OPEN ACCESS Freely available online

PLOS ONE

Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences

Matteo Fumagalli*

Department of Integrative Biology, University of California, Berkeley, California, United States of America

Abstract

Next-Generation Sequencing (NGS) technologies have dramatically revolutionised research in many fields of genetics. The ability to sequence many individuals from one or multiple populations at a genomic scale has greatly enhanced population genetics studies and made it a data-driven discipline. Recently, researchers have proposed statistical modelling to address genotyping uncertainty associated with NGS data. However, an ongoing debate is whether it is more beneficial to increase the number of sequenced individuals or the per-sample sequencing depth for estimating genetic variation. Through extensive simulations, I assessed the accuracy of estimating nucleotide diversity, detecting polymorphic sites, and predicting population structure under different experimental scenarios. Results show that the greatest accuracy for estimating population genetics parameters is achieved by employing a large sample size, despite single individuals being sequenced at low depth. Under some circumstances, the minimum sequencing depth for obtaining accurate estimates of allele frequencies and to identify polymorphic sites is $2X$, where both alleles are more likely to have been sequenced. On the other hand, inferences of population structure are more accurate at very large sample sizes, even with extremely low sequencing depth. This all points to the conclusion that under various experimental scenarios, in cost-limited population genetics studies, large sample sizes at low sequencing depth are desirable to achieve high accuracy. These findings will help researchers design their experimental set-ups and guide further investigation on the effect of protocol design for genetic research.

Citation: Fumagalli M (2013) Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences. PLoS ONE 8(11): e79667. doi:10.1371/journal.pone.0079667

Editor: Ludovic Orlando, Natural History Museum of Denmark, University of Copenhagen, Denmark

Received May 30, 2013; **Accepted** September 23, 2013; **Published** November 18, 2013

Funding: MF is supported by European Molecular Biology Organistation Long-Term Post-doctoral Fellowship (ALTF 229-2011). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: matteo.fumagalli@berkeley.edu

Introduction

One primary aim of population genetics studies is understanding the relative role of neutral and selective forces in shaping the overall genetic diversity of populations. This is often nowadays achieved by investigating the amount and patterns of genetic variation across multiple samples at a large genomic scale. However, until recently, studies relied on the analysis of sequencing data for short genomic regions or for a limited number of candidate genes, or on the analysis of genotypes from sparse Single Nucleotide Polymorphism (SNP) data. While the former approach produces accurate inferences, it targets a small fraction of the genome, and the latter provides insights at the genome-wide level but can be prone to considerable ascertainment bias, which has been shown to inflate certain results [1]. The main obstacle precluding more extensive analyses relates to high experimental costs.

In the last few years, new high-throughput DNA sequencing technologies have allowed researchers to generate large amounts of genetic data. Such Next-Generation-Sequencing (NGS) technologies are now a common tool in population genetics [2], medical genetics [3] and other genetic disciplines [4]. While NGS technologies may differ in their protocols, the data produced by them all have similar general characteristics [5]: short fragments of sequenced DNA known as "reads" are mapped to a reference genome or *de novo* aligned. The data on which all downstream analyses are performed typically consists of a collection of mapped reads covering a particular genomic position, with associated base and mapping quality scores. Each site in the alignment can be covered by a variable number of reads (a feature called "sequencing depth"). Individual genotypes are then inferred from the allelic state of the reads covering the site of interest (a procedure called "genotype calling"), while "SNP calling" refers to the process of identifying which sites are polymorphic in the sample, that is, have more than 1 base type at the site.

Sequencing depth is an important characteristic of the data. Genotypes called for sites with higher depth are likely to be more accurate, while lower sequencing depth leads to a non-negligible amount of genotyping uncertainty [6]. Since SNP calling proceeds from genotype calling, sequencing depth influences the detection of variable sites. Factors such as sequencing and mapping errors add to the uncertainty in genotype and SNP calling from NGS data.

Recently proposed methods that employ statistical models accommodate this uncertainty by using genotype likelihoods and have been successfully applied to empirical datasets (e.g. [7]). Such methods include those used for estimating allele frequencies at a single site [8–10] or jointly across multiple sites [9,11,12], mutation rates [13], and several population genetics summary statistics and parameters [11,12,14–17].

NGS technologies are a powerful tool for investigating the evolutionary forces that shape genomes. Many summary statistics used for analysing demography, natural selection, and population structure, are derived from estimates of nucleotide variation across multiple individuals [18]. The number of segregating sites and the allele frequencies at these sites are among the most important features of the data from an evolutionary perspective, and are the basis of commonly used neutrality tests [19–22].

Genetic structure is another extremely important feature of populations that can be discerned from population genetics data. Realising population structure provides insights into demographic history [23], and has practical use in clinical association studies [24]. Principal Component Analysis (PCA) is a long-standing statistical tool for examining genetic structure among individuals because it reduces highly-dimensional genetic data into a map of uncorrelated components based on the covariance among genotypes [25].

Population genetics inferences will become more accurate with greater sample sizes, that is, with more individuals representing a particular population. However, at a fixed research budget, sequencing more samples will lower the per-sample sequencing depth, and, as a consequence, increase the genotype uncertainty. Similarly, higher sequencing coverage will decrease genotyping uncertainty, but will also restrict the analysis to a smaller sample of individuals, which may be a poor representation of the genomic variation of the entire population. Recent whole-genome sequencing projects have adopted both the former [26–29] and the latter strategy [30].

It is therefore appealing to investigate the relationship between the accuracy in estimating within- and between-populations genetic variation and the sequencing experimental design. The sequencing strategy can easily be modelled in terms of the number of sequenced samples and the per-sample sequencing depth. Despite the extensive use of NGS data in population genetics, the effect on the accuracy of estimates of genetic variation by different sequencing strategies has yet to be thoroughly quantified.

Through simulation of sequencing data and by using state-of-the-art statistical methods for estimating genetic variation from NGS data, I quantified the accuracy of estimating the number of segregating sites, nucleotide diversity, allele frequencies, and population structure under a wide range of sequencing scenarios. These results will help researchers optimise their sequencing experiments.

Results and Discussion

Estimating Nucleotide Diversity

Extensive simulations were performed to evaluate the accuracy of estimating nucleotide diversity under various sequencing conditions and fixed experimental budget. The cost is assumed to be proportional to the total sequencing depth, which is a function of the number of individuals and target size. Therefore, experiments with equal cost will have equal total sequencing depth. Although this may be not strictly true, this assumption is a reasonable generalization given current NGS technologies.

A total of $1M$, and $100k$, sites of DNA sequencing data were simulated at an average per-sample sequencing depth of $1X$, $2X$, $10X$, and $50X$. Corresponding sample sizes were 1,000, 500, 100, and 20 diploid individuals, so that the product of the sample size and sequencing coverage was the same across scenarios. The standardised bias for estimates of the number of segregating sites (S) and the expected heterozygosity (H), between the case of known genotypes for all 1,000 individuals and the case of unknown genotypes for all or a fraction of individuals, was calculated. Sequence data was divided into 100 independent windows and the bias in the estimates for the population genetics statistics was computed for each region separately (see Methods).

The highest accuracy for estimating the number of segregating sites was achieved at a larger sample size despite the lower sequencing depth (Figure 1). In all scenarios, the true number of segregating sites in the population was underestimated, but this error approaches 0 in the $1X$ coverage condition. The error rapidly increases at higher sequencing depth and lower sample size. At $50X$ coverage for 20 individuals, the number of segregating sites is underestimated by up to 50%.

Secondly, estimates for the expected heterozygosity from simulated sequencing data were compared to estimates of heterozygosity with known genotypes. Heterozygosity is a function of allele frequency (see Methods). Heterozygosity is severely underestimated at high sequencing depth and small sample size, while an approximately unbiased estimate is achieved at 2X coverage for 500 sequenced individuals (Figure 1). Similar results are observed when simulating a larger number of sites with lower variability (Figure S1) or lower sequencing error rate (Figure S2).

When sequencing depth is low, under-estimating S and H can be attributed to the smaller probability of sequencing the alternate allele from heterozygotes. On the other hand, when sample sizes are small, S and H are under-estimated due to heterozygotes not being sampled. The results clearly show that, despite lower sequencing depths, larger sample sizes produce more accurate estimates of population genetics variation. Furthermore, increasing sample size affords greater accuracy for detecting nucleotide diversity outliers, with a sequencing depth of $2X$ for 500 individuals giving the highest correlation between true and estimated values (Table S1).

Under a simulated population expansion model (e.g. like in humans [31]), estimates of nucleotide diversity at high sequencing depth and small sample size were even more biased than under the constant population size model (Figure S3). Under population expansion, the site frequency spectrum is skewed towards low frequency variants, which are not captured well when sequencing only a small number of individuals. This effect increases the error when estimating nucleotide diversity.

The number of segregating sites and nucleotide diversity were also estimated under conditions in which genotype proportions deviated from Hardy-Weinberg Equilibrium (HWE) due to inbreeding. Specifically, an individual inbreeding coefficient of 0.3 was used for the simulations (see Methods). This inbreeding scenario is representative of highly structured populations, self-pollinating plants, and domesticated species. The highest accuracy in estimating the number of segregating sites and nucleotide diversity was achieved when employing many samples at low sequencing depth (Figure S4). The general decrease in accuracy when estimating average heterozygosity is caused by violation of the HWE assumption upon which the method used to estimate heterozygosity relies [11]. Further studies to generalise models for estimating allele frequencies from sequencing data when HWE does not hold are strongly encouraged [32].

Sequencing a large number of samples at the trade-off of lower individual coverage represents the optimal design for accurately inferring population nucleotide diversity. Under some scenarios, the highest accuracy for estimating the expected heterozygosity, which is a function of the sample allele frequency, is achieved at $2X$ sequencing depth, where both alleles are more likely to have been sequenced, versus 1X coverage. These findings are robust to different assumptions of population demography and mating system.

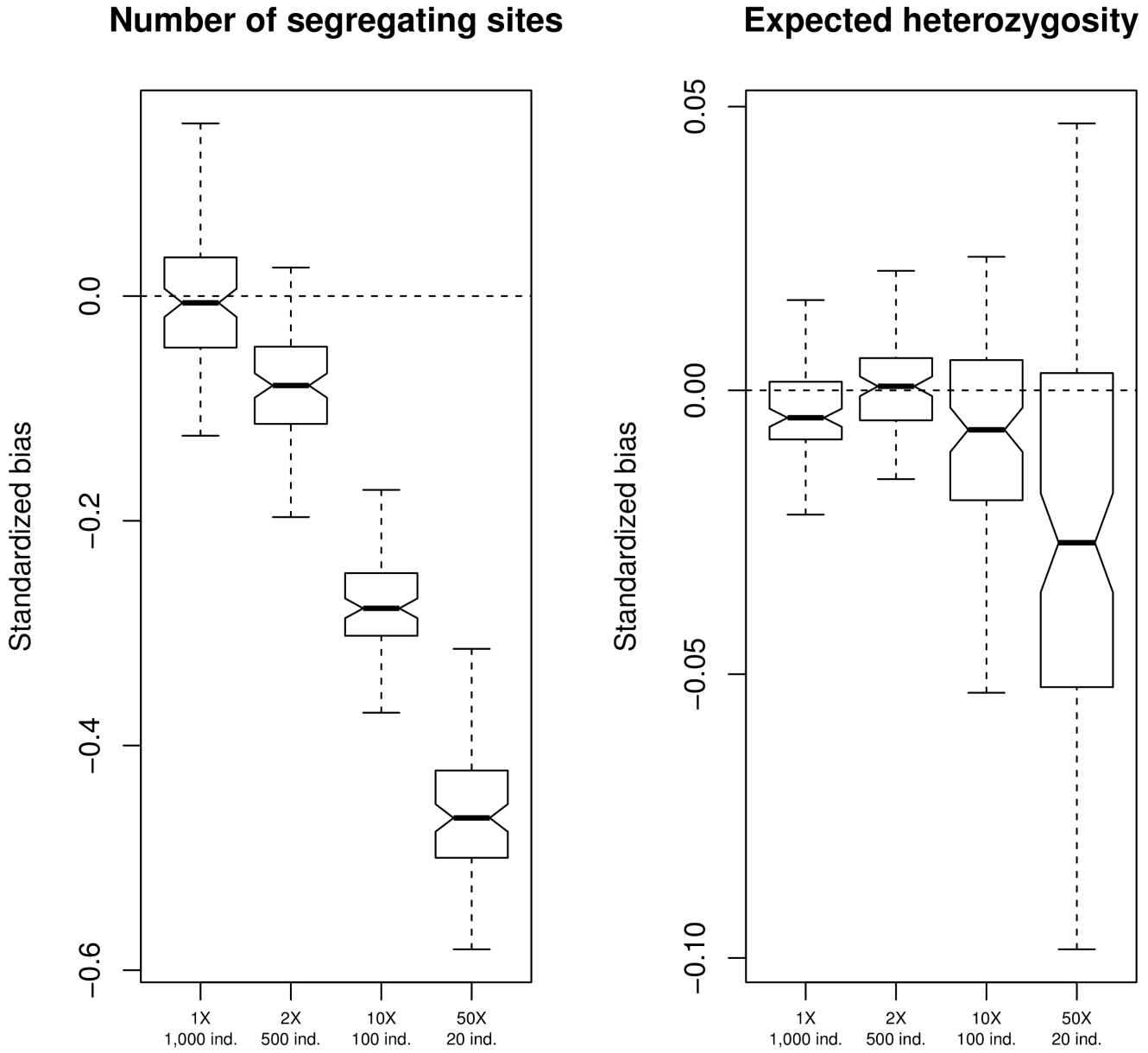

Figure 1. Nucleotide diversity estimation. Bias in the estimate of the number of segregating sites (left panel) and the expected heterozygosity (right panel) under different experimental scenarios. Sequencing depths are $1X$, $2X$, $10X$, and $50X$ and the corresponding sample sizes are 1,000, 500, 100, and 20 individuals. I simulated 100 regions of $1k$ independent sites, with a probability of each site being variable in the population equal to 0.1.

doi:10.1371/journal.pone.0079667.g001

Identifying Polymorphic Sites

SNP calling is the procedure for identifying which sites are polymorphic in a sample, and hence in the population from which the sample was drawn. The False Positive (FP) and False Negative (FN) rates, and Precision and Recall values (see Methods) were calculated under all experimental scenarios in order to assess SNP calling accuracy. FP measures how many non variable sites are misidentified as being polymorphic, while FN measures how many SNPs are not identified as being variable. Precision and Recall measure the proportion of relevant calls for FP and FN, respectively (see Methods). High values of Precision and Recall, and low values of FP and FN are desirable.

Precision and Recall values for SNP calling under different scenarios are shown in Table 1. A site was considered to be a SNP if its probability of being variable exceeded a given threshold, which was dynamically chosen to minimise the difference between the true and estimated number of variable sites in the entire population. This approach is not realistic outside of simulations, but guarantees an optimal equilibrium between FP and FN (i.e. their sum is approximately constant). As expected, Precision increases with higher sequencing depth. For instance, at $50X$, Precision is 1, indicating that all called SNPs are truly polymorphic. On the other hand, as sequencing depth increases and the sample size is reduced, Recall values decrease. This reflects the inability to call variable sites when heterozygous individuals are not sequenced. The highest Recall is obtained at $2X$ sequencing depth for 500 individuals, at which point the Precision is comparable to a scenario that uses depth of $10X$ for

Table 1. SNP calling Precision and Recall.

Sequencing depth	Sample size	Precision	Recall
1X	1,000	0.737 (0.0437)	0.749 (0.0472)
2X	500	0.778 (0.0461)	0.771 (0.0446)
10X	100	0.779 (0.0441)	0.725 (0.0408)
50X	20	1 (0)	0.540 (0.0582)

Precision and Recall values for detecting polymorphic sites at different scenarios of sequencing depth and sample size. Values are averaged across 100 different replicates and standard deviations are reported in parentheses.
doi:10.1371/journal.pone.0079667.t001

100 individuals. Similar results are obtained when filtering out sites with low total sequencing depth (see Methods) (Table S2). As expected, when identifying polymorphisms solely at the sequenced sample level, as opposed to the population level, both Precision and Recall increase with higher sequencing depth (Table S3).

These trends, as well as the distribution of FP and FN rates, are similar across all windows (Figure 2). The FP rate is higher in cases of low sequencing depth, especially at $1X$, while it is 0 at $50X$. The opposite effect is observed for FN rates, which are higher at $50X$; specifically, almost 50% of true SNPs are not detected. The median FN rate at $2X$ is the lowest among all tested experimental conditions (Figure 2). Similar results are obtained when simulating genotype frequencies not in HWE (Figure S5), and for a population under an expansion model, although, in the latter case, $1X$, $2X$, and $10X$ designs show comparable levels of accuracy (Figure S6).

Then, I performed SNP calling by assigning polymorphisms if the probability of being variable was greater than a fixed threshold, namely 0.95. This strategy is similar to common practice. For all scenarios, FP rates drop to 0, while FN rates increase and median values are above 30% (Figure S7). Indeed, SNPs are called only if high confidence is achieved. Similar results are obtained in case of population expansion (Figure S8) and deviation from HWE (Figure S9). A less stringent threshold for SNP calling reduces the FN rate (Figure S10), while a more stringent cut-off increases FN values (Figure S11).

SNP calling accuracy was also assessed when confined to common variants, defined as sites with a minor allele frequency greater than 0.01, which is equivalent to an absolute frequency of 20 chromosomes, out of 2,000 total chromosomes, bearing the alternate allele. SNPs were called if their probability of being variable was greater than 0.95. Notably, FN rates have a median equal to 0 in $1X$ and $2X$ cases, while it is close to 15% for $50X$ (Figure S12). Accuracy increases if rare variants, which are more likely not to be identified, are ignored. Similar results were obtained in the cases of population expansion (Figure S13) and deviation from HWE (Figure S14).

The results suggest that SNP calling is greatly influenced by the joint effect of sample size and sequencing depth. Generally, high sequencing depth provides greater Precision, while greater Recall is obtained with higher sample size. However, calling SNPs using a common strategy based on the probability of each site being variable reduces FP rates to 0 for all scenarios. Nevertheless, FN rates are always greater than 0 with small sample sizes. A sequencing depth lower than $2X$ precludes accurate identification of variable sites because of the lower chance of sequencing both alleles at the individual level. These findings are robust to different assumptions for population size changes and deviation from HWE. As expected, most of the misidentified true variable sites have low minor allele frequency. Indeed, SNP calling on common variants produces FN rates close to 0 for all sequencing configurations except at the lowest sample size.

Predicting Population Structure

I simulated sequencing data for multiple sub-populations to test the accuracy of inferring population structure under different sequencing depth and sample size conditions. Specifically, I simulated 3 populations of 40 individuals each, at different levels of genetic differentiation, with the per-sample sequencing depth set to $1X$, $2X$, $10X$, and $20X$, and corresponding sample sizes of 40, 20, 4, and 2 individuals from each of the 3 populations, so that total sequencing depth was equal across designs. One hundred simulations were performed under each sequencing scenario to account for variation in individual sub-sampling (see Methods).

The first 2 Principal Components (PCs) in a Principal Components Analysis (PCA) were used to train a predictive model of population structure on a 2-dimensional grid through a Support Vector Machine (SVM) technique. For each cell of the grid, I assigned a population based on the model trained from known genotypes and from sequencing data. The proportion of mislabelled cells, where the model from sequencing data predicts a different population than the model trained by known genotypes (see Methods), was recorded. Accuracy in predicting population structure is then inversely proportional to the fraction of mislabelled cells, and can be quantified on an arbitrary grid.

Results show that the design with 1X sequencing depth and 40 individuals sampled from each population achieves the highest accuracy in predicting population structure (Figure 3). This effect is more pronounced for cases involving low-to-medium genetic differentiation between populations. Under these conditions, sequencing less samples produces 50% more mislabelled cells, on average, than using all individuals at very low sequencing depth. Similar results were obtained with a less dense grid (Figure S15), and when simulating only variable sites (Figure S16). The latter finding suggests that monomorphic sites do not influence predictions even at low sequencing depth.

To illustrate the overall trend in distinguishing population structure, the inferred population structure was plotted over a grid for a single simulation, assuming low genetic differentiation among populations. For each scenario, a simulation having accuracy equal to the median for the entire distribution was chosen to represent the overall behaviour. Figure 4 shows that most of the mislabelled cells lie on the borders between populations. As already seen in Figure 3, an experimental design in which all individuals have been sequenced at low depth provides the greatest accuracy for predicting population structure.

Conclusions

For this study, extensive simulations were performed under a wide range of sequencing designs to test the joint effect of sequencing depth and sample size on population genetics inferences. The results suggest that at a fixed sequencing budget, it is desirable to sequence a large number of individuals, at the cost of reducing the per-sample sequencing depth.

To estimate allele frequencies and identify polymorphic sites, sequencing the largest possible sample size with at least a per-sample sequencing depth of $2X$ is recommended. Similarly, population structure is more accurately inferred at low depth with large sample sizes, and even at depth as low as 1X if a large enough sample size is used.

It is also important to consider that state-of-the-art statistical methods to estimate genetic variation from NGS data were used

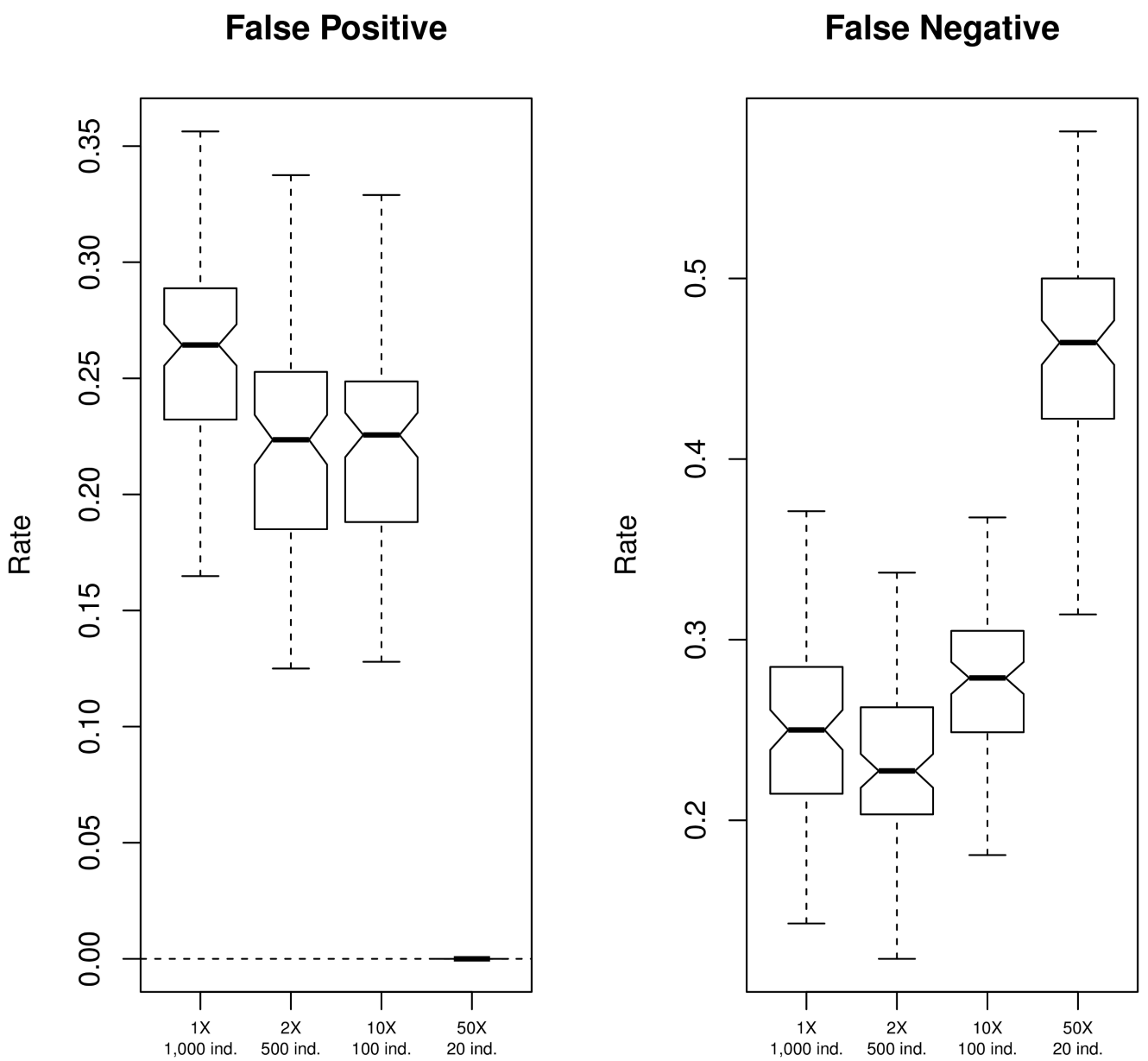

Figure 2. SNP calling accuracy. False Positive and False Negative rates in the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 1. Sites were identified as polymorphic if their probability of being variable was above a threshold, chosen to minimise the difference between the true and the estimated number of SNPs (see Methods).
doi:10.1371/journal.pone.0079667.g002

[11]. These approaches, based on genotype likelihoods, provide superior estimates to methods employing strict genotype calling [11,17,33], and therefore should be adopted in all population genetics studies using low-medium coverage sequencing data.

I believe that this study will assist researchers in their experimental design. The approach for testing the effect of experimental conditions on population genetics inferences used in this study can be extended to other fields in genomics and medical genetics.

Methods

Simulating Sequencing Data

Sequencing data was extensively simulated to assess the accuracy of estimating nucleotide variation and population structure under different experimental scenarios. Simulated individual genotypes were assigned assuming Hardy-Weinberg Equilibrium (HWE), and an inbreeding coefficient of 0 or 0.3, given an ancestral population allele frequency. This ancestral allele frequency was drawn from an exponential distribution, which is proportional to the expected allele frequency distribution under a standard neutral diffusion model [34]. To mimic the genomic effect of population expansion, I artificially skewed the expected allele frequency distribution towards low frequency variants by squaring, and then normalising, the values in the site frequency spectrum. The number of reads at each locus for each individual was drawn from a Poisson distribution [10,35]. Sequencing errors were randomly and uniformly introduced among reads at rates of 0.005 and 0.01, which are comparable to empirical error rates

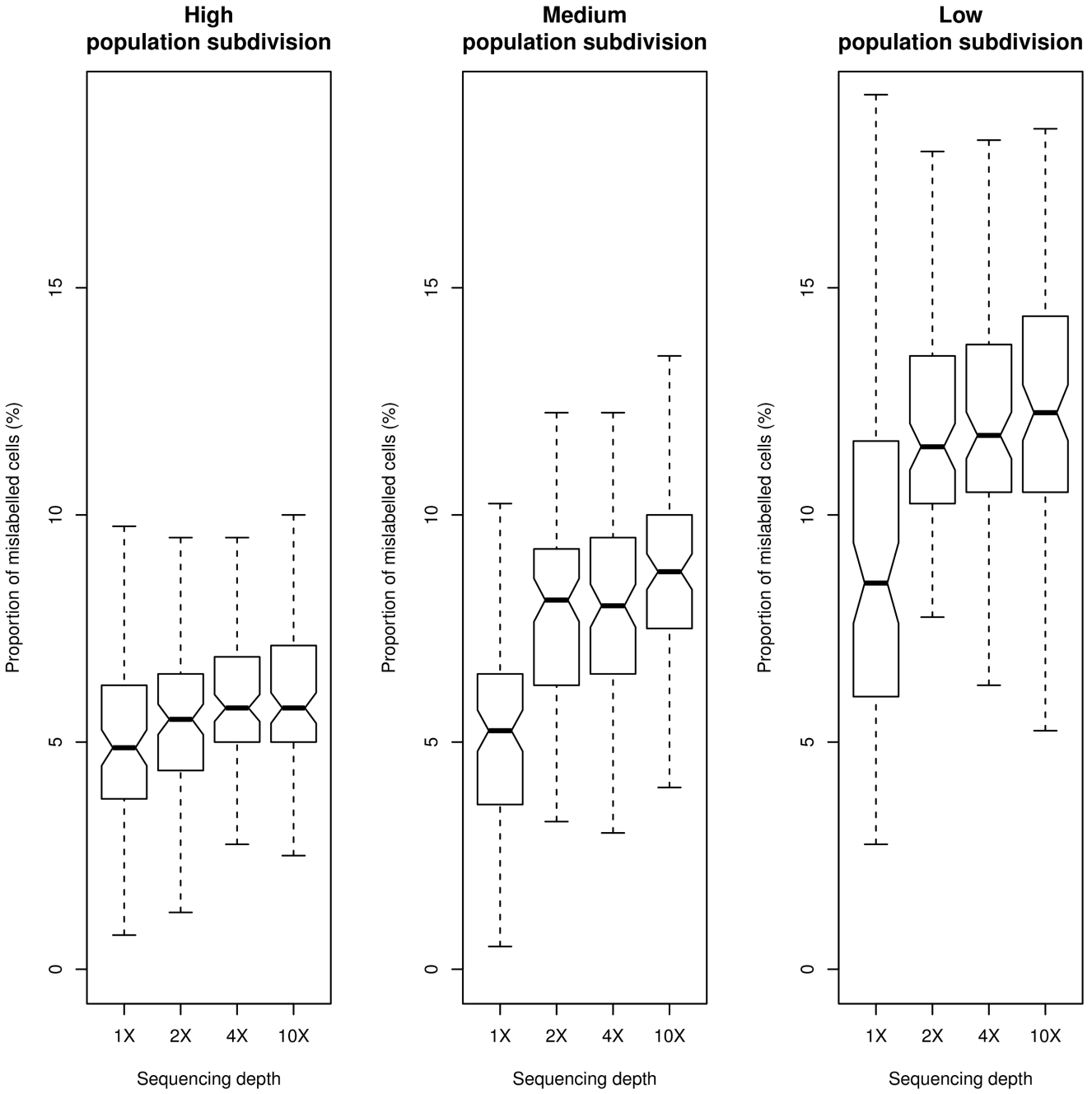

Figure 3. Population structure inference accuracy. Accuracy of population structure inference, measured as the proportion of cells over a $20x20$ grid where sub-populations have been wrongly assigned from sequencing data compared to the case of known genotypes for all individuals (see Methods). Sequencing depths are $1X$, $2X$, $10X$, and $20X$ and the corresponding sample sizes are 120, 60, 12, and 6 individuals. I simulated $20k$ independent sites, with a probability of each site being variable in the population equal to 0.1. Populations were simulated with high genetic subdivision (left panel, F_{ST} 0.4 and 0.1), medium genetic subdivision (mid panel, F_{ST} 0.3 and 0.05), low genetic subdivision (right panel, F_{ST} 0.1 and 0.02).

doi:10.1371/journal.pone.0079667.g003

[26,27]. The probability of a site being polymorphic in the population was set to 0.01, 0.1, and 1.

For analyses related to estimating within-population nucleotide diversity, the individual per-site mean sequencing depths (the average number of mapped reads) were set to $1X$, $2X$, $10X$ or $50X$ for different corresponding sample sizes in order to achieve a constant total sequencing depth of $2000X$ across all individuals. I simulated $1M$, and $100k$, independent diallelic sites for 1,000 individuals. The information content produced by these simulations is comparable to the output of current high-throughput sequencing machines.

To simulate population structure, sub-population allele frequencies were drawn from a Beta distribution [36] with mean equal to the ancestral population allele frequency [37]. To simulate data from 3 populations, allele frequencies for two sub-populations were drawn as just described and the first of these

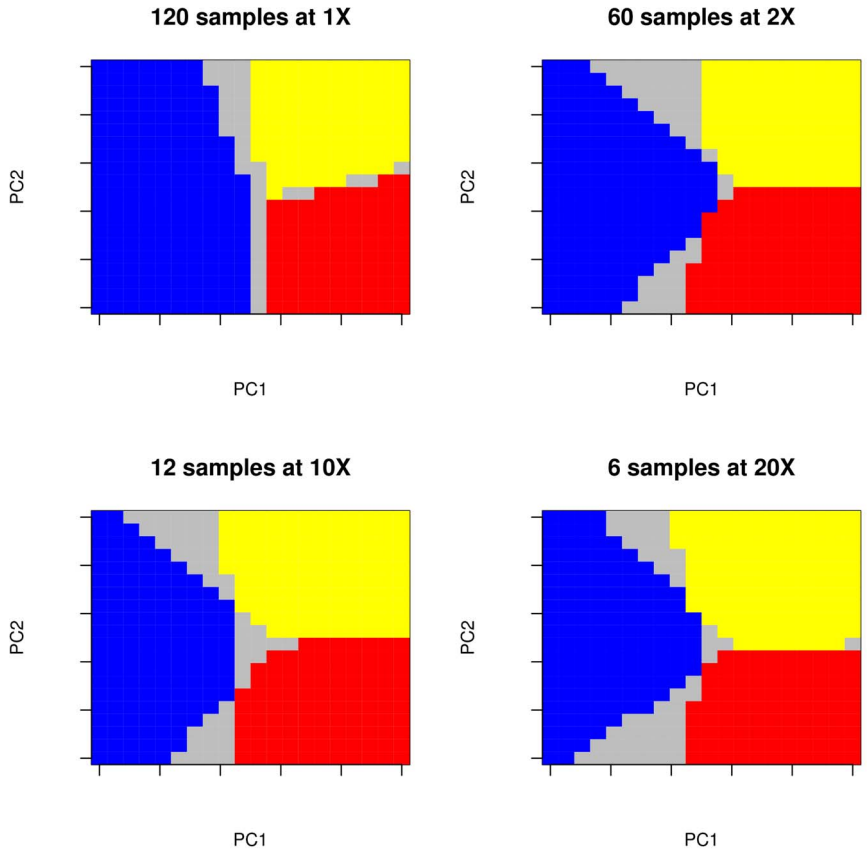

Figure 4. Population structure prediction. Population structure predicted over a $20x20$ grid for a single replicate under different experimental scenarios. Simulations were performed as described in Figure 3, in the case of low genetic subdivision. Grey cells represent locations where a different sub-population was predicted to be located from sequencing data compared to the case of known genotypes of all individuals. These particular replicates show a proportion of mislabelled cells equal to be the medium of the distribution. Note that replicates are not the same across the different tested scenarios.
doi:10.1371/journal.pone.0079667.g004

frequencies was assigned to sub-population 1. The second allele frequency was assigned as the ancestral allele frequency for sub-populations 2 and 3. To model variable degrees of genetic sub-division among populations in the Beta distribution [36], different values of F_{ST}, a common measure of population genetics differentiation [38], were assumed. I simulated population structure with low (F_{ST} values of 0.1 and 0.02), medium (F_{ST} values of 0.3 and 0.05), and high (F_{ST} values of 0.4 and 0.1) genetic sub-division.

For population structure analyses, I simulated 3 populations of 40 individuals each, and a total of $20k$ independent diallelic sites. Then, 40, 20, 4, or 2 individuals per population were sampled, with corresponding sequencing depth of $1X$, $2X$, $10X$, and $20X$, resulting in a total sequencing depth of $80X$ per population. Given that individuals can be sampled in many different combinations, I performed 100 replicates for each experimental scenario.

Computing Nucleotide Diversity from Sequencing Data

Accuracy for estimating nucleotide diversity from sequencing data was assessed by first dividing all $1M$, and $100k$, simulated sites into 100 $10k$, and $1k$, non-overlapping windows. For each window, I calculated the proportion of segregating sites (S) as the fraction of variable sites in the sample, and the expected heterozygosity (H). In the case of known genotypes, these quantities can be easily calculated across L sites as:

$$S = \sum_{s=1}^{L} I_s \tag{1}$$

where I_s is an indicator function equal to 1 when at least one individual is heterozygous at site s, and 0 otherwise, and

$$H = \sum_{s=1}^{L} 2f_s(1-f_s); \tag{2}$$

where f_s is the reference allele frequency for a site, s, in the sample.

When genotypes are unknown, they must be inferred from the mapped sequence read data. Current studies use genotype likelihoods to ultimately call genotypes when necessary. Genotype likelihoods are a function of both base calls and quality scores and are proportional to the probability of the observed data given a certain genotype, for a given site in an individual [12,39]. Bayesian methods have been proposed to calculate the posterior probability $P(G_s^{(z)}|X)$ of genotype G at site s for individual z given the observed data X [11,12]. The prior for obtaining these posteriors can be derived from an estimate of the allele frequency [11]. Similarly, empirical Bayes methods have been proposed to calculate the posterior probability $P(f_s|X)$ of the sample allele frequency f at site s [11,12]. $P(G_s^{(z)}|X)$ and $P(f_s|X)$ from

simulated sequencing reads were computed using ANGSD software (http://popgen.dk/angsd).

Nucleotide diversity indices were calculated in a way that accounts for genotyping uncertainty, rather than strictly assigning individual genotypes. This probabilistic framework has been successfully adopted to estimate population genetics parameters from low sequencing depth data [11,12,14,16,17]. Throughout the study, the ancestral and derived allelic state were assumed to be known, and "allele frequency" refers to the frequency of the derived allele. All motivations are still valid under the folded site frequency spectrum (when ancestral and derived state are unknown).

Estimates of S and H from sequencing data can be calculated as:

$$\hat{S} = \sum_{s=1}^{L} (1 - P(f_s = 0|X) - P(f_s = 2k|X)) \quad (3)$$

where k is the number of diploid individuals in the sample, and

$$\hat{H} = \sum_{s=1}^{L} \sum_{i=0}^{2k} (2(\frac{i}{2k})(\frac{2k-i}{2k}) P(f_s = i|X)) \quad (4)$$

where $P(f_s = 0|X)$, $P(f_s = 2k|X)$, $P(f_s = i|X)$ is the posterior probability of having 0, $2k$, and i chromosomes with the derived allele at site s, respectively [14].

Several experimental scenarios were explored by varying sequencing depth and sample size, while keeping their product (the total sequencing coverage) constant. 1,000, 500, 100, and 20 samples at $1X$, $2X$, $10X$, and $50X$, respectively were sub-sampled from the entire pool of 1,000 individuals. To assess the accuracy for estimating nucleotide variation under different experimental scenarios, the standardised bias between estimates obtained from known genotypes for all individuals and from unknown genotypes, for each window, was calculated as:

$$Bias(S) = \frac{\hat{S} - S}{S} \quad (5)$$

and

$$Bias(H) = \frac{\hat{H} - H}{H}. \quad (6)$$

Positive values of $Bias(S)$ and $Bias(H)$ therefore indicate over-estimation of true values, while negative values indicate under-estimation. To directly quantify the effect of this bias on population genetics estimates, I identified windows showing extremely low or high values of H from the empirical distribution of all 100 windows for each experimental scenario. The number of correctly identified outliers using sequencing data, and the correlation between H and $\hat{H}$ were used to measure estimation accuracy.

In the case of unknown genotypes, identifying variable sites in the sample can be achieved by detecting sites with a probability of being variable, calculated as $(1 - P(f_s = 0|X) - P(f_s = 2k|X))$ (see Equation 3), greater than a certain threshold. For each simulation, this threshold was dynamically chosen to minimise the difference between the number of true and estimated variable sites, in order to realise an optimal trade-off between SNP over-calling and SNP under-calling. Additional analyses were performed by setting the probability of being variable threshold to fixed values.

I evaluated the accuracy of SNP calling by computing False Positive (FP) and False Negative (FN) rates. Precision and Recall values were derived from these quantities. Precision is computed as the ratio of True Positive (TP) rates to (TP+FP), while Recall is the ratio of TP to (TP+FN). The average and standard deviation for Precision and Recall, and FP and FN rates, were calculated across all 100 windows to inspect their distribution.

Predicting Population Structure from Sequencing Data

I assessed the prediction accuracy of population structure under different experimental scenarios. Specifically, I compared the predicted population structure in the case of known genotypes from all individuals to the structure determined from the sequencing data for the entire pool of individuals, or a subset of it, at a fixed total sequencing depth. A total of 120 individuals with known genotypes were sampled from 3 different sub-populations. Sample sizes of 40, 20, 4, and 2 individuals from each of the 3 populations, at $1X$, $2X$, $10X$, and $20X$ sequencing depth, respectively, were examined.

Principal Component Analyses (PCA) was used to inspect population genetics structure. The PCA is ultimately based on a covariance matrix of individual genotypes [40]. In the original latter approach, the denominator normalises the allele frequency variance. However, this normalisation over-weights low frequency variants and is therefore not suitable for NGS data, for which estimates of rare variants are usually less confident. Thus, the normalisation was not applied, without loss of generalisation throughout all analyses.

In cases where the genotypic covariance matrix had to be inferred directly from the sequencing data, previously proposed methods [17] were followed. Briefly, the posterior probability for the covariance matrix is approximated from the genotype posterior probabilities at each site for each individual. The covariance matrix is finally weighted by the probability of each site of being variable. This approach has been shown to perform well in cases of low sequencing depth and converges to standard genotype calling methods in cases of high sequencing depth [17]. Eigenvector decomposition of the covariance matrix is then performed to obtain the first 2 Principal Components (PCs). Given the simulation scheme used, these PCs contain the full information on population structure, while other PCs are likely to represent only stochastic noise. Procrustes Analysis techniques [41] were used to compare PCs obtained from the case of known genotypes and the case of unknown genotypes. Specifically, the PCs coordinates derived from unknown genotypes were rotated and scaled to minimise the distance to the corresponding coordinates of PCs computed from known genotypes.

A Support Vector Machine (SVM) algorithm was adopted to model and predict population structure over a 2-dimensional grid. SVM receives a training set of features and categories, and trains a machine to model the relationship between them. PCs coordinates were set as uncorrelated features and the population labelling at each set of coordinates as categories, and a model, for both the case of known genotypes and unknown genotypes, was estimated.

From these models, I predicted the population structure over a grid of $20x20$ cells, as well as $10x10$ cells, from the model estimated from known and unknown genotypes separately. In other words, for each cell of the grid I predicted which population is located at that particular set of coordinates. I used the same grid, obtained by equally partitioning the PCs plane from known genotypes, for both models. Finally, the proportion of mislabelled populations between the model from known genotypes and from

unknown genotypes over the entire grid was used as a measure of population structure prediction accuracy.

Programs to simulate sequencing data and to perform all described analyses are available at https://github.com/mfumagalli/ngsTools. All statistical analyses were performed in the R environment (www.r-project.org).

Supporting Information

Figure S1 Nucleotide diversity estimation with lower level of polymorphisms. Bias in the estimate of the number of segregating sites (left panel) and the expected heterozygosity (right panel) under different experimental scenarios. Simulation were performed as described in Figure 1. I simulated 100 segments of $10k$ independent sites with the probability of each site being variable in the population equal to 0.01.
(TIF)

Figure S2 Nucleotide diversity estimation with lower sequencing error rate. Bias in the estimate of the number of segregating sites (left panel) and the expected heterozygosity (right panel) under different experimental scenarios. Simulations were performed as described in Figure 1. The sequencing error rate was set to 0.005.
(TIF)

Figure S3 Nucleotide diversity estimation under population size expansion. Bias in the estimate of the number of segregating sites (left panel) and the expected heterozygosity (right panel) under different experimental scenarios. Simulations were performed as described in Figure 1. Populations were simulated under a size expansion model.
(TIF)

Figure S4 Nucleotide diversity estimation with inbreeding. Bias in the estimate of the number of segregating sites (left panel) and the expected heterozygosity (right panel) under different experimental scenarios. Simulations were performed as described in Figure 1. Genotypes were simulated assuming an individual inbreeding coefficient of 0.3.
(TIF)

Figure S5 SNP calling accuracy with inbreeding. False Positive and False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Genotypes were simulated assuming an individual inbreeding coefficient of 0.3.
(TIF)

Figure S6 SNP calling accuracy under population size expansion. False Positive and False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Populations were simulated under a size expansion model.
(TIF)

Figure S7 SNP calling accuracy using a fixed cut-off. False Positive and False negative rates in the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Sites were identified as polymorphic if their probability of being variable was above 0.95.
(TIF)

Figure S8 SNP calling accuracy using a fixed cut-off under population size expansion. False Positive and False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Sites were identified as polymorphic if their probability of being variable was above 0.95. Populations were simulated under a size expansion model.
(TIF)

Figure S9 SNP calling accuracy using a fixed cut-off with inbreeding. False Positive and False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Sites were identified as polymorphic if their probability of being variable was above 0.95. Genotypes were simulated assuming an individual inbreeding coefficient of 0.3.
(TIF)

Figure S10 SNP calling accuracy using a less stringent fixed cut-off. False Positive and False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Sites were identified as polymorphic if their probability of being variable was above 0.90.
(TIF)

Figure S11 SNP calling accuracy using a more stringent fixed cut-off. False Positive and False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Sites were identified as polymorphic if their probability of being variable was above 0.99.
(TIF)

Figure S12 SNP calling accuracy for common variants using a fixed cut-off. False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure 2. Sites were identified as polymorphic if their probability of being variable was above 0.95. Only sites with a true sample allele frequency greater than 0.01 were retained. Outliers are plotted as circles.
(TIF)

Figure S13 SNP calling accuracy for common variants using a fixed cut-off under population size expansion. False negative rates in the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure S12. Populations were simulated under an expansion size model.
(TIF)

Figure S14 SNP calling accuracy for common variants using a fixed cut-off with inbreeding. False negative rates for the identification of polymorphic sites under different experimental scenarios. Simulations were performed as described in Figure S12. Genotypes were simulated assuming an individual inbreeding coefficient of 0.3.
(TIF)

Figure S15 Population structure inference accuracy over a less dense grid. Accuracy of population structure inference, measured as the proportion of the cells over a $10x10$ grid where sub-populations have been wrongly assigned compared to the case of known genotypes for all individuals (see Methods). Simulations were performed as described in Figure 3. Populations were simulated with high genetic subdivision (upper left panel, F_{ST} 0.4 and 0.1), medium genetic subdivision (upper right panel, F_{ST} 0.3 and 0.05), low genetic subdivision (lower left panel, F_{ST} 0.1 and 0.02). I also simulated $2k$ independent variable sites at medium genetic subdivision (lower right panel).
(TIF)

Figure S16 Population structure inference accuracy with all sites variable in the population. Accuracy of population structure inference, measured as the proportion of the cells over a $20x20$ grid where sub-populations have been wrongly assigned compared to the case of known genotypes for all individuals (see Methods). Simulations were performed as described in Figure 3. I simulated $2k$ independent variable sites at medium genetic subdivision (F_{ST} 0.3 and 0.05).
(TIF)

Table S1 Power to detect outliers in the distribution of nucleotide diversity. Accuracy of detecting outliers in the distribution of nucleotide diversity. Simulations were performed as described in Figure 1. The number of top and bottom (5 or 10 out of 100) windows from the distribution of H calculated from known genotypes that were correctly identified using sequencing data. Wilcoxon-test correlation between H and $\dot{H}$ (see Methods) is also shown.
(PDF)

Table S2 SNP calling Precision and Recall with data filtering. Precision and Recall values for detecting polymorphic sites at different scenarios of sequencing depth and sample size. Analyses were performed as described in Table 1. Sites with a total sequencing depth below the $10th$ percentile were discarded.
(PDF)

Table S3 SNP calling Precision and Recall for the sample. Precision and Recall values for detecting polymorphic sites at different scenarios of sequencing depth and sample size. Analyses were performed as described in Table 1. Accuracy was estimated by comparing true and estimated SNPs variable in the specific sample size, and not in the entire population of 1,000 individuals.
(PDF)

Acknowledgments

I am extremely grateful to Tyler Linderoth for editing the final version of the manuscript, to Filipe G. Vieira and Thorfinn Sand Korneliussen for technical support, and to Erik Garrison and one anonymous reviewer for insightful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: MF. Performed the experiments: MF. Analyzed the data: MF. Contributed reagents/materials/analysis tools: MF. Wrote the paper: MF.

References

1. Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in snp chips affect measures of population divergence. Molecular biology and evolution 27: 2534–2547.
2. Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. Genome research 20: 291–300.
3. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, et al. (2011) Exome sequencing as a tool for mendelian disease gene discovery. Nature reviewsGenetics 12: 745–755.
4. Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. Trends in genetics : TIG 26: 177–187.
5. Metzker ML (2010) Sequencing technologies - the next generation. Nature reviews Genetics 11: 31–46.
6. Crawford JE, Lazzaro BP (2012) Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. Frontiers in genetics 3: 66.
7. Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, et al. (2013) Unlocking the vault: next generation museum population genomics. Molecular Ecology under review.
8. Lynch M (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182: 295–301.
9. Keightley PD, Halligan DL (2011) Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. Genetics 188: 931–940.
10. Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, et al. (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. BMC bioinformatics 12: 231.
11. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PloS one 7: e37558.
12. Li H (2011) A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England) 27: 2987–2993.
13. Kang CJ, Marjoram P (2011) Inference of population mutation rate and detection of segregating sites from next-generation sequence data. Genetics 189: 595–605.
14. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science (New York, NY) 329: 75–78.
15. Gompert Z, Buerkle CA (2011) A hierarchical bayesian model for next-generation population genomics. Genetics 187: 903–917.
16. Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, et al. (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. Evolution; international journal of organic evolution 66: 2167–2181.
17. Fumagalli M, Vieira FG, Korneliussen TS, Linderoth TP, Huerta-Sanchez E, et al. (2013) Quantifying population genetic differentiation from next-generation sequencing data. Genetics Epub ahead of print.
18. Nielsen R (2005)Molecular signatures of natural selection. Annual Review of Genetics 39: 197–218.
19. Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.
20. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics 123: 585–595.
21. Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from dna sequence data. Genetics 132: 583–589.
22. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693–709.
23. Jay F, Sjdin P, Jakobsson M, Blum MG (2013) Anisotropic isolation by distance: the main orientations of human genetic differentiation. Molecular biology and evolution 30: 513–525.
24. Price AL, Patterson NJ, Plenge RM,WeinblattME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38: 904–909.
25. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in europeans. Science (New York, NY) 201: 786–792.
26. Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
27. Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.
28. Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, et al. (2012) A fine-scale chimpanzee genetic map from population sequencing. Science (New York, NY) 336: 193–198.
29. Huang X, Kurata N, Wei X, Wang ZX, Wang A, et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490: 497–501.
30. Wong LP, Ong RT, Poh WT, Liu X, Chen P, et al. (2013) Deep whole-genome sequencing of 100 southeast asian malays. American Journal of Human Genetics 92: 52–66.
31. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740–743.
32. Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R (2013) Estimating inbreeding coefficients from ngs data: impact on genotype calling and allele frequency estimation. Genome Research Epub ahead of print.
33. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and snp calling from next-generation sequencing data. Nature reviewsGenetics 12: 443–451.
34. Ewens W (2004) Mathematical Population Genetics: Theoretical Introduction. Springer.
35. Kim SY, Li Y, Guo Y, Li R, Holmkvist J, et al. (2010) Design of association studies with pooled or un-pooled next-generation sequencing data. Genetic epidemiology 34: 479–491.
36. Balding DJ (2003) Likelihood-based inference for genetic correlation coefficients. Theoretical population biology 63: 221–230.
37. Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theoretical population biology 60: 227–237.
38. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting f(st). Nature reviewsGenetics 10: 639–650.
39. Li H, Ruan J, Durbin R (2008) Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome research 18: 1851–1858.

40. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS genetics 2: e190.

41. Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, et al. (2010) Comparing spatial maps of human population-genetic variation using procrustes analysis. Statistical applications in genetics and molecular biology 9: Article 13.