Genetics: Early Online, published on August 26, 2013 as 10.1534/genetics.113.154740

Quantifying population genetic differentiation from Next-Generation Sequencing data

Matteo Fumagalli[*], Filipe G. Vieira[*], Thorfinn Sand Korneliussen[§], Tyler Linderoth[*], Emilia Huerta-Sánchez[*], Anders Albrechtsen[†], Rasmus Nielsen[*,†,‡]

[*]Department of Integrative Biology, University of California, Berkeley, CA, USA, 94720
[§]Centre for GeoGenetics, Natural History Museum of Denmark and Department of Biology, University of Copenhagen, Copenhagen, Denmark, 2100
[†]Department of Biology, University of Copenhagen, Copenhagen, Denmark, 2200
[‡]Department of Statistics, University of California, Berkeley, CA, USA, 94720

Running Head: Genetic variation from NGS data

Keywords: Next-Generation Sequencing, F_{ST}, Principal Components Analysis

Corresponding author
Matteo Fumagalli
Department of Integrative Biology
University of California
4134 Valley Life Sciences Building
Berkeley, CA 94720, USA
Phone +1 510 643 0060
email: matteo.fumagalli@berkeley.edu

Abstract

Over the last few years, new high-throughput DNA sequencing technologies have dramatically increased speed and reduced sequencing costs. However, the use of these sequencing technologies is often challenged by errors and biases associated with the bioinformatical methods used for analyzing the data. In particular, the use of naïve methods to identify polymorphic sites and infer genotypes can inflate downstream analyses. Recently, explicit modelling of genotype probability distributions has been proposed as a method for taking genotype call uncertainty into account. Based on this idea, we propose a novel method for quantifying population genetic differentiation from next-generation sequencing data. In addition, we present a strategy to investigate population structure via Principal Components Analysis.

Through extensive simulations, we compare the new method herein proposed to approaches based on genotype calling and demonstrate a marked improvement in estimation accuracy for a wide range of conditions.

We apply the method to a large-scale genomic data set of domesticated and wild silkworms sequenced at low coverage. We find that we can infer the fine-scale genetic structure of the sampled individuals, suggesting that employing this new method is useful for investigating the genetic relationships of populations sampled at low coverage.

Introduction

Determining the level of genetic variation within and between species or populations is necessary to study the effects of mutation, natural selection and genetic drift. In the last few years, faster and cheaper high-throughput DNA sequencing technologies have provided us with an unprecedented amount of large-scale genetic data. These Next-Generation Sequencing (NGS) technologies are now commonly used in population genetic studies, and provide us with the perfect opportunity to investigate the evolutionary forces affecting genetic variation.

Currently, available NGS technologies differ in their protocol design (reviewed in Metzker, 2010) but all produce data with similar general features. Briefly, the sequencing output consists of relatively short streches (e.g. currently about 50-100 base pairs for Illumina machines) of sequenced DNA, commonly called "reads". These small segments of DNA are then aligned to a reference genome or assembled into scaffolds in *de novo* assembly when a reference genome is not available.

These technologies have greatly improved sequencing efforts in both model and non-model organisms, but they have also introduced new challenges because many of the datasets produced using these methods are sequenced at low coverage (a position in the genome is only covered by few sequencing reads), and raw sequencing error rates are often higher than observed using Sanger sequencing. Under such circumstances, it is often difficult to distinguish between a variable site and a sequencing error, making the identification of variable sites in the sample (a procedure known as "SNP calling") non-trivial and prone to error. Also, determining the genotype for each individual ("genotype calling") can be unreliable due to uncertainty whether both the parental chromosomes were sampled. Therefore, sequencing errors and uncertainty in the genotype calls may lead to a biased allele frequency distribution (Johnson and Slatkin, 2008; Hellmann et al., 2008).

Accurate estimation of the site frequency spectrum (SFS), however, is important for population genetic inferences of demography, natural selection, and population structure. Indeed, many summary statistics for evolutionary inferences are functions of the sample allele frequencies (Nielsen, 2005). Higher sequencing coverage lowers the uncertainty, but with a fixed budget, researchers need to choose between sequencing fewer samples at higher coverage or sequencing more samples at low to medium coverage. The latter was the preferred option for many recent large-scale sequencing population genetic studies (1000 Genomes Project Consortium, 2010; 1000 Genomes Project Consortium et al., 2012; Auton et al., 2012; Huang et al., 2012).

Naïve methods for estimating allele frequencies, that are primarily based on direct counting of sequencing reads, provide inaccurate estimates of local nucleotide diversity (Nielsen et al., 2011). Consequently, there have been numerous efforts to use statistical models to analyze NGS data in order to provide more accurate estimates of allele frequencies. To this end, Maximum Likelihood (ML) methods and Bayesian methods have been developed for estimating the allele frequency at any given site (Lynch, 2009; Keightley and Halligan, 2011; Kim et al., 2011) or the entire distribution of allele frequencies jointly across multiple sites (Nielsen et al., 2012; Li, 2011; Keightley and Halligan, 2011). Bayesian methods incorporate base quality scores and statistical uncertainty to obtain posterior probabilities associated with each genotype. Recent studies incorporate this probabilistic approach to estimate population genetic parameters from NGS data (Yi et al., 2010; Gompert and Buerkle, 2011;

Kang and Marjoram, 2011; Li, 2011; Gompert et al., 2012).

Thanks to these approaches, genome-wide scans of positive selection have been possible in samples sequenced at moderate coverage. For example, in Yi et al. (2010), 50 Tibetan individuals were sequenced to identify the regions of the genome involved in the adaptation to high altitude. Species of rice (Xu et al., 2011), chicken (Rubin et al., 2010) and silkworm (Xia et al., 2009) have also been sequenced at low coverage to identify functional differences between domesticated and wild populations.

In genome-wide scans for selection, it is often informative to summarize genetic variation using population differentiation statistics, such as F_{ST} (Wright, 1951), to identify particular regions of the genome that are highly differentiated relative to the rest of the genome. F_{ST} can also be informative about the divergence time between two populations. Another powerful tool for the analysis of genetic data is Principal Component Analysis (PCA). This data reduction method is a convenient way to visualize the data, derive corrections for population stratification in association studies, and investigate specific features of population history and differentiation. Both PCA and F_{ST} have been used extensively for the last 30 years, and continue to be valuable tools in summarizing genetic variation.

However, as we will show, traditional methods for computing F_{ST} and performing PCA result in biases when applied to genotype calls from low or moderate coverage NGS data. Therefore, we propose a new method to estimate F_{ST} from NGS data that accounts for uncertainty in the genotype calls. Furthermore, we also show that population structure can be investigated with PCA under the proposed probabilistic framework that accounts for sequencing errors. These new methods outperform previous approaches, especially in the case of low-coverage sequencing data as determined from simulated sequences. Finally, we demonstrate the power of the proposed methods by applying it to a previously published dataset of wild and domesticated samples of *Bombyx mori* (Xia et al., 2009).

The methods developed in this study contribute to the current tool-kit for population genetic analyses of Next-Generation Sequencing data, and can be applied to both model and non-model organisms.

Materials and Methods

Measuring genetic differentiation between populations

F_{ST} is a measure of population genetic differentiation which quantifies the proportion of variance in allele frequencies among populations relative to the total variance (the sum of the variance within individuals, within populations, and between populations). Several estimators of F_{ST} have been proposed through the years (reviewed in Weir and Hill, 2002; Holsinger and Weir, 2009).

There is considerable debate about definitions of F_{ST}. Some researchers consider F_{ST} to be a model parameter (e.g. Balding and Nichols, 1995; Nicholson et al., 2002; Holsinger et al., 2002), while others consider it to be a statistic (e.g. Reynolds et al., 1983; Weir and Cockerham, 1984; Hudson et al., 1992). Even when considering F_{ST} as a parameter, there is considerable discussion about what model it is a parameter of, and how it should be estimated (Marchini and Cardon, 2002; Balding, 2003). The objective of this paper is not to compare these approaches, which differ both in what they estimate and in how the estimation procedure works. We will remain agnostic with regard to the debate on interpretation and definition of F_{ST}, although we will use the word 'estimator' throughout. Instead, we will show how some of the most commonly applied estimators of F_{ST} can be modified in the presence of low and medium coverage data to more accurately reflect what the original F_{ST} estimators were intended to capture, i.e. the objective will be to derive estimators applicable to NGS data that produce results similar to those that would have been obtained from the original estimator based on full genotype data without any errors. As a note, other estimators, not considered here, could potentially be modified in a similar fashion.

Method-of-moments estimation

We will start by considering the most simple method-of-moments estimators of F_{ST}. They do not rely on any assumptions about the shape of the sampling distribution, beyond the moments used to estimate the parameters, and they are easy to implement through simple algebraic expressions. For these reasons, method-of-moments estimators are popular and often used.

Our first aim is to extend the method-of-moments F_{ST} estimator proposed by Reynolds et al. (1983), as this is one of the most popular and well-motivated estimators of F_{ST}, to take into account genotyping uncertainty. Assuming a bi-allelic SNP, with non-reference allele at estimated frequencies of $\hat{p}_i$, $\hat{p}_j$, and $\hat{p}$ for population i, j, and pooled, the genetic variance between and within populations at site s is respectively:

$$a_s = \frac{4n_i(\hat{p}_{(i,s)} - \hat{p}_s)^2 + 4n_j(\hat{p}_{(j,s)} - \hat{p}_s)^2 - b_s}{2(\frac{2n_i n_j}{n_i+n_j})} \tag{1}$$

and

$$b_s = \frac{n_i \alpha_{(i,s)} + n_j \alpha_{(j,s)}}{n_i + n_j - 1} \tag{2}$$

where n_i and n_j are the number of sampled individuals per population, $\alpha_{(i,s)} = 2\hat{p}_{(i,s)}(1 - \hat{p}_{(i,s)})$, and $\alpha_{(j,s)} = 2\hat{p}_{(j,s)}(1 - \hat{p}_{(j,s)})$. Table 1 describes nomenclature used throughout this manuscript.

The estimate of F_{ST} for a single site is then:

$$F_{ST} = \frac{a_s}{a_s + b_s} \tag{3}$$

while for a *locus* of m sites is:

$$F_{ST}^{(locus)} = \frac{\sum_{s=1}^{m} a_s}{\sum_{s=1}^{m} (a_s + b_s)}. \tag{4}$$

Maximum Likelihood Estimation

Maximum Likelihood (ML) methods for estimating F_{ST} require the specification of a sampling probability distribution. Once this distribution is defined, one can maximize a likelihood function to obtain ML estimators for the parameters of the distribution. ML estimators of F_{ST} have been very popular, particularly for detecting signatures of adaptive natural selection among populations (e.g. Beaumont and Balding, 2004; Riebler et al., 2008; Foll and Gaggiotti, 2008).

Assuming a bi-allelic site s with beta-distributed allele frequencies, the probability of the sample allele frequencies $\hat{p}_{(i,s)}$ at population i can be expressed as a beta-binomial distribution with parameters $2n_i$ (sample size), F_{ST}, and $p_{anc,s}$, the ancestral population allele frequency. This parameterization assumes divergence from a common ancestral population, and that the subsequent divergence is well-modeled by the beta-distribution. The marginal sampling distribution in population i is then given by (Balding and Nichols, 1995; Balding, 2003):

$$P(\hat{p}_{(i,s)} = \frac{k}{2n_i} | p_{anc,s}, F_{ST}) = \binom{2n_i}{k} \frac{B(k+\alpha, 2n_i - k + \beta)}{B(\alpha, \beta)} \tag{5}$$

where k is the count of the non-reference (or derived) allele, B is the Beta function and

$$\alpha = \frac{p_{anc,s}(1 - F_{ST})}{F_{ST}} \tag{6}$$

and

$$\beta = \frac{(1 - p_{anc,s})(1 - F_{ST})}{F_{ST}}. \tag{7}$$

The full likelihood function is the product of this sampling distribution for all populations, as the populations are independent conditional on $p_{anc,s}$. For two populations i and j, we have:

$$P(\hat{p}_{(i,s)} = \frac{k}{2n_i}, \hat{p}_{(j,s)} = \frac{z}{2n_j} | p_{anc,s}, F_{ST}) = P(\hat{p}_{(i,s)} = \frac{k}{2n_i} | p_{anc,s}, F_{ST}) P(\hat{p}_{(j,s)} = \frac{z}{2n_j} | p_{anc,s}, F_{ST}) \tag{8}$$

where the subscripts on n and $\hat{p}$ indicate population identity. We numerically maximize Equation 8 using the Broyden - Fletcher - Goldfarb - Shanno (BFGS) algorithm (Fletcher, 1987; Press et al., 2007).

Quantifying population genetic differentiation by calling genotypes

A naïve strategy for estimating sample allele frequencies and F_{ST} is to first call genotypes at each site, and then simply count the occurrence of non-reference or derived alleles among all individuals.

We first assessed the accuracy of several genotype calling strategies (Supporting Text). These methods include approaches based on direct counts of read bases, on genotype likelihoods, and on genotype posterior probabilities. One promising approach is to use Bayesian methods to assign individual genotypes by computing genotype posterior probabilities $P(G|X)$ from genotype likelihoods and a specific prior $P(G)$ on genotype G. Bayes' theorem is used to calculate $P(G|X)$, the posterior probability of genotype G given the observed data X (1000 Genomes Project Consortium, 2010). The prior can be defined using extraneous data, such as the reference sequence, sequences in a data base, an estimate of the allele frequency and/or inbreeding coefficients, etc. (e.g. 1000 Genomes Project Consortium, 2010; Li, 2011; Nielsen et al., 2012).

We calculate genotype posterior probabilities at site s for individual w, $P(G_{(w,s)}|X_{(w,s)})$ as

$$P(G_{(w,s)}|X_{(w,s)}) = \frac{P(X_{(w,s)}|G_{(w,s)})P(G_{(w,s)})}{\sum_{G=0}^{2} P(X_{(w,s)}|G_{(w,s)})P(G_{(w,s)})} \tag{9}$$

where $P(X_{(w,s)}|G_{(w,s)})$ are the genotype likelihoods and $P(G_{(w,s)})$ is the prior probability of genotype G at site s under Hardy-Weinberg Equilibrium (HWE). The prior is calculated from estimates of the per-site population allele frequencies using the method described in Kim et al. (2011). To call genotypes, the genotype with the highest posterior probability was chosen for each individual.

Results show that calling genotypes from genotype posterior probabilities provides the most stable and accurate genotype and SNP calling accuracy at almost all tested experimental scenarios (Tables S1-S3). We adopted this strategy to call genotypes throughout the rest of the study. Specifically, we counted non-reference alleles from these called genotypes to infer allele frequencies and computed a method-of-moments estimator of F_{ST}, which we labeled $\hat{F}_{ST.GC}$ (Equations 10-11). We adopted this genotype calling strategy to compute a ML estimator of F_{ST}, $\hat{F}_{ST.ML.GC}$ (Equations 5, 8).

An alternative strategy for computing F_{ST} is to avoid genotype calling altogether so that inference is based directly on the posterior probabilities (e.g. Yi et al., 2010; Nielsen et al., 2012). We will describe such methods in the following sections.

Quantifying population genetic differentiation without calling genotypes

Here we propose to use a Bayesian probabilistic framework to estimate F_{ST} from posterior probabilities of sample allele frequencies of each population at each site without calling specific genotypes. In our applications, we compute a Maximum-Likelihood estimate of the Site Frequency Spectrum from genotype likelihoods, as previously proposed by Nielsen et al. (2012). Using this ML estimate of the SFS as a prior in an Empirical Bayes approach, we estimate the posterior probability for all possible allele frequencies at each site (Nielsen et al., 2012).

Method-of-moments estimation

Let $\pi_i^{(k)} = P(\hat{p_i} = k/(2n_i)|Y_{(i,s)})$ be the posterior probability that a site in population i has derived sample allele frequency $\hat{p_i} = k/(2n_i)$, in a sample of n_i diploid individuals, given the read data $Y_{(i,s)}$. This probability can be calculated from the genotype probabilities using the algorithm in Nielsen et al. (2012). Allele labeling with respect to the derived allele is arbitrary and any other labeling of alleles could have been chosen if identification of the ancestral and derived state is not possible.

From these quantities, we compute the posterior expectation of the genetic variance between and within populations (see Equations 1 and 2) at site s as:

$$E[a_s|Y_s] = \sum_{k=0}^{2n_i}\sum_{z=0}^{2n_j} a_{(i,j)}^{(k,z)} \pi_{(i,j,s)}^{(k,z)} \tag{10}$$

and

$$E[b_s|Y_s] = \sum_{k=0}^{2n_i}\sum_{z=0}^{2n_j} b_{(i,j)}^{(k,z)} \pi_{(i,j,s)}^{(k,z)} \tag{11}$$

where $a_{(i,j)}^{(k,z)}$ and $b_{(i,j)}^{(k,z)}$ are genetic variances from Reynolds et al. (1983) formula, with k and z derived alleles in populations i and j respectively, and Y_s is the sequencing data at site s. The total expected variance, $E[c_s|Y_s]$, at each site, is then $E[c_s|Y_s] = E[a_s|Y_s] + E[b_s|Y_s]$.

The estimate of F_{ST} for a single site is given by the ratio of $E[a_s|Y_s]$ to $E[c_s|Y_s]$ (Equation 3). However, since the two variance components are not independent and this calculation involves the expectation of a ratio, we approximate it using the delta method (Rice, 2008; Rice and Papadopoulos, 2009) to obtain the following estimator of F_{ST} at site s:

$$\hat{F}_{ST.Ev} = E[\frac{a_s}{c_s}|c_s \neq 0, Y_s] = \frac{E[a_s|Y_s]}{E[c_s|Y_s]} + \sum_{u=1}^{\infty}(-1)^u \frac{E[a_s|Y_s]\langle c_u\rangle + \langle a, c_u\rangle}{E[c|Y_s]^{i+1}} \tag{12}$$

where $\langle c_u\rangle$ is the $u-th$ central moment of c_s and $\langle a, c_u\rangle$ is the mixed central moment, which can be calculated as:

$$\langle c_u\rangle = E[(c_s - E[c_s])^u|Y_s] = \sum_{k=0}^{2n_i}\sum_{z=0}^{2n_j}(c_{(i,j)}^{(k,z)} - E[c_s|Y_s])^u \pi_{(i,j,s)}^{(k,z)} \tag{13}$$

and

$$\langle a, c_u\rangle = E[(a_s - E[a_s|Y_s])(c_s - E[c_s|Y_s])^u|Y_s] = \sum_{k=0}^{2n_i}\sum_{z=0}^{2n_j}(a_{(i,j)}^{(k,z)} - E[a_s|Y_s])(c_{(i,j)}^{(k,z)} - E[c_s|Y_s])^u \pi_{(i,j,s)}^{(k,z)} \tag{14}$$

where $c_{(i,j)}^{(k,z)}$ is the total genetic variance from Reynolds et al. (1983) formula, with k and z derived alleles in populations i and j respectively. For computational purposes, we only use the first central and mixed central moments.

$\pi_{(i,j,s)}^{(k,z)}$ can be calculated using maximum likelihood similarly to the method used for calculating $\pi_{(i,s)}^{(k)}$ for a single population (Nielsen et al., 2012). However, this calculation may not be desirable due to the high variance associated with the estimation of so many parameters.

An alternative approach is to compute an estimate of the 2-Dimensional Site Frequency Spectrum (2D-SFS), $S_{(i,j)}^{(k,z)}$, as:

$$S_{(i,j)}^{(k,z)} = \frac{1}{\sum_{s=0}^{m}\sum_{k=0}^{2n_i}\sum_{z=0}^{2n_j}(h_{(i,s)}^{(k)} h_{(j,s)}^{(z)})} \sum_{s=0}^{m} h_{(i,s)}^{(k)} h_{(j,s)}^{(z)} \tag{15}$$

where $h^{(k)}_{(i,s)}$ and $h^{(z)}_{(j,s)}$ are the marginal likelihoods of observing k and z non-reference alleles at population i and j, respectively, at site s, as presented in Nielsen et al. (2012).

$S^{(k,z)}_{(i,j)}$ is then used as a prior to compute the posterior probability of quantities of interest. For instance, the expectation of the genetic variance between populations (see Equation 10) can be computed as:

$$E[a_s|Y_s] = \sum_{k=0}^{2n_i} \sum_{z=0}^{2n_j} a^{(k,z)}_{(i,j)} h^{(k)}_{(i,s)} h^{(z)}_{(j,s)} S^{(k,z)}_{(i,j)}. \tag{16}$$

Finally, a method-of-moments estimator of F_{ST} over m sites is given by Equation 4. When analyzing multiple sites, we do not add the correction factor to the ratio of $E[a|X]$ to $E[c|X]$ at each site because, for a large number of sites, the error introduced by taking the ratio of two non-independent expectations will be minimal. We also tested the performance of other methods to estimate F_{ST} from sequencing data derived from the expectations of sample allele frequencies (Supporting Text).

These methods can be extended to non-pairwise definitions of F_{ST} (Weir, 1996). These formulations require the estimation of a joint SFS among all populations, which can be estimated in a similar fashion as in Equation 15.

Maximum Likelihood Estimation

We also extend the procedure for ML estimation of F_{ST} and p_{anc} under the Beta-Binomial distribution (Balding and Nichols, 1995; Balding, 2003) (Equation 8) to the case of unknown genotypes. These estimates, which we call $F_{ST.ML}$, are obtained by maximizing the following likelihood function:

$$\begin{aligned} & P(Y_{(i,s)}, Y_{(j,s)}|p_{anc,s}, F_{ST}) \\ & = \sum_{k=0}^{2n_i} \sum_{z=0}^{2n_j} P(Y_{(i,s)}|\hat{p}_{(i,s)} = \frac{k}{2n_i}) P(\hat{p}_{(i,s)} = \frac{k}{2n_i}|p_{anc,s}, F_{ST}) P(Y_{(j,s)}|\hat{p}_{(j,s)} = \frac{z}{2n_j}) P(\hat{p}_{(j,s)} = \frac{z}{2n_j}|p_{anc,s}, F_{ST}) \\ & = \sum_{k=0}^{2n_i} \sum_{z=0}^{2n_j} h^{(k)}_{(i,s)} P(\hat{p}_{(i,s)} = \frac{k}{2n_i}|p_{anc,s}, F_{ST}) h^{(z)}_{(j,s)} P(\hat{p}_{(j,s)} = \frac{z}{2n_j}|p_{anc,s}, F_{ST}) \end{aligned} \tag{17}$$

where $Y_{(i,s)}$ and $Y_{(j,s)}$ are the observed read data at site s for population i and j, respectively, and $h^{(k)}_{(i,s)}$ and $h^{(z)}_{(j,s)}$ are again the marginal likelihoods of the sample allele frequency for population i and j, computed as in Nielsen et al. (2012).

Principal Components Analysis

A similar approach to the one used for correcting estimates of F_{ST} can be used in Principal Component Analyses (PCA). The now-standard method for calculation PCA in population genetics is based on Patterson et al. (2006). For n individuals and m sites a normalized covariance matrix C is calculated as:

$$C_{(w,y)} = \frac{1}{m} \sum_{s=1}^{m} \frac{(G_{(w,s)} - 2\hat{p}_s)(G_{(y,s)} - 2\hat{p}_s)}{\hat{p}_s(1 - \hat{p}_s)} \tag{18}$$

where $\hat{p_s}$ is the derived allele frequency at site s (the labelling is again arbitrary) and $G_{(w,s)}$ is the number of derived alleles for individual w at site s ($G \in \{0, 1, 2\}$ in the diploid case). The denominator is inserted to account for genetic drift and normalizes the standardized allele frequencies to have the same variance (Patterson et al., 2006). However, other normalizations can be chosen. An eigenvector decomposition of C is then computed.

We propose to compute an estimate of $C_{(w,y)}$ by integrating over the posterior genotype probabilities at site s for individual w, $P(G_{(w,s)}|X_{(w,s)})$, and y, $P(G_{(y,s)}|X_{(y,s)})$, which both can be calculated as in Equation 9. The prior is calculated using the sample allele frequencies $\hat{p}_s$ at site s as in Kim et al. (2011). Therefore, $P(G_{(w,s)} = 2) = \hat{p}_s^2$, $P(G_{(w,s)} = 1) = 2\hat{p}_s(1-\hat{p}_s)$, $P(G_{(w,s)} = 0) = (1 - \hat{p}_s)^2$, where $G_{(w,s)}$ is the number of derived alleles for individual i at site s. Missing genotype data is then implicitly incorporated in a Bayesian manner using the prior from the sample allele frequencies.

Additionally, the C matrix is weighted by the probability of each site being variable. This is motivated by the fact that, at low to medium sequencing coverage, sites which have a small probability of being variable in the sample can have a small but non-negligible contribution to the matrix C. As there are several orders of magnitude more invariable than variable sites, this can have a profound effect on the analyses, even when weighting with genotype probabilities. Instead of using an arbitrary discrete SNP calling, or minor allele frequency, cut-off, we propose to weight sites according to their probability of being variable.

We, therefore, estimate the matrix C as follows (for $w \neq y$):

$$C_{(w,y)} = \frac{1}{\sum_{s=1}^{m} P_{var,s}} \sum_{s=1}^{m} \frac{(\sum_{G_{(w,s)}=0}^{2} \sum_{G_{(y,s)}=0}^{2} (G_{(w,s)} - 2\hat{p}_s)(G_{(y,s)} - 2\hat{p}_s) P(G_{(w,s)}|X_{(w,s)}) P(G_{(y,s)}|X_{(y,s)})) P_{var,s}}{\hat{p}_s(1 - \hat{p}_s)} \tag{19}$$

where the probability of site s being variable, $P_{var,s}$, is computed as:

$$P_{var,s} = 1 - (\pi_s^{(0)} + \pi_s^{(2n)}). \tag{20}$$

We emphasize that this approach does not provide a form of Bayesian PCA analysis. Rather, it is a modification of the Patterson et al. (2006) approach for PCA analysis in the context of population genetics, modified to incorporate uncertainty in genotype calls by using an appropriate weighting of different genotypes using their respective posterior probabilities.

Notice that we estimate the joint posterior of the genotype probabilities for the two individuals using the product of their marginal genotype probabilities, i.e. we estimate $P(G_{(w,s)}, G_{(y,s)}|X_{(w,s)}, X_{(y,s)})$ by $P(G_{(w,s)}|X_{(w,s)})P(G_{(y,s)}|X_{(y,s)})$. $P(G_{(w,s)}|X_{(w,s)})$ and $P(G_{(y,s)}|X_{(y,s)})$ are not independent as they are correlated through the underlying estimate of genotype frequencies affecting the prior. However, as these analyses are carried-out conditional on an estimated allele frequency, the approximation is accurate, although it ignores the sampling variance in the estimate of the allele frequency. Conditional on the allele frequency, $P(G_{(w,s)}|X_{(w,s)})$ and $P(G_{(y,s)}|X_{(y,s)})$ are independent.

We also notice that:

$$E[\sum_{G_{(w,s)}=0}^{2}\sum_{G_{(y,s)}=0}^{2}(G_{(w,s)}-2\hat{p}_s)(G_{(y,s)}-2\hat{p}_s)P(G_{(w,s)}|X_{(w,s)})P(G_{(y,s)}|X_{(y,s)})]=0 \tag{21}$$

for unrelated individuals under HWE assuming known allele frequencies and a HWE-derived prior for the genotype probabilities. This shows that the covariance function for unrelated individuals is in fact expected to be zero using this estimator, a necessary and desirable property for the method to perform well. Proof of Equation 21 is provided in the Appendix. As we shall argue, the resulting PCA is greatly improved over naïve methods using genotype calling under all the explored scenarios.

This approach could be extended to different strategies to perform PCA from a matrix of genotype posterior probabilities, for instance ML methods that account for noise contributions of each variable (Wentzell et al., 1997), or Bayesian methods that use external information about the data (Nounou et al., 2002).

Simulating sequencing data for multiple populations

We performed simulations to compare the performance of these methods to estimate population genetic differentiation, as well as to quantify the genotyping and SNP calling accuracy, under a broad range of experimental conditions. As done in previous studies (Kim et al., 2010, 2011), we simulated sequencing data rather than raw sequencing reads for computational efficiency. We treated sites as independent of each other and simulated genotypes for each individual assuming HWE and a specific population allele frequency. Specifically, we repeated the following procedure for each site.

First, for each site, we drew an ancestral allele frequency p_{anc} from a distribution in $[5\times10^{-3}, 1-(5\times10^{-3})]$ with density proportional to $1/x$. This distribution is the expected allele frequency distribution under a standard neutral infinite sites model, truncated at the boundaries corresponding to a population size of 200 individuals (see e.g. Ewens, 2004). We then simulated allele frequencies for two populations using the Balding-Nichols model (Balding and Nichols, 1995) with mean equal to p_{anc}, as in previous studies (Pritchard and Donnelly, 2001; Price et al., 2006). We simulated two independent samples, conditionally on F_{ST} and p_{anc}, from this distribution to obtain allele frequencies for two populations (see Equation 5). From these population allele frequencies, we assigned genotypes according to HWE for each individual.

To simulate data from 3 populations, we first drew population allele frequencies from the Balding-Nichols model for two populations as described above. We then assigned the first allele frequency to population 1 and used the second allele frequency as the ancestral allele frequency for populations 2 and 3. We then drew two population allele frequencies from the Balding-Nichols model for a different value of F_{ST} and assigned these allele frequencies to populations 2 and 3.

In order to simulate NGS data, the number of reads at each locus for each individual was simulated from a Poisson distribution as in Kim et al. (2010, 2011). Additionally, errors were randomly introduced uniformly among nucleotides at a rate of 0.0075. This value is comparable to error rates found in previous studies (1000 Genomes Project Consortium, 2010; Li et al., 2010; Yi et al., 2010). The probability of a site being polymorphic, P_{var}, was varied from 0.02 to 1.

We computed genotype likelihoods from simulated sequencing reads. Genotype likelihoods depend on both base calls and quality scores and are proportional to the probability, $P(X|G)$, of the observed read data, X, at a site for each individual given a certain genotype G. In the simplest possible case, for read z at site s, we calculated the genotype likelihood of a particular base v, $L_{(z,v,s)}$ with $v \in \{A, C, G, T\}$ as $L_{(z,v,s)} = (1-e)$ if v is the observed base at read z, and $L_{(z,v,s)} = e/3$ otherwise. Here e is the sequencing error used in the simulation setting. There are many other methods for estimating e, including methods for estimating it directly from the data (e.g. Kim et al., 2011). Genotype likelihoods at site s for individual w are then calculated by taking the product of the likelihoods over all r reads:

$$P(X_{(w,s)}|G_{(w,s)} = v_1v_2) = \frac{1}{2^r}\prod_{z=1}^{r}(L_{(z,v_1,s)} + L_{(z,v_2,s)}). \tag{22}$$

Using this procedure, we computed genotype likelihoods for each individual at each site for all 10 possible genotypes. We then computed posterior probabilities of genotypes and sample allele frequencies, as previously described (see Equation 9).

When calling genotypes, we assigned genotypes with a posterior probability lower than 0.90 as missing data. We removed sites where more than half of the individuals had missing genotypes. With this procedure, we filtered approx. 25% of the total sites at $2X$ sequencing coverage. We computed F_{ST} only on non-missing genotypes, while for PCA we imputed missing data with genotypes with the highest posterior probability.

To assess the accuracy of the per-site estimates of F_{ST}, we simulated two datasets of 10k and 1k sites for each experimental scenario to evaluate method-of-moments and ML estimates respectively, with F_{ST} varying from 0.01 to 0.4, and with P_{var} equal to 1. We verified convergence of optimization algorithms for ML estimators of F_{ST} and discarded sites where this condition was not met. We also simulated 1M sites by concatenating 100 sets of 10k simulated sites with F_{ST} values drawn from a Normal distribution $N(0.2, 0.2)$ truncated at 0.02 and 0.90, and P_{var} equal to 0.10 to assess the accuracy of multiple-sites estimates of F_{ST}. We simulated 20 individuals per population at low (2X), medium (6X), and high (20X) sequencing coverage.

In order to evaluate the performance of different methods for estimating F_{ST}, we calculated two measures of deviation from the true F_{ST} over m sites: the Root-Mean-Square Deviation (RMSD):

$$RMSD = \sqrt{\frac{1}{m}\sum_{s=1}^{m}(\hat{F}_{ST}^{(s)} - F_{ST}^{(s)})^2} \tag{23}$$

and mean bias:

$$\text{Mean bias} = \frac{1}{m}\sum_{s=1}^{m}(\hat{F}_{ST}^{(s)} - F_{ST}^{(s)}). \tag{24}$$

where $F_{ST}^{(s)}$ and $\hat{F}_{ST}^{(s)}$ is the estimated F_{ST} at site s from the case of known genotypes and sequencing data, respectively.

To evaluate the accuracy of the PCA method, we simulated 10k sites for each scenario with values of F_{ST} ranging from 0.02 to 0.4 and with P_{var} equal to 0.02, 0.1 or 1. We simulated 3 populations with 20 individuals each at 2X, 6X, and 20X sequencing coverage. We performed 10 distinct simulations for each experimental condition to assure robustness of our results. We assessed the accuracy of inferred PCA plots using Procrustes Analysis (Wang et al., 2010). Briefly, we measured the deviation of PC1 and PC2 computed from the case of known genotypes and the case of unknown genotypes using Sum-of-Squares (SS), where SS values closer to 0 indicate better fits.

Applications to real data

We analyzed a dataset of wild and domesticated species of silkworm, *Bombyx mori* (Xia et al., 2009). The data consisted of 40 samples representing 29 domesticated lineages and 11 wild lineages. Domesticated lineages are phenotypically and geographically separated into subgroups while all wild lineages are from China. Samples were sequenced at an approximate mean per-site coverage of 3X. We analyzed chromosome 2 using the original genotype likelihoods by removing sites where we had no information for at least one individual. Details on the calculation of genotype likelihoods can be found in the original paper (Xia et al., 2009). Approximately, $200,000$ sites were analyzed in total.

We computed posterior probabilities of sample allele frequencies and genotypes using ANGSD software (available at `popgen.dk/angsd`). We then performed PCA and estimated F_{ST} using the new proposed methods implemented in a set of C/C++ programs (available at `https://github.com/mfumagalli/ngsTools`). All statistical analyses were performed in the R environment (`www.r-project.org`).

Results

Quantifying population genetic differentiation from sequencing data

We performed extensive simulations to evaluate the accuracy of estimating F_{ST} using different methods and under different conditions. We first evaluated the accuracy of method-of-moments estimates of per-site F_{ST} based on called genotypes. Specifically, we assign genotypes for each individual based on the the highest genotype posterior probability ($\hat{F}_{ST.GC}$) (see Material and Methods). This approach is representative of strategies currently used for genotype calling, and it provides better genotype and SNP calling accuracies than other genotype calling strategies examined here (Tables S1-S3).

We then obtain a method-of-moments estimator of F_{ST} from NGS data without calling genotypes by using posterior probabilities of sample allele frequencies, which allows us to compute expected genetic variance components between and within populations (see Material and Methods). Here, we employ Equation 15 to estimate the 2D-SFS and use it as a prior as in Equation 16. We call this estimator $\hat{F}_{ST.Ev}$.

Results show that this new method performs substantially better than the method based on genotype calling under the experimental conditions explored in this study, especially at low sequencing coverage (Figure 1). $\hat{F}_{ST.Ev}$ tends to underestimate the true value of F_{ST} at 2X coverage, but this bias is reduced at 6X coverage (Figure 1). We observe accuracy in our estimates that are comparable to that of methods based on genotype calling for high coverage sequencing data. We obtain similar results when using the true 2D-SFS as a prior (Figure S1). We also observe that at 2X coverage, $\hat{F}_{ST.Ev}$ is more accurate for estimating F_{ST} than estimators based on computing the expected allele frequency for each population (see Supporting Text, Table S4), which overestimates F_{ST} (Figure S2).

Next, we compared the accuracy of a ML estimator of F_{ST} from called genotypes under the Balding-Nichols model, $F_{ST.ML.GC}$, to the proposed estimator based on the full likelihood under the same model while taking genotype calling uncertainty into account, $F_{ST.ML}$ (see Material and Methods). The results show that $F_{ST.ML}$ outperforms the method based on calling genotypes at 2X and 6X coverage (Figure 2). For higher sequencing coverage, both methods perform very similarly. We also observe that ML estimates of the ancestral population allele frequency are highly correlated with the true values (Figure S3).

We also test the accuracy of estimating multiple-sites F_{ST} on 10k sites from a larger set of 1M simulated positions where only 10% of the sites are variable in the population (see Material and Methods). For this particular analysis we chose the method-of-moments estimator because of its natural extension to multiple-sites estimation (Equation 4). At 2X sequencing coverage we underestimate the true F_{ST} (Figure S4). This bias diminishes at 6X and disappears at 20X. When we use the true 2D-SFS as a prior at 2X sequencing coverage, we underestimate the true F_{ST} when this value is above the whole-region average (approximately equal to 0.25), while we overestimate the true F_{ST} when this value is below the whole-region average (Figure S4). This bias is derived from using the 2D-SFS estimated from the entire region as a prior. At 6X and 20X sequencing coverage we observed unbiased estimates using the true 2D-SFS as a prior (Figure S4).

Principal Components Analysis

In traditional PCA, genotypes are called at each site for each individual. We explore an alternative approach based on the genotype posterior probabilities for each individual at each site (see Material and Methods).

At low sequencing coverage, the new method, which does not rely on SNP nor genotype calling, produces PCA plot results that are essentially identical to those that use known genotypes (Figure 3). By contrast, direct genotype calling at low sequencing coverage, generally leads to a loss in the ability to cluster individuals according to populations, which is a problem that may persists even after removing outlier individuals (Figure 3).

We replicated these findings under many different experimental conditions and for multiple independent simulations, and assessed the accuracy of PCA plots using Sum-of-Squares (SS) values from PC1 and PC2 computed from known genotypes (see Material and Methods). The new method provides better accuracy than the method based on genotype calling for all tested scenarios, even at medium sequencing coverage (Figure S5). Generally, we obtain lower SS values without normalization of the standardized allele frequencies (see Equation 18), and the new method still outperforms an approach based on called genotypes at low sequencing coverage (Figure S6). We next simulated only variable sites data at high sequencing coverage in order to produce an ideal scenario for genotype calling. As expected, procedures based on calling genotypes lead to accurate PCA results under these conditions (Figure S7).

Notably, weighting each site by its probability of being variable gives higher accuracy than simply weighing all sites equally, especially when there are only few variable sites in the sample (Figure S8). This proposed method also performs better than an approach based on computing expected genotypes from genotype posterior probabilities (Skotte et al., 2012; Gompert et al., 2012) for low coverage data (Figure S9). We also simulated one population with no genetic structure but where half of the individuals were sequenced at low coverage (2X) while the rest were sequenced at high coverage (20X). We still observe an improvement in the accuracy of the inferred PCA plots (Figure S10).

Analysis of real data

To illustrate the performance of the herein proposed methods, we applied them to a dataset of low coverage sequencing data. Specifically, we investigate the population structure of wild and domesticated silkworm samples (Xia et al., 2009). Despite using only a single chromosome of the entire silkworm dataset, we were able to detect fine scale population genetic structure. Indeed, the first component of the PCA plot generated using the new method, which takes statistical uncertainty in genotype calling into account, shows a clear separation between wild and domesticated lineages (Figure 4A). Moreover, the second component divides the different lineages of domesticated silkworms into their subgroups (Figure 4A). The first two principal components explain 6.8% and 5.2% of the total genetic variation, respectively. Of note is that we achieve a better separation among the subgroups than in the original study using whole-genome sequence data, where several subgroups appear to be intermixed (Xia et al., 2009).

We then applied naïve strategies of performing PCA based on called genotypes using the maximum genotype likelihood or genotype posterior probability at each site for each individual. Results show several outlier individuals which may be the effect of systematically misassigned heterozygous sites (Figure S11). However, when only including sites with estimated allele frequency greater or equal to two, and using genotype calling based on genotype posterior probabilities, we see a more accurate representation of the genetic structure. (Figure S11). A similar result using this allele frequency-filtered data set is obtained using the new proposed method that does not rely on genotype calling (Figure S11). Nonetheless, the new method applied to all of the data provides larger fractions of explained variance than the method based on genotype calling (Figure S12).

Finally, we estimated F_{ST} between wild and domesticated samples for 20kb non-overlapping genomic windows. We used the folded 2D-SFS due to uncertainty in assigning the ancestral and derived state of alleles. The distribution of the estimated F_{ST} values in 20kb windows has mode around 0.4 (Figure 4B), which is larger than what was found in the original study (Xia et al., 2009).

Discussion

Next-Generation Sequencing (NGS) technologies are now an essential tool for population genetic studies. However, genotyping uncertainty associated with low sequencing coverage and high sequencing error can drastically bias downstream analyses (Nielsen et al., 2011). A recent study assessed the power to detect selective events and infer demographic scenarios as a function of sequencing coverage and error (Crawford and Lazzaro, 2012). The results of the study show that weak selective events are hardly detectable, and inferences of population size changes are systematically biased for of low-coverage data (less than 10X) (Crawford and Lazzaro, 2012). Interestingly, the authors determined that population genetic differentiation was underestimated, even at medium to high sequencing coverage, suggesting that multi-population analyses are even more sensitive to inaccuracy of NGS data (Crawford and Lazzaro, 2012).

In this study, we take full advantage of a recently proposed Bayesian approach for taking sequencing data uncertainty into account (Nielsen et al., 2012; Li, 2011). This method involves computing posterior probabilities for each genotype and all possible sample allele frequencies from genotype likelihoods. Estimation of classic population genetic parameters within this new probabilistic framework has previously been suggested (Nielsen et al., 2012; Yi et al., 2010; Li, 2011) and in some cases implemented (Yi et al., 2010; Gompert and Buerkle, 2011; Kang and Marjoram, 2011). For instance, Gompert and Buerkle (2011) proposed a hierarchical Bayes model for genomic population structure. Their method accounts for uncertainty in sampling sequencing reads and measured population differentiation in terms of haplotype distances. Also, Skotte et al. (2012) and Gompert et al. (2012) used genotype expectations rather than called genotypes for the analysis of population structure. Here, we developed new methods for quantifying population genetic differentiation in terms of F_{ST} without relying on SNP or genotype calling. We simulated NGS data to assess the accuracy of these new estimators under a wide range of experimental scenarios.

Herein proposed methods for computing method-of-moments F_{ST} estimators, based on computing the posterior expected genetic variance components (see Material and Methods), offer a solution to the lack of accuracy for low coverage data and outperform other examined estimators under all tested conditions (Figure 1). While the improvement offered by the new method is greatest and most noticeable at low coverage, even at medium sequencing coverage, it results in less biased estimates of F_{ST} (Figure 1). Similarly, ML estimation of F_{ST} that accounts for uncertainty in genotype calls outperforms a method based on genotype calling at low and medium coverage (Figure 2). These findings suggest that the framework presented in this study can be easily extended to other F_{ST} estimators. Overall, these results highlight the importance of taking statistical uncertainty into account when computing population genetic differentiation from NGS data. The great improvement in accuracy for low coverage data can be explained by the fact that we do not call SNPs or genotypes. We can thus avoid introducing errors during these processes which can be particularly problematic for downstream analyses.

Errors introduced by calling SNPs and genotypes for low coverage and quality data can be even more evident when investigating population structure with Principal Components Analysis. Simple genotype calling provides very little ability to accurately identify structure using PCA for low coverage data. However, the new method based on genotype posterior probabilities provides PCA plots that are almost identical to cases where true genotypes are known (Figure 3). Accuracy in identifying population structure can be recovered when calling genotypes by removing outlier individuals, low quality sites, and low frequency variants, but at the price of losing potential important information. Skoglund and Jakobsson (2011) investigated population structure by randomly sampled one read from each individual at each position. In this way they could compare modern, high quality data with the low-pass ancient data. A disadvantage of this method is the loss of information associated with using only a single read from each individual, especially in the presence of sequencing errors.

We applied methods proposed in this study to a dataset comprising 40 silkworm samples sequenced at low-coverage (Xia et al., 2009). We only used a single chromosome of the original dataset and we did not apply any criteria for SNP calling. Despite this, we were able to obtain a fine-scale map of population genetic structure, clearly separating wild and domesticated lineages of silkworm samples (Figure 4A). The first principal component separates domesticated and wild varieties, while the second component accurately divides the domesticated lineage into subgroups. Genotype calling from genotype posterior probabilities can provide an overall similar representation of the genetic structure when using a conservative initial filtering of data.

Genotype calling using stringent data filtering and a conservative approach for SNP calling and rare variants removal may be sufficient to give an overall picture of the genetic population structure, for example a reasonably representative PCA. Other analyses, such as estimation of F_{ST}, that rely on accurate estimates of allele frequencies may be more difficult to rescue by conservative filtering because a fixed cut-off for SNP calling cannot provide unbiased estimates of allele frequencies (e.g. Johnson and Slatkin, 2008). Furthermore, the accuracy of genotype calling can be improved for human data by using imputation or haplotype based genotype calling methods (e.g. Zhi et al., 2012), although such approaches are not as easily applicable to most other species. The poor performance of PCA after calling genotypes may largely be a result of inaccuracies in SNP calling rather than a consequence of erroneous genotype calls at variable sites. However, when simulating sequences with a larger proportion of polymorphic sites the new method still outperforms traditional methods, even in case of an uneven sequencing coverage among individuals (Figure S10). While more sophisticated approaches have been developed to perform accurate SNP calling (e.g. Kim et al., 2011), calling polymorphic sites using all individuals may result in ascertainment biases which can influence estimates of population structure and divergence (Albrechtsen et al., 2010). Additionally, a stringent SNP calling strategy implies that a large amount of data is discarded from the analyses, potentially leading to loss of important features of the data. For example, low-frequency variants, which are more likely to be removed in a conservative SNP calling strategy, can effectively distinguish closely related populations. Moreover, highly differentiated SNPs among populations, which may be related to genetic adaptation, might be lost in some analyses.

Like any other method for SNP calling and allele frequency estimation, the approach herein discussed is sensitive to the underlying base calling algorithm and to the accuracy of quality scores. By improving accuracy and quality scores, current and future base callers can both reduce sequencing costs and increase accuracy of all downstream analyses of genetic variation. Furthermore, data filtering is a complex procedure when sequencing quality is low (e.g. Minoche et al., 2011). Many other protocols, other than the ones used in this manuscript, can be adopted in order to minimize the genotypes assignment bias.

We implemented the new proposed methods for estimating F_{ST} and perform PCA from NGS data in a fast, portable, and memory-efficient set of C/C++ programs, and distributed on a public repository for shared development. These programs are directly integrated with ANGSD (`popgen.dk/angsd`), a software for the analysis of NGS data, and easily integrable with other common software such as SAMtools (Li et al., 2009) or GATK (McKenna et al., 2010). The computational cost associated with the new methods are slightly higher than standard approaches (Tables S5-S6). However, the increased computational burden is mostly associated with the computation of sample allele frequency posterior probabilities, which can be used for additional analyses. Notably, the computational cost should not be prohibitive for any existing data sets.

As NGS technologies become more ubiquitous and affordable, the frequency of large-scale population genetic and quantitative studies will certainly increase. The methods presented in this paper provide tools for investigating genetic variation for multiple populations at large scales directly from high-throughput sequencing data.

Acknowledgments

We are grateful to Michael DeGiorgio and Gaston Sanchez at the University of California, Berkeley, USA, for helpful discussions. We would like to thank Shiping Liu and Zengli Yan at Beijing Genomics Institute, Shenzhen, China, for testing previous versions of the programs, and three anonymous reviewers for insightful comments on the manuscript. MF is supported by EMBO Long-Term Post-doctoral Fellowship (ALTF 229-2011). TL is supported by NIH Genomics Training Grant (Grant T32HG000047-13). EHS is supported by NSF DBI-0906065 and NIH grant 3R01HG03229-08S2. RN is supported by NIH grant 3R01HG03229-07.

Appendix

Let G_i and X_i be random variables representing the genotype and read data, respectively, from individual i. Likewise, let g_i be a realization of G_i, and let x_i be a realization of X_i, $i = 1, 2$.We then wish to prove that

$$E_{X_1,X_2}[\sum_{g_1}\sum_{g_2}(g_1 - E[G_1])(g_2 - E[G_2])p(g_1|x_1)p(g_2|x_2)] = 0 \tag{25}$$

where the sums, here and in the following, are over all supported values of the variable under the summation sign. We use a simplified notation so that expectation operators implicitly are taken with respect to the random variable(s) inside the argument of the expectation operator, except when otherwise indicated by the use of subscripts. Also, we use the shorthand notation $p(g_i)$ for the probability of the random variable G_i taken on the value g_i.

First notice that, for unrelated individuals, G_1 and G_2 are independent and that X_1 and X_2 are independent assuming a fixed known allele frequency and assuming random mating. Next also notice that

$$E_{X_i}[g_i p(g_i|x_i)] = g_i p(g_i) \tag{26}$$

and

$$E_{X_i}[p(g_i|x_i)] = p(g_i). \tag{27}$$

Then

$$\begin{aligned}
& E_{X_1,X_2}\left[\sum_{g_1}\sum_{g_2}(g_1 - E[G_1])(g_2 - E[G_2])p(g_1|x_1)p(g_2|x_2)\right] \\
= & \sum_{g_1}\sum_{g_2}(E_{X_1,X_2}\left[g_1 g_2 p(g_1|x_1)p(g_2|x_2)\right] - E_{X_1,X_2}\left[g_1 E[G_2]p(g_1|x_1)p(g_2|x_2)\right] \\
& + E_{X_1,X_2}\left[g_2 E[G_1]p(g_1|x_1)p(g_2|x_2)\right] + E_{X_1,X_2}\left[g_1 g_2 p(g_1|x_1)p(g_2|x_2)\right]) \\
= & \sum_{g_1}\sum_{g_2}(E_{X_1}[g_1 p(g_1|x_1)]E_{X_2}[g_2 p(g_2|x_2)] \\
& - E_{X_1}[g_1 p(g_1|x_1)]E_{X_2}[E[G_2]p(g_2|x_2)] \\
& - E_{X_2}[g_2 p(g_2|x_2)]E_{X_1}[E[G_1]p(g_1|x_1)] \\
& + E_{X_1}[g_1 p(g_1|x_1)]E_{X_2}[g_2 p(g_2|x_2)]) \\
= & \sum_{g_1}\sum_{g_2}(g_1 g_2 p(g_1)p(g_2) - g_1 p(g_1)E[G_2]p(g_2) \\
& - g_2 p(g_2)E[G_1]p(g_1) + E[G_1]E[G_2]p(g_1)p(g_2)) \\
= & E[G_1]E[G_2] - E[G_1]E[G_2] - E[G_2]E[G_1] + E[G_1]E[G_2] \\
= & 0.
\end{aligned}$$

The interchange of summations in the first step is justified because all sums are finite. The second equality is true because of the independence assumption. The third equality is verified by substitution of the expressions in 26 and 27. The fourth equality follows from the independence assumption and the definition of expectation.

References

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 28 2010.

1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 1 2012.

A. Albrechtsen, F. C. Nielsen, and R. Nielsen. Ascertainment biases in snp chips affect measures of population divergence. *Molecular biology and evolution*, 27(11):2534–2547, Nov 2010.

A. Auton, A. Fledel-Alon, S. Pfeifer, O. Venn, L. Segurel, T. Street, E. M. Leffler, R. Bowden, I. Aneas, J. Broxholme, P. Humburg, Z. Iqbal, G. Lunter, J. Maller, R. D. Hernandez,

C. Melton, A. Venkat, M. A. Nobrega, R. Bontrop, S. Myers, P. Donnelly, M. Przeworski, and G. McVean. A fine-scale chimpanzee genetic map from population sequencing. *Science (New York, N.Y.)*, 336(6078):193–198, Apr 13 2012.

D. J. Balding. Likelihood-based inference for genetic correlation coefficients. *Theoretical population biology*, 63(3):221–230, May 2003.

D. J. Balding and R. A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12, 1995.

M. A. Beaumont and D. J. Balding. Identifying adaptive genetic divergence among populations from genome scans. *Molecular ecology*, 13(4):969–980, Apr 2004.

J. E. Crawford and B. P. Lazzaro. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Frontiers in genetics*, 3:66, 2012.

W. Ewens. *Mathematical Population Genetics: Theoretical Introduction*. Springer, 2004.

R. Fletcher. *Practical methods of optimization; (2nd ed.)*. Wiley-Interscience New York, NY, USA, 1987.

M. Foll and O. Gaggiotti. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics*, 180(2):977–993, Oct 2008.

Z. Gompert and C. A. Buerkle. A hierarchical bayesian model for next-generation population genomics. *Genetics*, 187(3):903–917, Mar 2011.

Z. Gompert, L. K. Lucas, C. C. Nice, J. A. Fordyce, M. L. Forister, and C. A. Buerkle. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution; international journal of organic evolution*, 66(7):2167–2181, Jul 2012.

I. Hellmann, Y. Mang, Z. Gu, P. Li, F. M. de la Vega, A. G. Clark, and R. Nielsen. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome research*, 18(7):1020–1029, Jul 2008.

K. E. Holsinger and B. S. Weir. Genetics in geographically structured populations: defining, estimating and interpreting f(st). *Nature reviews.Genetics*, 10(9):639–650, Sep 2009.

K. E. Holsinger, P. O. Lewis, and D. K. Dey. A bayesian approach to inferring population structure from dominant markers. *Molecular ecology*, 11(7):1157–1164, Jul 2002. JID: 9214478; 0 (Genetic Markers); ppublish.

X. Huang, N. Kurata, X. Wei, Z. X. Wang, A. Wang, Q. Zhao, Y. Zhao, K. Liu, H. Lu, W. Li, Y. Guo, Y. Lu, C. Zhou, D. Fan, Q. Weng, C. Zhu, T. Huang, L. Zhang, Y. Wang, L. Feng, H. Furuumi, T. Kubo, T. Miyabayashi, X. Yuan, Q. Xu, G. Dong, Q. Zhan, C. Li, A. Fujiyama, A. Toyoda, T. Lu, Q. Feng, Q. Qian, J. Li, and B. Han. A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 490(7421):497–501, Oct 25 2012.

R. R. Hudson, M. Slatkin, and W. P. Maddison. Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2):583–589, Oct 1992.

P. L. Johnson and M. Slatkin. Accounting for bias from sequencing error in population genetic estimates. *Molecular biology and evolution*, 25(1):199–206, Jan 2008.

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

P. D. Keightley and D. L. Halligan. 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, Aug 2011.

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

S. Y. Kim, K. E. Lohmueller, A. Albrechtsen, Y. Li, T. Korneliussen, G. Tian, N. Grarup, T. Jiang, G. Andersen, D. Witte, T. Jorgensen, T. Hansen, O. Pedersen, J. Wang, and R. Nielsen. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics*, 12:231, Jun 11 2011.

H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21):2987–2993, Nov 1 2011.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, Aug 15 2009.

Y. Li, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, H. Cao, T. Korneliussen, N. Grarup, Y. Guo, I. Hellman, X. Jin, Q. Li, J. Liu, X. Liu, T. Sparso, M. Tang, H. Wu, R. Wu, C. Yu, H. Zheng, A. Astrup, L. Bolund, J. Holmkvist, T. Jorgensen, K. Kristiansen, O. Schmitz, T. W. Schwartz, X. Zhang, R. Li, H. Yang, J. Wang, T. Hansen, O. Pedersen, R. Nielsen, and J. Wang. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature genetics*, 42(11):969–972, Nov 2010.

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

J.L. Marchini and L. Cardon. Discussion on the meeting on statistical modelling and analysis of genetic data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):737–775, Oct 2002.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, Sep 2010.

M. L. Metzker. Sequencing technologies - the next generation. *Nature reviews.Genetics*, 11 (1):31–46, Jan 2010.

A. E. Minoche, J. C. Dohm, and H. Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biology*, 12(11):R112, Nov 2011.

G. Nicholson, A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson, and P. Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64: 695–715, 2002.

R. Nielsen. Molecular signatures of natural selection. *Annual Review of Genetics*, 39:197–218, 2005.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and snp calling from next-generation sequencing data. *Nature reviews.Genetics*, 12(6):443–451, Jun 2011.

R. Nielsen, T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang. Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS one*, 7(7):e37558, 2012.

M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen. Bayesian principal component analysis. *Journal of chemometrics*, 16:576–595, 2002.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, Dec 2006.

W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes: The Art of Scientific Computing, third edition.* Cambridge University Press, 2007.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, Aug 2006.

J. K. Pritchard and P. Donnelly. Case-control studies of association in structured or admixed populations. *Theoretical population biology*, 60(3):227–237, Nov 2001.

J. Reynolds, B. S. Weir, and C. C. Cockerham. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 105(3):767–779, Nov 1983.

S. H. Rice. A stochastic version of the price equation reveals the interplay of deterministic and stochastic processes in evolution. *BMC evolutionary biology*, 8:262, Sep 25 2008.

S. H. Rice and A. Papadopoulos. Evolution with stochastic fitness and stochastic migration. *PloS one*, 4(10):e7130, Oct 9 2009.

A. Riebler, L. Held, and W. Stephan. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, 178(3):1817–1829, Mar 2008.

C. J. Rubin, M. C. Zody, J. Eriksson, J. R. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallbook, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, 464(7288):587–591, Mar 25 2010.

P. Skoglund and M. Jakobsson. Archaic human ancestry in east asia. *Proceedings of the National Academy of Sciences of the United States of America*, 108(45):18301–18306, Nov 8 2011.

L. Skotte, T. S. Korneliussen, and A. Albrechtsen. Association testing for next-generation sequencing data using score statistics. *Genetic epidemiology*, 36(5):430–437, Jul 2012.

C. Wang, Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, A. B. Singleton, and N. A. Rosenberg. Comparing spatial maps of human population-genetic variation using procrustes analysis. *Statistical applications in genetics and molecular biology*, 9(1):Article 13, 2010.

B. S. Weir. *Genetic Data Analysis II.* Sinauer, 1996.

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

B. S. Weir and W. G. Hill. Estimating f-statistics. *Annual Review of Genetics*, 36:721–750, 2002.

P. D. Wentzell, D. Andrews, D. C. Hamilton, F. Faber, and B. R. Kowalski. Maximum likelihood principal component analysis. *Journal of chemometrics*, 11:339–366, 1997.

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

Q. Xia, Y. Guo, Z. Zhang, D. Li, Z. Xuan, Z. Li, F. Dai, Y. Li, D. Cheng, R. Li, T. Cheng, T. Jiang, C. Becquet, X. Xu, C. Liu, X. Zha, W. Fan, Y. Lin, Y. Shen, L. Jiang, J. Jensen, I. Hellmann, S. Tang, P. Zhao, H. Xu, C. Yu, G. Zhang, J. Li, J. Cao, S. Liu, N. He, Y. Zhou, H. Liu, J. Zhao, C. Ye, Z. Du, G. Pan, A. Zhao, H. Shao, W. Zeng, P. Wu, C. Li, M. Pan, J. Li, X. Yin, D. Li, J. Wang, H. Zheng, W. Wang, X. Zhang, S. Li, H. Yang, C. Lu, R. Nielsen, Z. Zhou, J. Wang, Z. Xiang, and J. Wang. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (bombyx). *Science (New York, N.Y.)*, 326(5951):433–436, Oct 16 2009.

X. Xu, X. Liu, S. Ge, J. D. Jensen, F. Hu, X. Li, Y. Dong, R. N. Gutenkunst, L. Fang, L. Huang, J. Li, W. He, G. Zhang, X. Zheng, F. Zhang, Y. Li, C. Yu, K. Kristiansen,

X. Zhang, J. Wang, M. Wright, S. McCouch, R. Nielsen, J. Wang, and W. Wang. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature biotechnology*, 30(1):105–111, Dec 11 2011.

X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, 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, Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, and J. Wang. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*, 329(5987):75–78, Jul 2 2010.

D. Zhi, J. Wu, N. Liu, and K. Zhang. Genotype calling from next-generation sequencing data using haplotype information of reads. *Bioinformatics*, 28(7):938–946, Apr 2012.

Tables

Notation	Description
$p_{(i,s)}$, p_s	population allele frequency in population i and pooled, respectively, at site s
$p_{anc,s}$	ancestral population allele frequency at site s
$\hat{p}_{(i,s)}$, $\hat{p}_s$	estimated population allele frequency from allele counts at population i and pooled, respectively, at site s
n_i, n	number of sampled individuals at population i, and pooled, respectively
m	number of sites
r_s	number of sequencing reads at site s
$v_{z,s}$	base at sequencing read z at site s
$L_{(z,v,s)}$	likelihood of base v at read z and site s
$G_{(w,s)}$	genotype at site s for individual w; $G \in \{0, 1, 2\}$
$X_{(w,s)}$	data (sequencing reads) at site s for individual w
$Y_{(i,s)}$	data (sequencing reads) at site s for population i
$h^{(k)}_{(i,s)}$	marginal likelihood of k non-reference alleles for population i at site s
$\pi^{(k)}_{(i,s)}$, $\pi^{(k)}_s$	posterior probability of k non-reference alleles for population i and pooled, respectively, at site s
$\pi^{(k,z)}_{(i,j,s)}$	joint posterior probability of k and z non-reference alleles for population i and j, respectively, at site s
$a^{(k,z)}_{(i,j)}$, $b^{(k,z)}_{(i,j)}$, $c^{(k,z)}_{(i,j)}$	genetic variance between (a) and within (b) populations and total (c) assuming k and z non-reference alleles at population i and j, respectively
$S^{(k,z)}_{(i,j)}$	joint allele proportions for k and z non-reference alleles at population i and j, respectively
$C_{(w,y)}$	normalized matrix for PCA for individual w and y
s	index for sites
k	index for samples (allele frequencies)
w, y	indexes for individuals
z	index for sequencing reads
$P(\cdot)$	probability function
$B(\cdot)$	Beta function

Table 1: **Nomenclature used in the manuscript.**

Figures

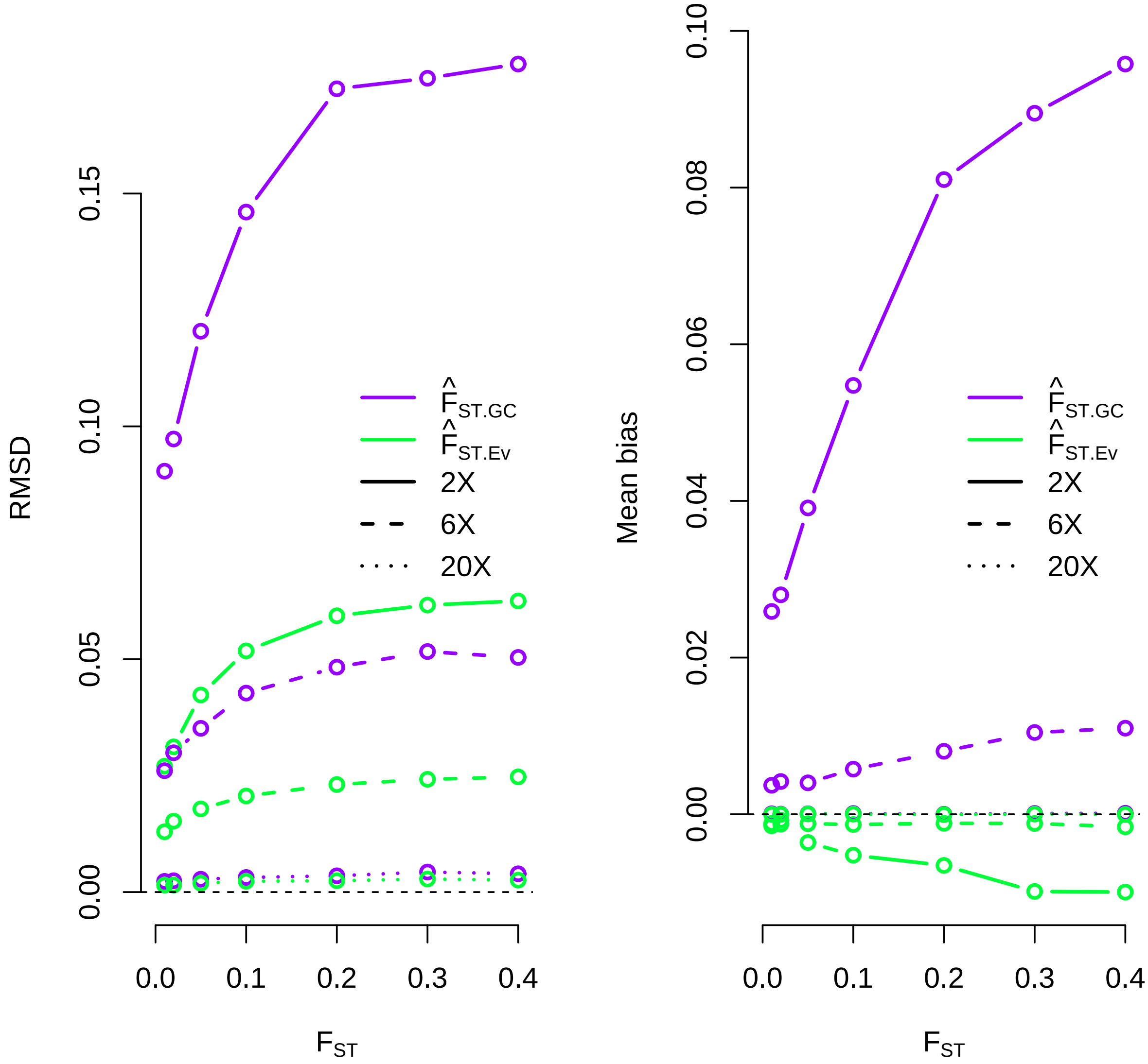

Figure 1: RMSD (left panel) and mean bias (right panel) for method-of-moments estimates of F_{ST} under different sequencing coverage (2X, 6X and 20X). We compared the accuracy of the new method which does not rely on genotype calling ($\hat{F}_{ST.Ev}$) and a method based on allele frequencies estimated from called genotypes ($\hat{F}_{ST.GC}$) (see Material and Methods). We simulated 20 individuals for each population and $10,000$ sites for each scenario.

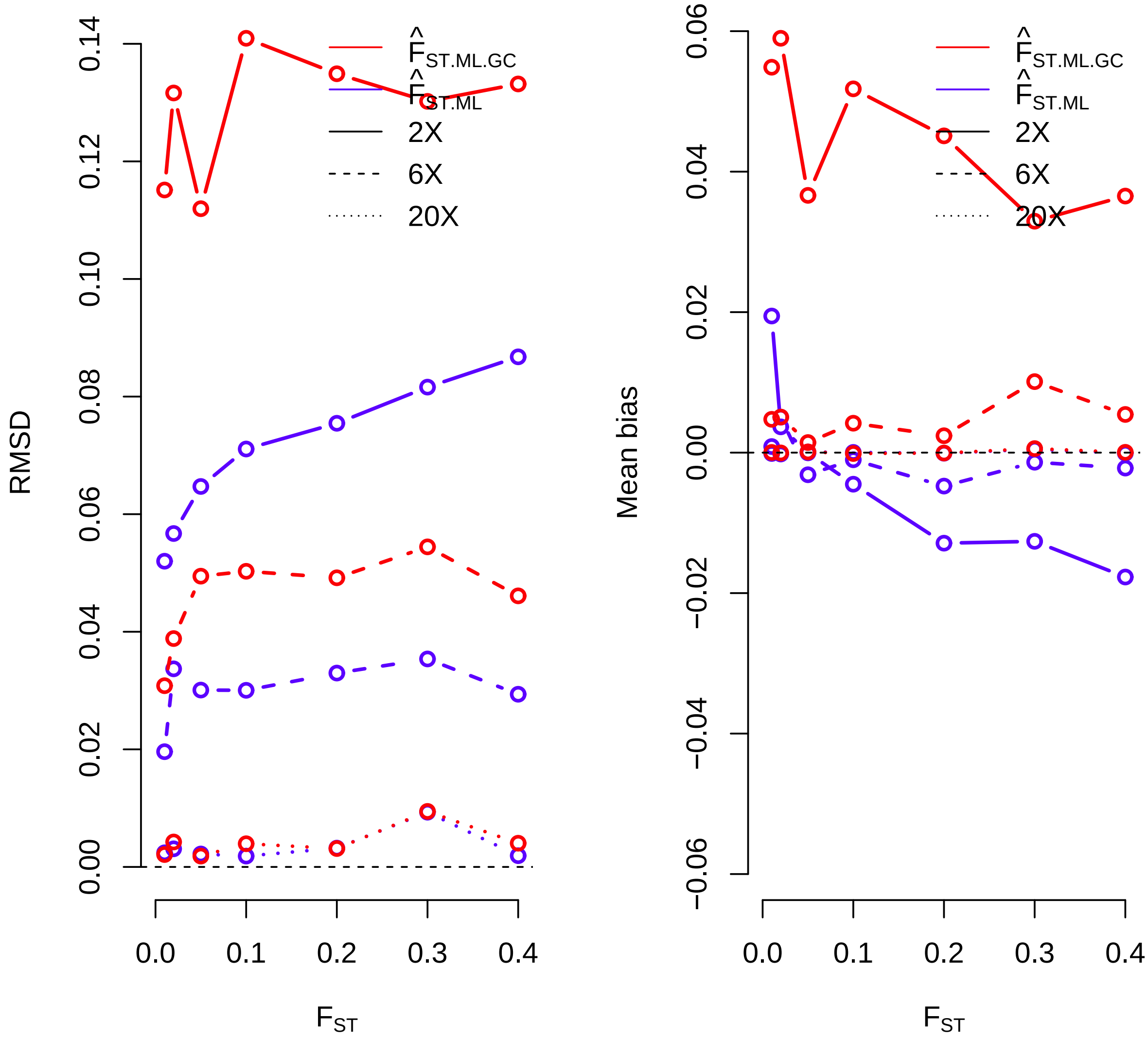

Figure 2: RMSD (left panel) and mean bias (right panel) for Maximum Likelihood estimates of F_{ST} under different sequencing coverage (2X, 6X and 20X). We compared accuracy of the new method which does not rely on genotype calling ($\hat{F}_{ST.ML}$) and the standard method applied to called genotypes ($\hat{F}_{ST.ML.GC}$) (see Material and Methods). We simulated 20 individuals for each population and $1,000$ sites for each scenario.

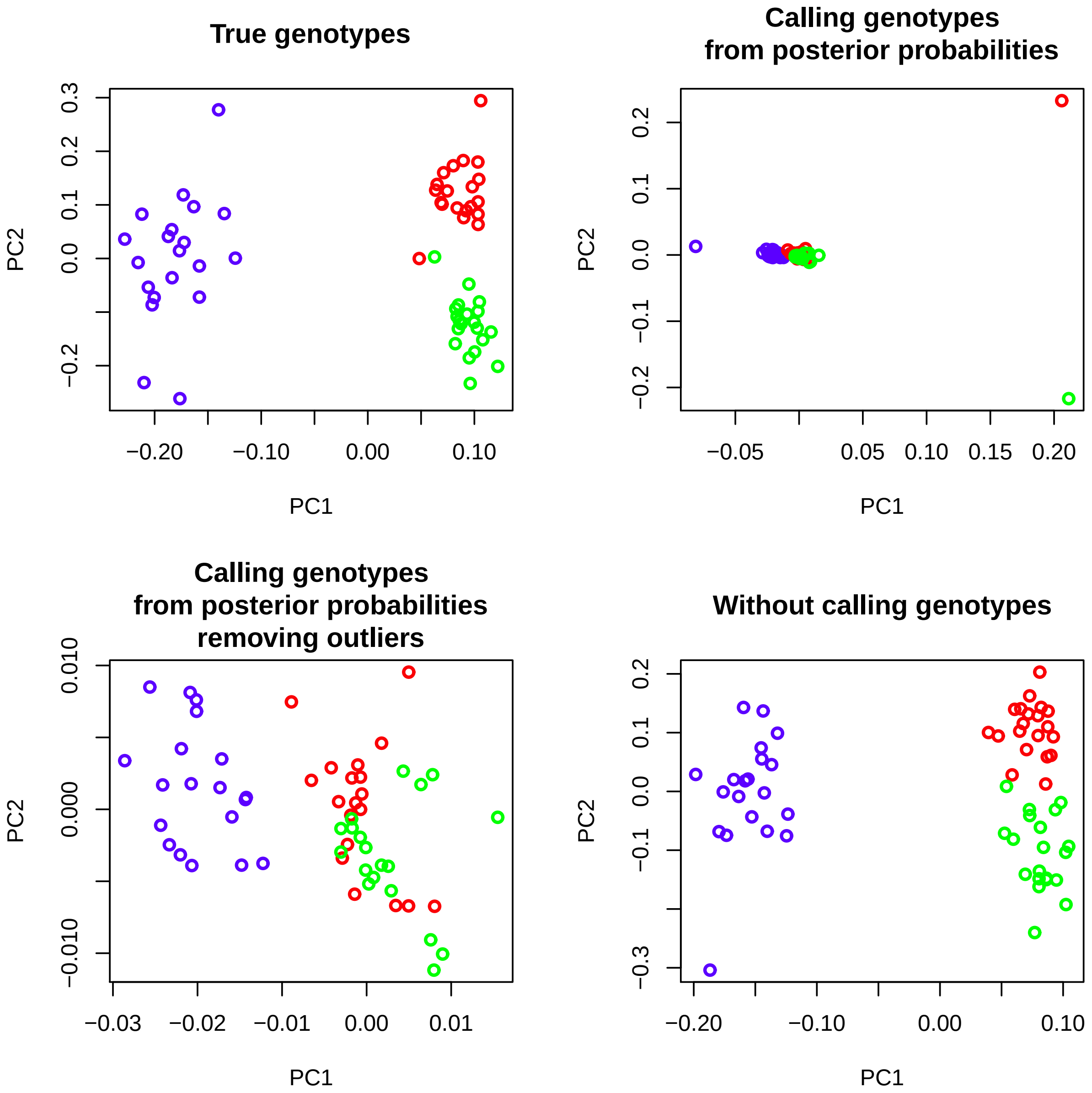

Figure 3: PCA plots from known genotypes, called genotypes using genotype posterior probabilities with or without outlier individuals, and using the new method without calling genotypes (see Material and Methods). We simulated 3 populations of 20 individuals each at 2X sequencing coverage. Colors are coded according to each simulated population. Blue and green/red populations are differentiated by an F_{ST} of 0.4 while green and red populations are differentiated by an F_{ST} of 0.15. We simulated $10,000$ sites with 10% of sites being variable in the population.

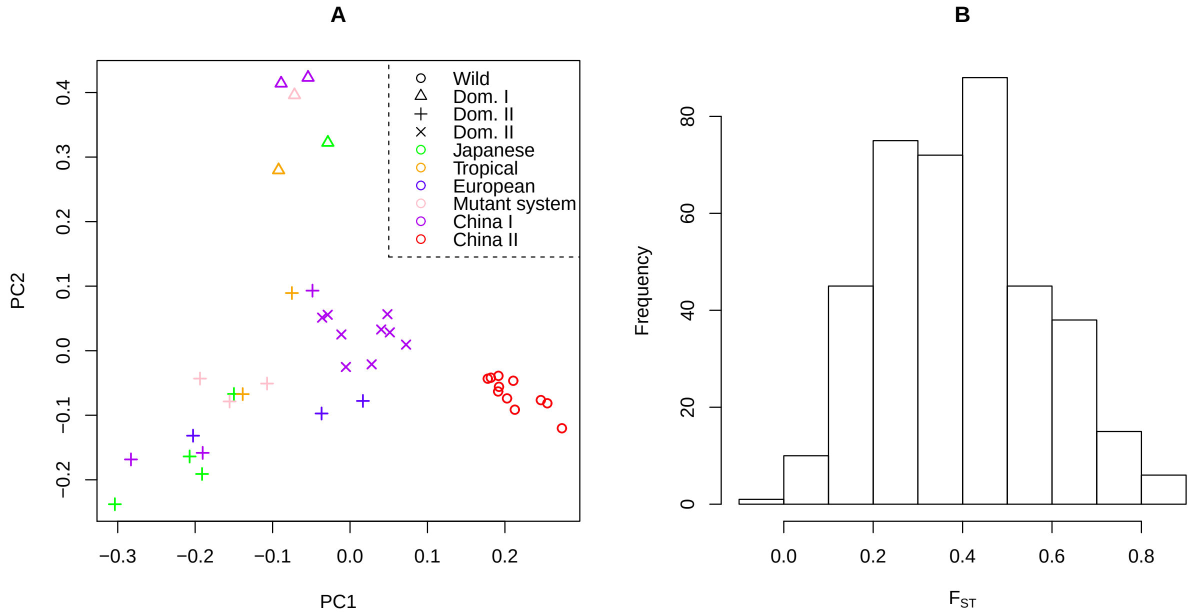

Figure 4: (A) PCA plot for wild and domesticated *Bombyx mori* samples using the method proposed in this study. (B) Distribution of F_{ST} between wild and domesticated *Bombyx mori* samples over 20kb genomic windows.