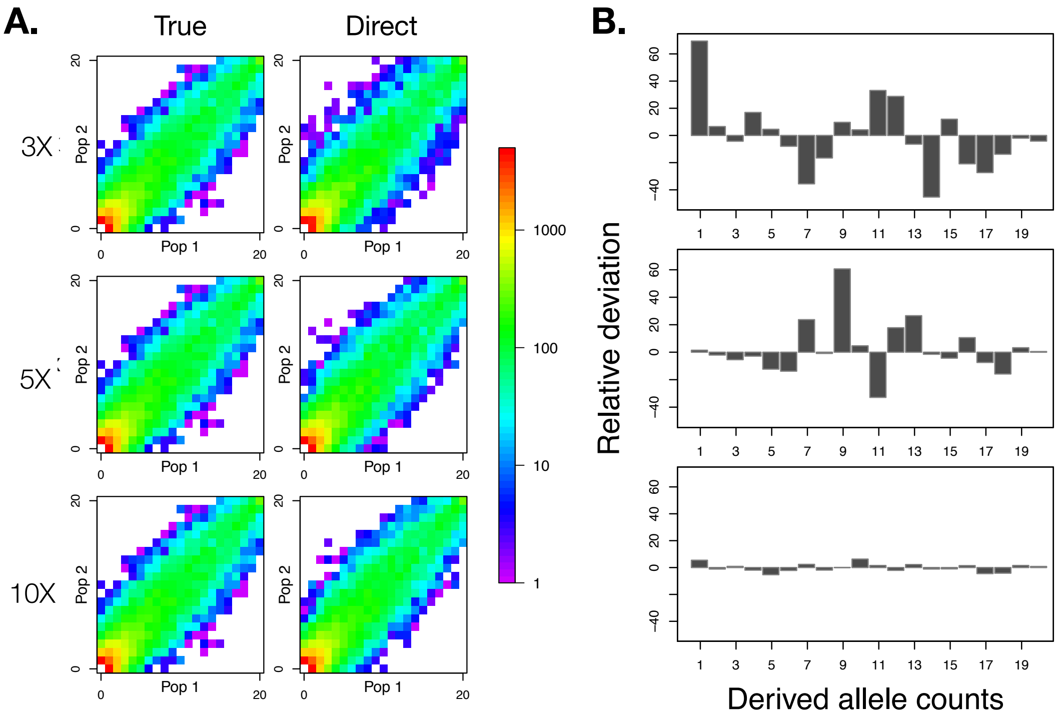


Figure 6.1: Evaluation of the accuracy of the inferred 2-dimensional SFS. We simulated 100 replicates of sequencing data for 10 diploid individuals for each population each from genomic regions of length 100Kb under the standard model. **A**. Shapes of the inferred 2-dimensional SFS (right) compared with the ground-truth SFS (left) for coverage 3X (top), 5X (middle), and 10X (bottom). **B**. Relative deviation of a fraction of sites with the derived allele count of 1-20.

Direct estimation method vs. multisample calling with genotype imputation With the sequencing data from the 1000 Genomes Project, we observed a lack of rare variants in the SFS inferred from the genotype calls (using the VCF file) compared to the SFS inferred directly from the aligned short-read sequencing data (using a set of the BAM files). To conclude that this difference is not the result of artifacts, we need to verify the results using simulations. We can simulate sequencing data with the same sample size used in the 1000 Genomes Project using the inferred human demographic model (such as one in Nelson et al.), and then compare the inferred SFS with the direct estimation method to the inferred SFS with the call-based approach (genotypes are inferred with a multisample calling pipeline and genotype imputation).

BIBLIOGRAPHY

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65.

Achaz, G. (2008). Testing for neutrality in samples with sequencing errors. *Genetics*, **179**(3), 1409–1424.

Achaz, G. (2009). Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, **183**(1), 249–258.

Andersen, E. C., Gerke, J. P., Shapiro, J. A., Crissman, J. R., Ghosh, R., Bloom, J. S., Felix, M. A., and Kruglyak, L. (2012). Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nat. Genet.*, **44**(3), 285–290.

Andolfatto, P. (2007). Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. *Genome Res.*, **17**(12), 1755–1762.

Axelsson, E., Ratnakumar, A., Arendt, M. L., Maqbool, K., Webster, M. T., Perloski, M., Liberg, O., Arnemo, J. M., Hedhammar, A., and Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, **495**(7441), 360–364.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**(1), 379–406.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y. P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., and Langley, C. H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.*, **5**(11), e310.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met*, **57**(1), 289–300.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**(2), 573–580.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–59.

Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., Lohmueller, K. E., Zhao, K., Brisbin, A., Parker, H. G., vonHoldt, B. M., Cargill, M., Auton, A., Reynolds, A., Elkahloun, A. G., Castelhano, M., Mosher, D. S., Sutter, N. B., Johnson, G. S., Novembre, J., Hubisz, M. J., Siepel, A., Wayne, R. K., Bustamante, C. D., and Ostrander, E. A. (2010). A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol.*, **8**(8), e1000451.

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**(2), 210–223.

Browning, B. L. and Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.*, **85**(6), 847–861.

Casey, J. P., Magalhaes, T., Conroy, J. M., Regan, R., Shah, N., Anney, R., Shields, D. C., Abrahams, B. S., Almeida, J., Bacchelli, E., Bailey, A. J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bolton, P. F., Bourgeron, T., Brennan, S., Cali, P., Correia, C., Corsello, C., Coutanche, M., Dawson, G., de Jonge, M., Delorme, R., Duketis, E., Duque, F., Estes, A., Farrar, P., Fernandez, B. A., Folstein, S. E., Foley, S., Fombonne, E., Freitag, C. M., Gilbert, J., Gillberg, C., Glessner, J. T., Green, J., Guter, S. J., Hakonarson, H., Holt, R., Hughes, G., Hus, V., Igliozzi, R., Kim, C., Klauck, S. M., Kolevzon, A., Lamb, J. A., Leboyer, M., Le Couteur, A., Leventhal, B. L., Lord, C., Lund, S. C., Maestrini, E., Mantoulan, C., Marshall, C. R., McConachie, H., McDougle, C. J., McGrath, J., McMahon, W. M., Merikangas, A., Miller, J., Minopoli, F., Mirza, G. K., Munson, J., Nelson, S. F., Nygren, G., Oliveira, G., Pagnamenta, A. T., Papanikolaou, K., Parr, J. R., Parrini, B., Pickles, A., Pinto, D., Piven, J., Posey, D. J., Poustka, A., Poustka, F., Ragoussis, J., Roge, B., Rutter, M. L., Sequeira, A. F., Soorya, L., Sousa, I., Sykes, N., Stoppioni, V., Tancredi, R., Tauber, M., Thompson, A. P., Thomson, S., Tsiantis, J., Van Engeland, H., Vincent, J. B., Volkmar, F., Vorstman, J. A., Wallace, S., Wang, K., Wassink, T. H., White, K., Wing, K., Wittemeyer, K., Yaspan, B. L., Zwaigenbaum, L., Betancur, C., Buxbaum, J. D., Cantor, R. M., Cook, E. H., Coon, H., Cuccaro, M. L., Geschwind, D. H., Haines, J. L., Hallmayer, J., Monaco, A. P., Nurnberger, J. I., Pericak-Vance, M. A., Schellenberg, G. D., Scherer, S. W., Sutcliffe, J. S., Szatmari, P., Vieland, V. J., Wijsman, E. M., Green, A., Gill, M., Gallagher, L., Vicente, A., and Ennis, S. (2012). A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Hum. Genet.*, **131**(4), 565–579.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.*, **22**(11), 3124–3140.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., Wheeler, D. A., Sabo, A., Lusk, C., Weiss, K. G., Akbar, H., Cree, A., Hawes, A. C., Newsham, I., Varghese, R. T., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Iv, W. H., Templeton, A. R., Boerwinkle, E., Gibbs, R., and Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*, **1**, 131.

Cubelos, B., Sebastian-Serrano, A., Beccari, L., Calcagnotto, M. E., Cisneros, E., Kim, S., Dopazo, A., Alvarez-Dolado, M., Redondo, J. M., Bovolenta, P., Walsh, C. A., and Nieto, M. (2010). Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron*, **66**(4), 523–535.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and Group, . G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics*, **27**(15), 2156–2158.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**(5), 491–498.

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

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet.*, **9**(10), e1003905.

Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, **155**(3), 1405–1413.

Franco, S. J., Gil-Sanz, C., Martinez-Garay, I., Espinosa, A., Harkins-Perry, S. R., Ramos, C., and Muller, U. (2012). Fate-restricted neural progenitors in the mammalian cerebral cortex. *Science*, **337**(6095), 746–749.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., Akey, J. M., Gabriel,

S. B., Altshuler, D. M., Goncalo R, A., Allayee, H., Cresci, S., Daly, M. J., de Bakker, P. I., DePristo, M. A., Do, R., Donnelly, P., Farlow, D. N., Fennell, T., Garimella, K., Hazen, S. L., Hu, Y., Jordan, D. M., Jun, G., Kathiresan, S., Kang, H. M., Kiezun, A., Lettre, G., Li, B., Li, M., Newton-Cheh, C. H., Padmanabhan, S., Peloso, G., Pulit, S., Rader, D. J., Reich, D., Reilly, M. P., Rivas, M. A., Schwartz, S., Hutchinson, F., Scott, L., Siscovick, D. S., Spertus, J. A., Stitziel, N. O., Stoletzki, N., Sunyaev, S. R., Voight, B. F., Willer, C. J., Rich, S. S., Akylbekova, E., Atwood, L. D., Ballantyne, C. M., Barbalic, M., Barr, R., Benjamin, E. J., Bis, J., Boerwinkle, E., Bowden, D. W., Brody, J., Budoff, M., Burke, G., Buxbaum, S., Carr, J., Chen, D. T., Chen, I. Y., Chen, W. M., Concannon, P., Crosby, J., Cupples, L. A., D'Agostino, R., DeStefano, A. L., Dreisbach, A., Dupuis, J., Durda, J., Ellis, J., Folsom, A. R., Fornage, M., Fox, C. S., Fox, E., Funari, V., Ganesh, S. K., Gardin, J., Goff, D., Gordon, O., Grody, W., Gross, M., Guo, X., Hall, I. M., Heard-Costa, N. L., Heckbert, S. R., Heintz, N., Herrington, D. M., Hickson, D., Huang, J., Hwang, S. J., Jacobs, D. R., Jenny, N. S., Johnson, A. D., Johnson, C. W., Kawut, S., Kronmal, R., Kurz, R., Lange, E. M., Lange, L. A., Larson, M., Lawson, M., Lewis, C. E., Levy, D., Li, D., Lin, H., Liu, C., Liu, J., Liu, K., Liu, X., Liu, Y., Longstreth, W. T., Loria, C., Lumley, T., Lunetta, K., Mackey, A. J., Mackey, R., Manichaikul, A., Maxwell, T., McKnight, B., Meigs, J. B., Morrison, A. C., Musani, S. K., Mychaleckyj, J. C., Nettleton, J. A., North, K., O'Donnell, C. J., O'Leary, D., Ong, F., Palmas, W., Pankow, J. S., Pankratz, N. D., Paul, S., Perez, M., Person, S. D., Polak, J., Post, W. S., Psaty, B. M., Quinlan, A. R., Raffel, L. J., Ramachandran, V. S., Reiner, A. P., Rice, K., Rotter, J. I., Sanders, J. P., Schreiner, P., Seshadri, S., Shea, S., Sidney, S., Silverstein, K., Siscovick, D. S., Smith, N. L., Sotoodehnia, N., Srinivasan, A., Taylor, H. A., Taylor, K., Thomas, F., Tracy, R. P., Tsai, M. Y., Volcik, K. A., Wassel, C. L., Watson, K., Wei, G., White, W., Wiggins, K. L., Wilk, J. B., Williams, O., Wilson, G., Wilson, J. G., Wolf, P., Zakai, N. A., Hardy, J., Meschia, J. F., Nalls, M., Rich, S. S., Singleton, A., Worrall, B., Bamshad, M. J., Barnes, K. C., Abdulhamid, I., Accurso, F., Anbar, R., Beaty, T., Bigham, A., Black, P., Bleecker, E., Buckingham, K., Cairns, A. M., Chen,

W. M., Caplan, D., Chatfield, B., Chidekel, A., Cho, M., Christiani, D. C., Crapo, J. D., Crouch, J., Daley, D., Dang, A., Dang, H., De Paula, A., DeCelie-Germana, J., Dozor, A., Drumm, M., Dyson, M., Emerson, J., Emond, M. J., Ferkol, T., Fink, R., Foster, C., Froh, D., Gao, L., Gershan, W., Gibson, R. L., Godwin, E., Gondor, M., Gutierrez, H., Hansel, N. N., Hassoun, P. M., Hiatt, P., Hokanson, J. E., Howenstine, M., Hummer, L. K., Kanga, J., Kim, Y., Knowles, M. R., Konstan, M., Lahiri, T., Laird, N., Lange, C., Lin, L., Lin, X., Louie, T. L., Lynch, D., Make, B., Martin, T. R., Mathai, S. C., Mathias, R. A., McNamara, J., McNamara, S., Meyers, D., Millard, S., Mogayzel, P., Moss, R., Murray, T., Nielson, D., Noyes, B., O'Neal, W., Orenstein, D., O'Sullivan, B., Pace, R., Pare, P., Parker, H., Passero, M. A., Perkett, E., Prestridge, A., Rafaels, N. M., Ramsey, B., Regan, E., Ren, C., Retsch-Bogart, G., Rock, M., Rosen, A., Rosenfeld, M., Ruczinski, I., Sanford, A., Schaeffer, D., Sell, C., Sheehan, D., Silverman, E. K., Sin, D., Spencer, T., Stonebraker, J., Tabor, H. K., Varlotta, L., Vergara, C. I., Weiss, R., Wigley, F., Wise, R. A., Wright, F. A., Wurfel, M. M., Zanni, R., Zou, F., Nickerson, D. A., Rieder, M. J., Green, P., Shendure, J., Akey, J. M., Bamshad, M. J., Bustamante, C. D., Crosslin, D. R., Eichler, E. E., Fox, P. K., Fu, W., Gordon, A., Gravel, S., Jarvik, G. P., Johnsen, J. M., Kan, M., Kenny, E. E., Kidd, J. M., Lara-Garduno, F., Leal, S. M., Liu, D. J., McGee, S., O'Connor, T. D., Paeper, B., Robertson, P. D., Smith, J. D., Tennessen, J. A., Turner, E. H., Wang, G., Jackson, R., North, K., Peters, U., Carlson, C. S., Anderson, G., Anton-Culver, H., Assimes, T. L., Auer, P. L., Hutchinson, F., Beresford, S., Hutchinson, F., Bizon, C., Black, H., Brunner, R., Brzyski, R., Burwen, D., Caan, B., Carty, C. L., Chlebowski, R., Cummings, S., Curb, J. D., Eaton, C. B., Ford, L., Franceschini, N., Fullerton, S. M., Gass, M., Geller, N., Heiss, G., Howard, B. V., Hsu, L., Hutter, C. M., Ioannidis, J., Jiao, S., Johnson, K. C., Kooperberg, C., Kuller, L., LaCroix, A., Lakshminarayan, K., Lane, D., Lange, E. M., Lange, L. A., Lasser, N., LeBlanc, E., Lewis, C. E., Li, K. P., Limacher, M., Lin, D. Y., Logsdon, B. A., Ludlam, S., Manson, J. E., Margolis, K., Martin, L., McGowan, J., Monda, K. L., Kotchen, J. M., Nathan, L., Ockene, J., O'Sullivan, M. J., Phillips, L. S., Prentice, R. L., Reiner, A. P., Hutchinson, F., Robbins, J., Robinson, J. G., Rossouw, J. E., Sangi-Haghpeykar, H., Sarto, G. E., Shumaker, S., Simon, M. S., Stefanick, M. L., Stein, E., Tang, H., Taylor, K. C., Thomson, C. A., Thornton, T. A., Van Horn, L., Vitolins, M., Wactawski-Wende, J., Wallace, R., Wassertheil-Smoller, S., Zeng, D., Applebaum-Bowden, D., Feolo, M., Gan, W., Paltoo, D. N., Rossouw, J. E., Sholinsky, P., and Sturcke, A. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**(7431), 216–220.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147**(2), 915–925.

Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, **133**(3), 693–709.

Gray, M. M., Granka, J. M., Bustamante, C. D., Sutter, N. B., Boyko, A. R., Zhu, L., Ostrander, E. A., and Wayne, R. K. (2009). Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, **181**(4), 1493–1505.

Green, C. B., Douris, N., Kojima, S., Strayer, C. A., Fogerty, J., Lourim, D., Keller, S. R., and Besharse, J. C. (2007). Loss of Nocturnin, a circadian deadenylase, confers resistance to hepatic steatosis and diet-induced obesity. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(23), 9888–9893.

Grossman, S. R., Shlyakhter, I., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**(5967), 883–886.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, **5**(10), e1000695.

Han, E., Sinsheimer, J. S., and Novembre, J. (2014). Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol. Biol. Evol.*, **31**(3), 723–735.

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

Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, **12**(11), 756–766.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., and Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.*, **6**(2), e1000862.

Hoopes, B. C., Rimbault, M., Liebers, D., Ostrander, E. A., and Sutter, N. B. (2012). The insulin-like growth factor 1 receptor (IGF1R) contributes to reduced size in dogs. *Mamm. Genome*, **23**(11-12), 780–790.

Innan, H. and Kim, Y. (2004). Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. U.S.A.*, **101**(29), 10667–10672.

Innan, H. and Kim, Y. (2008). Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics*, **179**(3), 1713–1720.

Johnson, P. L. and Slatkin, M. (2008). Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.*, **25**(1), 199–206.

Jordan, G. and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.*, **29**(4), 1125–1139.

Kang, C. J. and Marjoram, P. (2011). Inference of population mutation rate and detection of segregating sites from next-generation sequence data. *Genetics*, **189**(2), 595–605.

Kawai, M. and Rosen, C. J. (2010). PPAR: a circadian transcription factor in adipogenesis and osteogenesis. *Nat Rev Endocrinol*, **6**(11), 629–636.

Kawai, M., Green, C. B., Lecka-Czernik, B., Douris, N., Gilbert, M. R., Kojima, S., Ackert-Bicknell, C., Garg, N., Horowitz, M. C., Adamo, M. L., Clemmons, D. R., and Rosen, C. J. (2010a). A circadian-regulated gene, Nocturnin, promotes adipogenesis by stimulating PPAR-gamma nuclear translocation. *Proc. Natl. Acad. Sci. U.S.A.*, **107**(23), 10508–10513.

Kawai, M., Delany, A. M., Green, C. B., Adamo, M. L., and Rosen, C. J. (2010b). Nocturnin suppresses igf1 expression in bone by targeting the 3' untranslated region of igf1 mRNA. *Endocrinology*, **151**(10), 4861–4870.

Keightley, P. D. and Halligan, D. L. (2011). Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics*, **188**(4), 931–940.

Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**(6082), 740–743.

Kim, S. Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet. Epidemiol.*, **34**(5), 479–491.

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, **12**, 231.

Larson, G., Karlsson, E. K., Perri, A., Webster, M. T., Ho, S. Y., Peters, J., Stahl, P. W., Piper, P. J., Lingaas, F., Fredholm, M., Comstock, K. E., Modiano, J. F., Schelling, C., Agoulnik, A. I., Leegwater, P. A., Dobney, K., Vigne, J. D., Vila, C., Andersson, L., and Lindblad-Toh,

K. (2012). Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc. Natl. Acad. Sci. U.S.A.*, **109**(23), 8878–8883.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21), 2987–2993.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**(11), 1851–1858.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009b). Genotype imputation. *Annu Rev Genomics Hum Genet*, **10**, 387–406.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., DeJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C. W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K. P., Parker, H. G., Pollinger, J. P., Searle, S. M., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A. C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J. P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., and Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**(7069), 803–819.

Liu, X., Maxwell, T. J., Boerwinkle, E., and Fu, Y. X. (2009). Inferring population mutation rate and sequencing error rate using the SNP frequency spectrum in a sample of DNA sequences. *Mol. Biol. Evol.*, **26**(7), 1479–1490.

Liu, X., Fu, Y. X., Maxwell, T. J., and Boerwinkle, E. (2010). Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. *Genome Res.*, **20**(1), 101–109.

Lynch, M. (2008). Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.*, **25**(11), 2409–2419.

Lynch, M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, **182**(1), 295–301.

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R., Barron, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., Harris, Z., Javaid, M., Jayaseelan, J. C., Jhangiani, S. N., Jordan, K. W., Lara, F., Lawrence, F., Lee, S. L., Librado, P., Linheiro, R. S., Lyman, R. F., Mackey, A. J., Munidasa, M., Muzny, D. M., Nazareth, L., Newsham, I., Perales, L., Pu, L. L., Qu, C., Ramia, M., Reid, J. G., Rollmann, S. M., Rozas, J., Saada, N., Turlapati, L., Worley, K. C., Wu, Y. Q., Yamamoto, A., Zhu, Y., Bergman, C. M., Thornton, K. R., Mittelman, D., and Gibbs, R. A. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, **482**(7384), 173–178.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**(7), 499–511.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–380.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**(9), 1297–1303.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**(1), 31–46.

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S. A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M. D., Nangle, K., Wang, J., Abecasis, G., Cardon, L. R., Zollner, S., Whittaker, J. C., Chissoe, S. L., Novembre, J., and Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**(6090), 100–104.

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

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**(6), 443–451.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS ONE*, **7**(7), e37558.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., Sullivan, P. F., Bergen, S., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Purcell, S. M., Haas, D. W., Liang, L., Sunyaev, S., Patterson, N., de Bakker, P. I., Reich, D., and Price, A. L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.*, **44**(6), 631–635.

Pennings, P. S. and Hermisson, J. (2006a). Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.*, **23**(5), 1076–1084.

Pennings, P. S. and Hermisson, J. (2006b). Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.*, **2**(12), e186.

Pool, J. E., Hellmann, I., Jensen, J. D., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Res.*, **20**(3), 291–300.

Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.*, **20**(4), R208–215.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.*, **39**(Web Server issue), W307–315.

Rutten, K., Misner, D. L., Works, M., Blokland, A., Novak, T. J., Santarelli, L., and Wallace, T. L. (2008). Enhanced long-term potentiation and impaired learning in phosphodiesterase 4D-knockout (PDE4D) mice. *Eur. J. Neurosci.*, **28**(3), 625–632.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Johnson, T. A., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**(7164), 913–918.

Savolainen, P., Leitner, T., Wilton, A. N., Matisoo-Smith, E., and Lundeberg, J. (2004). A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **101**(33), 12387–12390.

Schmidt, A., Wolde, M., Thiele, C., Fest, W., Kratzin, H., Podtelejnikov, A. V., Witke, W., Huttner, W. B., and Soling, H. D. (1999). Endophilin I mediates synaptic vesicle formation by transfer of arachidonate to lysophosphatidic acid. *Nature*, **401**(6749), 133–141.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**(10), 1135–1145.

Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, **141**(1), 413–429.

Smit, A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0. *Nucleic Acids Res.*

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

Sutter, N. B., Bustamante, C. D., Chase, K., Gray, M. M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P. G., Quignon, P., Johnson, G. S., Parker, H. G., Fretwell, N., Mosher, D. S., Lawler, D. F., Satyaraj, E., Nordborg, M., Lark, K. G., Wayne, R. K., and Ostrander, E. A. (2007). A Single IGF1 Allele Is a Major Determinant of Small Size in Dogs. *Science*, **316**(5821), 112–115.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2), 437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**(3), 585–595.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., and Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**(6090), 64–69.

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., Sigurdsson, S., Fall, T., Seppala, E. H., Hansen, M. S., Lawley, C. T., Karlsson, E. K., Bannasch, D., Vila, C., Lohi, H., Galibert, F., Fredholm, M., Haggstrom, J., Hedhammar, A., Andre, C., Lindblad-Toh, K., Hitte, C., Webster, M. T., Battaille, G., Clercx, C., Druet, T., Farnir, F., Faust, N., Georges, M., Lequarre, A. S., Merveille, A. C., Momozawa, Y., Pamplona, M. R., Peeters, D., Ahlgren, K., Andersson, L., Arendt, M., Barrio, A. M., Carlborg, O., Johansson, C., Kampe, O., Lindblad-Toh, K., Meadows, J., Pielberg, G., Ratnakumar, A., Tengvall, K., Webster, M., Andersson, G., Bjornerfeldt, S., Gustafsson, S., Haggstrom, J., Hansson, J., Hedhammar, A., Hoglund, K., Kierczak, M., Ljungvall, I., Melin, M., Sundberg, K., Kennedy, L., Massey, J., Ollier, W., Rothwell, S., Ahonen, S., Hytonen, M., Kyostila, K., Lohi, H., Nevalainen, E., Fredholm, M., Madsen, M. B., Mogensen, M. S., Courtay-Cahen, C., Mellersh, C., Starkey, M., Dahlgren, S., Lingaas, F., Storengen, L. M., Wiik, C., Boerkamp, K., Fieten, H., Leegwater, P., Rutteman, G., Van Steenbeek, F., Leeb, T., Owczarek-Lipska, M., Aguirre-Hernandez, J., Sargan, D., Andre, C., Hitte, C., Thomas, A., Chetboul, V., Gouni, V., Tiret, L., Catchpole, B., Hendricks, A., Bonastre, A. S., Quilez, J., Callanan, S., Nolan, C., Dukes-McEwan, J., Copeland-Stephenson, H., Favrot, C., Wess, G., Wolf, J., Millar, K., Boland, A., Lathrop, M., Zelenika, D., Quinnell, R. J., and Philipp, U. (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.*, **7**(10), e1002316.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.*, **4**(3), e72.

Vonholdt, B. M., Pollinger, J. P., Lohmueller, K. E., Han, E., Parker, H. G., Quignon, P., Degenhardt, J. D., Boyko, A. R., Earl, D. A., Auton, A., Reynolds, A., Bryc, K., Brisbin, A., Knowles, J. C., Mosher, D. S., Spady, T. C., Elkahloun, A., Geffen, E., Pilot, M., Jedrzejewski, W., Greco, C., Randi, E., Bannasch, D., Wilton, A., Shearman, J., Musiani, M., Cargill, M., Jones, P. G., Qian, Z., Huang, W., Ding, Z. L., Zhang, Y. P., Bustamante, C. D., Ostrander, E. A., Novembre, J., and Wayne, R. K. (2010). Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature*, **464**(7290), 898–902.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, **7**(2), 256–276.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, **38**(6), 1358–1370.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z., Pool, J., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, P., Ni, G., Tian, J., Xu, X., Liu, T., Jiang, R., Wu, G., Zhou, M., Tang, J., Qin, T., Wang, S., Feng, G., Li, J., Huasang, W., Luosang, F., Wang, Y., Chen, X., Wang, Z., Z, Z., Li, G., Bianba, X., Yang, S., Wang, G., Tang, Y., Gao, Z., Chen, L., Luo, Z., Gusang, Q., Cao, W., Zhang, X., Ouyang, H., Ren, H., Liang, Y., Zheng, J., Huang, L., Li, K., Bolund, Y., Kristiansen, Y., Li, X., Zhang, R., Zhang, S., Li, H., Li, R., Yang, J., Nielsen, J., and Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, (5987), 75–8.