OPEN ACCESS Freely available online

PLOS ONE

SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data

Rasmus Nielsen[1,2,3]*, Thorfinn Korneliussen[3], Anders Albrechtsen[3], Yingrui Li[1], Jun Wang[1,3]*

1 BGI-Shenzhen, Shenzhen, China, **2** Departments of Integrative Biology and Statistics, University of California, Berkeley, California, United States of America, **3** Department of Biology, University of Copenhagen, Copenhagen, Denmark

Abstract

We present a statistical framework for estimation and application of sample allele frequency spectra from New-Generation Sequencing (NGS) data. In this method, we first estimate the allele frequency spectrum using maximum likelihood. In contrast to previous methods, the likelihood function is calculated using a dynamic programming algorithm and numerically optimized using analytical derivatives. We then use a Bayesian method for estimating the sample allele frequency in a single site, and show how the method can be used for genotype calling and SNP calling. We also show how the method can be extended to various other cases including cases with deviations from Hardy-Weinberg equilibrium. We evaluate the statistical properties of the methods using simulations and by application to a real data set.

Citation: 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(7): e37558. doi:10.1371/journal.pone.0037558

Editor: Philip Awadalla, University of Montreal, Canada

Received December 27, 2011; **Accepted** April 25, 2012; **Published** July 24, 2012

Funding: Funding was provided by United States National Institutes of Health grant R01-HG003229. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rasmus_nielsen@berkeley.edu (RN); wangj@genomics.org.cn (JW)

Introduction

The biological sciences have been transformed by the emergence of New-Generation Sequencing (NGS) technologies providing cheap and reliable large scale sequencing (e.g, [1]). These technologies are used for *de novo* genome sequencing (e.g., [2]), in human disease genetics and diagnostics (e.g., [3,4]), in gene expression analyses (e.g., [5]), in population genetic studies (e.g., [6]), and in many other applications. In this paper, we will mostly be interested in population genetic applications. However, the methods used in this paper may also be helpful for genotype and SNP calling in other studies based on multiple individuals, such as association mapping studies.

Many NGS studies (e.g., [6,7,8] are based on medium to low coverage, i.e. coverage at $<$20X. While the price of NGS is declining, the demand for larger sample sizes is similarly increasing, suggesting that low or medium sequencing coverage may be the design of choice for many future studies in the years to come. In such data, genotype calling for each individual is associated with statistical uncertainty. There are two reasons for this. First, in heterozygous individuals, both alleles may not have been sampled. Secondly, the high raw error rates often associated with NGS may cause a significant amount of homozygous genotypes to be wrongly inferred as heterozygous, if genotype calling is based on just absence/presence of an allele. In most NGS, the error rate is at least 0.1% even after stringent filtering based on quality scores (e.g., [9]). In 5X data, an error will then appear in at least 0.5% of all homozygotes, i.e. at a level comparable to the SNP level. If multiple individuals are sampled, most SNPs will then in fact be errors. For this reason, more stringent criteria are typically used for calling SNPs and for calling heterozygote individuals. Some of these might in effect correspond to requiring the minor allele to be observed twice in an individual to be called. If such a criterion is applied, the chance of calling a heterozygous individual as homozygous in 5X data is approx. 0.375. More clever algorithms can be designed for calling SNPs and for calling genotypes than this (e.g., [10,11,12,13]), but if the coverage is low, they will be sharing the basic features outlined here: a trade-off between including too many SNPs and under-calling true heterozygotes. As a result, low coverage and medium coverage NGS data tends to provide biased estimates of the distribution of allele frequencies ([14,15,16,17] In this paper, we will explore the implications of this for population genetic inferences. We will also present and evaluate a set of algorithms for providing more precise SNP calls, genotype calls, and estimates of allele frequency. The strategy presented in this paper is to estimate the distribution of sample allele frequencies, the so-called Site Frequency Spectrum (SFS), jointly for all individuals and for all sites without calling individual genotypes. When first a good estimate of the SFS has been obtained, better priors can be defined for allele frequencies leading to improved genotype calling and SNP calling. For population genetic inferences, the SFS is in itself of primary interest, and population genetic inferences can proceed directly from the estimated SFS without using individual genotype calls. For example common estimators of effective population sizes and mutation rates, such as Watterson's estimator [18] and π [19] are simple functions of the SFS. Many methods for detecting natural selection, such as Tajima's D [19] are also simple functions of the SFS. Finally, methods for

estimating demographic parameters (e.g. [20]) and quantifying population subdivision using F_{ST} (e.g., [21]) also proceed from estimates of the SFS. For population genetic inferences from next-generation sequencing data, obtaining reliable estimates of the SFS is, therefore, fundamental.

We test the new methods using simulations and apply them to data from 200 previously sequenced human exomes. The methods developed here are available in the program package Analyses of Next-Generation Sequencing Data (ANGSD) downloadable from http://popgen.dk/software/angsd.html.

Methods

The SFS describes the distribution of allele frequencies. Let the proportion of SNPs, with a derived allele frequency of $i/2k$ in a sample of k diploid individuals, be p_i. The SFS is then given by the vector $(p_1, p_2, \ldots p_{2k-1})$. We here consider an expanded version of the SFS: the vector $\mathbf{P} = (p_0, p_1, \ldots p_{2k})$, i.e. we also consider sites in the alignment that are fixed. The zero category then represent sites in which all individuals are homozygous for the ancestral allele, and the $2k$ category represents sites that are fixed for the derived allele. The SFS also exists in a so-called folded version, $\mathbf{P}^* = (p_0^*, p_1^*, \ldots p_k^*)$, in which $p_i^* = p_i + p_{2k-i}$ for $i < k$ and $p_i^* = p_i$ for $i = k$. The folded version of the SFS is often used when no reliable information can be used to determine which allele is ancestral and which is derived.

As a note of notation, we distinguish between population allele frequencies and sample allele frequencies by denoting the former by p, as in the preceding section, and the latter by f. Most of the methods discussed in this paper concerns sample allele frequencies, but we also occasionally discuss the use of population allele frequencies. A number of previous papers have focused on population allele frequencies, including [22,23]. The methods presented here differ from those methods by focusing on sample allele frequencies, except otherwise stated.

Calculation of recalibrated quality scores and genotype likelihoods

Any method for SNP calling and allele frequency estimation must rely on a base calling algorithm and a method for calculating quality scores. A quality score is a function of the probability of the most likely base in a particular read given the observed data. It is typically reported using a phred scaling, i.e. as the $\log_{10}$ likelihood ratio relative to the most common base. Standard next-generation sequencing methods provide such quality scores associated with each base call. However, the raw quality scores are often not very accurate and must be re-calibrated taking observed error rates in the data into account. The objective of this paper is not to explore different methods for calculating and calibrating quality scores. The methods presented here can be used based on any method for calculating quality scores. However, in our data analyses we use a method similar to the method currently implemented in SOAPsnp [11]. In brief, the raw quality scores from Illumina reads are calibrated taking the observed allelic type and sequencing cycle (coordinate on read) into account. Using observed mismatch rates, the empirical probability of observing the data in a position of a read given the raw quality scores, the sequencing cycle, and the true allelic state can then be calculated. We interpret the probability calculated for read i of a particular site as a likelihood of a particular allele, $L_b^{(i)}$, $b \in B$, $B = \{A, C, T, G\}$. The genotype likelihood, in a site covered by r reads, can then be obtained as the product of individual allelic likelihood values (e.g., [24]):

$$p(X|G=bh) = \prod_{i=1}^{r} \left(L_b^{(i)}/2 + L_h^{(i)}/2 \right), b, h \in B. \quad (1)$$

Notice here that there is an implicit assumption regarding independence of reads in Equation (1). However, this is not the same as assuming Hardy-Weinberg Equilibrium (HWE) as the probability is calculated conditional on the genotype. Posterior probabilities will, in contrast, depend either on HWE assumptions or on an explicit modeling of deviations from HWE. It is also important to notice that the modeling of the error structure in the data is done through the calculation of the genotype likelihood.

Likelihood function for the allele frequency spectrum

For low coverage data, estimation of allele frequency for a particular site can be associated with great uncertainty. Likewise, SNP calling for rare SNPs can be difficult. However, as shown in the Results section based on methods developed here, the joint estimation for multiple sites in the genome of the distribution of allele frequencies, and the number of SNPs can be carried out with quite high accuracy.

Consider a statistical model in which the sample allele frequencies are free parameters, i.e. for k individuals there are $2k+1$ possible sample allele frequencies including 0 and 1. The vector of parameters is then $\mathbf{P} = (p_0, p_1, \ldots p_{2k})$ defined on the unit simplex $\{(p_0, \ldots, p_{2k}) \in \mathbb{R}^{2k+1} \mid \sum_{i=0}^{2k} p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\}$. These sample allele frequencies define the SFS with fixed ancestral and derived alleles included. The ith sample allele frequency, p_i, is the proportion of sites in the sample in which the derived allele has a frequency of $i/2$k in the sample, $i = 0,1,..,2k$. As the sample allele frequencies must sum to one, there are $2k$ parameters to estimate. Estimation of these $2k$ parameters assumes that the ancestral state of each SNP can be identified using outgroups (e.g. other primates for humans). However, if the identification of ancestral state is uncertain, the frequency spectrum can be folded, i.e. the number of observations in category i and category $2k$-i can be added together as described in the results section. For next-generation sequencing data $\mathbf{P}$ is not known, but must be estimated from the data. An estimate of $\mathbf{P}$ also provides an estimate of the fraction of variable sites (SNPs) in the sample as $1 - p_0 - p_{2k}$. Notice that there is here an implicit assumption that at most two nucleotides are present in the locus. We will later describe how to take into account the presence of more than two nucleotides, but will for now assume that there are at most two alleles, an ancestral allele (a) and a derived allele (A), and that they can be unambiguously identified in each site, except for sites with only one allele.

Assuming that genotype likelihoods can be calculated as discussed above, a likelihood function for $\mathbf{P}$ can be defined as a function of the genotype likelihood values. Let $X_d^{(v)}$ and $G_d^{(v)} \in \{0, 1, 2\}$ be the observed data and the unknown genotype, respectively, for individual d in site v. The genotype counts the number of derived alleles, i.e. $G_d^{(v)} = 0$ implies an aa genotype. The genotype likelihood for individual d in site v can then, with this expanded notation, be written as $p(X_d^{(v)}|G_d^{(v)})$. If the genotypes were known, the sampling probability, as a function of $\mathbf{P}$, in site v would be found by taking the product of the probability of the data given the sample allele frequency multiplied by the probability of the sample allele frequency, given $\mathbf{P}$, and then summing over all possible values of the sample allele frequency:

$$
\begin{aligned}
p(X^{(v)},G^{(v)}|\mathbf{P}) &= \sum_{j=0}^{2k} p(S_A=j|\mathbf{P})p(X^{(v)},G^{(v)}|S_A=j) \\
&= \sum_{j=0}^{2k} p_j p(X^{(v)},G^{(v)}|S_A=j) \\
&= \sum_{j=0}^{2k} p_j p(X^{(v)}|G^{(v)})p(G^{(v)}|S_A=j) \\
&= \sum_{j=0}^{2k} p_j \left[\prod_{d=1}^{k} p(X_d^{(v)}|G_d^{(v)})\right] p(G^{(v)}|S_A=j)
\end{aligned} \tag{2}
$$

where the function

$$
p(G^{(v)}|S_A=j) = \begin{cases} \binom{2k}{j}^{-1} 2^{\sum_{d=1}^{k} I(G_d^{(v)}=1)} & if \sum_{d=1}^{k} G_d^{(v)} = j, \\ 0 \text{ e l se} \end{cases} \tag{3}
$$

is the combinatorial probability that a sample contains the labeled genotype vector $G^{(v)} = \left(G_1^{(v)}, G_2^{(v)}, \ldots G_k^{(v)}\right)$ given that it contains a total of S_A alleles of the derived type. This expression assumes Hardy-Weinberg equilibrium.

However, the true genotypes are not known. The likelihood function for $\mathbf{P}$ must, therefore, be obtained by summing over all the unknown genotypes:

$$
\begin{aligned}
p(X^{(v)}|\mathbf{P}) &= \sum_{G^{(v)}} p(X^{(v)},G^{(v)}|\mathbf{P}) \\
&= \sum_{G(v)} \sum_{j=0}^{2k} p_j p(G^{(v)}|S_A=j) \prod_{d=1}^{k} p(X_d^{(v)}|G_d^{(v)}) \\
&= \sum_{j=0}^{2k} p_j \sum_{G^{(v)}} p(G^{(v)}|S_A=j) \prod_{d=1}^{k} p(X_d^{(v)}|G_d^{(v)}) \\
&= \sum_{j=0}^{2k} p_j \sum_{G_1^{(v)}} \cdots \sum_{G_k^{(v)}} p(G^{(v)}|S_A=j) \prod_{d=1}^{k} p(X_d^{(v)}|G_d^{(v)})
\end{aligned}
$$

Assuming independence among sites, we then multiply the likelihood among all

sites and obtain:

$$
L(\mathbf{P}) = \prod_{v} \left(\sum_{j=0}^{2k} Pj \sum_{G_1^{(v)}} \cdots \sum_{G_k^{(v)}} p(G^{(v)}|S_A=j) \prod_{d=1}^{k} p(X_d^{(v)}|G_d^{(v)}) \right). \tag{4}
$$

This likelihood function is the one underlying the EM algorithm applied in [24] and is, if ignoring the difference in handling of errors, also effectively identical to the likelihood function used in [25]. While it might initially appear very challenging to calculate this function for large values of k and v directly, a simple dynamic programming algorithm can be devised that greatly simplifies these calculations.

Direct evaluation of the likelihood function

The first step in the algorithm is to calculate the likelihood function for each site, $L_v(\mathbf{P})$, separately. In the following we will describe this algorithm, suppressing the index for site v in the notation to enhance readability:

Initialization:

Set $h_0 = p(X_1|G_1=0), h_1 = 2p(X_1|G_1=1), h_2 = p(X_1|G_1=2)$, and $h_j = 0$ for $j = 3,4,\ldots,2k$.

Recursion

For $d = 2, 3, \ldots, k$:

For $j = 2d, 2d\text{-}1, \ldots, 2$:

Set $h_j = p(X_d|G_d=2)h_{j-2} + 2p(X_d|G_d=1)h_{j-1} + p(X_d|G_d=0)h_j$

Set $h_1 = p(X_d \mid G_d=0)h_1 + p(X_d \mid G_d=1)h_0$

Set $h_0 = p(X_d \mid G_d=0)h_0$

Termination

Set $h_j = h_j \binom{2k}{j}^{-1}$ for $j = 0,1,2,\ldots,2k$.

The likelihood function can then be expressed as

$$
L(\mathbf{P}) = \prod_{v} \left(\sum_{j=0}^{2k} p_j h_j^{(v)} \right) \tag{5}
$$

where $h_j^{(v)}$ is the value of h_j calculated for the vth site $(= p(X^{(v)}|S_A=j))$. By tabulating the values of $h_j^{(v)}$ in a table of size $(2k+1)\times S$, where S is the total number of sites, the likelihood function can be re-calculated very fast for different values of $\mathbf{P}$. Notice that the computational speed is $O(k^2 S)$. Similar algorithms have been used for single site inferences in [8] and [26].

We have here assumed an unfolded (polarized) frequency spectrum. However, the algorithm can also be applied directly to folded data, but with a $k+1$ dimensional parameter space instead of a $2k+1$ dimensional parameter space.

Optimization

After tabulation of values of $h_j^{(v)}$ we optimize the likelihood function for $\mathbf{P}$ using the BFGS algorithm [27]. In order to do that we transform the parameter space from $2k+1$ to $2k$ parameters. The transformation used is

$$
p_0 = 1/\left(1 + \sum_{i=1}^{2k} \theta_i\right) \text{ and } p_j = \theta_j/\left(1 + \sum_{i=1}^{2k} \theta_i\right), j \in \{1,\ldots,2k\}. \tag{6}
$$

We then optimize the log likelihood function with respect to the transformed parameters $\theta = (\theta_1 \ldots \theta_{2k})$ using analytical derivatives. Application of standard calculus techniques lead to the following derivatives of the log likelihood function for the transformed parameters:

$$
\frac{\partial \ell(\mathbf{q})}{\partial \theta_i} = \sum_{v} -\left(1 + \sum_{j=1}^{2k} \theta_j\right)^{-1} + \frac{h_i^{(v)}}{h_0^{(v)} + \sum_{j=1}^{2k} \theta_j h_j^{(v)}}. \tag{7}
$$

The BFGS algorithm can then be applied to $\boldsymbol{\theta}$, and the estimates of the natural parameters, $\mathbf{P}$, can be found by using the transformation in eq. (6).

Unknown derived allele

The representation given above assumes that the ancestral and derived (if it exists) alleles always can be unambiguously identified.

However, for most next-generation sequencing data, there might be considerable difficulties in separating errors from true low frequency alleles. If the ancestral allele is the common allele, identification of the derived allele will then be ambiguous. The frequency spectrum is only properly defined for di-allelic loci. The approach we will take to this problem is to assume that all loci are truly di-allelic, and errors are responsible when more than two alleles are observed. For most human data, mutation rates are so low that this should be a reasonable approximation. The likelihood function can then be modified by calculating the likelihood for each locus assuming any of the three possible derived alleles, and then adding these likelihood values together, weighted by the probability that each possible derived allele is truly the derived allele. This probability has been set to 1/3 in all analyses presented in this paper. But we note that the inference method could potentially be improved by instead using empirical substitution matrices for this weighting.

We also note that a situation might arise where the inferred ancestral allele is not observed in the data, but two other alleles are segregating, both at high frequency. In these cases the unfolded frequency spectrum is not well-defined. Such loci are typically ignored in population genetic analyses, and will also be ignored here.

SNP calling and Empirical Bayes estimation of allele frequencies at individual sites

To estimate the sample allele frequency in a single site, we could in theory sum the posterior expectation of the marginal allele frequency calculated for each individual together for all individuals. However, in most applications it will be desirable to obtain the joint posterior distribution for the allele frequency, as downstream inferences then can be performed by integrating over this distribution.

The ML estimates can be used directly in inferences in individual sites for SNP calling, genotype calling, and estimation of allele frequencies. In particular, an Empirical Bayes (EB) method in which the ML estimates are used to make inferences for each individual site might have desirable properties. The posterior probability of the allele frequency in a particular site is given by

$$p(S_m=j|X)=\frac{p(X|S_m=j)p(S_m=j)}{\sum_{i=1}^{2k}p(X|S_m=i)p(S_m=i)},j=0,1,2,..,2k \quad (8)$$

as in [24] which using the algorithm from the previous section can be calculated as

$$p(S_m=j|X)=\frac{h_jp_j}{\sum_{r=0}^{2k}h_rp_j},j=0,1,2,..,2k. \quad (9)$$

A point estimate of the sample allele frequency can then be obtained as $arg\ max_j\{p(S_m=j \mid X)\}$. As we often will be interested in SNP calling and genotype calling in all sites, and not only in sites in which the ancestral base is among the segregating nucleotides, inferences can be done using the folded, rather than the unfolded, frequency spectrum. To calculate the posterior probability, we then need to sum over foldings, and over assignments of derived and ancestral alleles:

$$p(S_m=j|X)=\frac{h_j(p_j+p_{2k-j})+h_{2k-j}(p_j+p_{2k-j})}{\sum_{r=0}^{k-1}h_r(p_r+p_{2k-r})+2h_kp_k+\sum_{r=k+1}^{2k}h_r(p_r+p_{2k-r})},j=0,1,2,..,k-1$$

and

$$p(S_m=k|X)=\frac{2h_kp_k}{\sum_{r=0}^{k-1}h_r(p_r+p_{2k-r})+2h_kp_k+\sum_{r=k+1}^{2k}h_r(p_r+p_{2k-r})} \quad (10)$$

Finally, if we wish to take into account uncertainty in the assignments of ancestral and derived alleles, we need to sum over all possible pairs of segregating alleles:

$$p(S_m=j|X)=\frac{\sum_{(a,b)}\left(h_j^{(ab)}p_j+h_{2k-j}^{(ab)}p_{2k-j}\right)}{\sum_{(a,b)}\left(\sum_{r=0}^{k-1}h_r^{(ab)}p_r+h_k^{(ab)}p_k+\sum_{r=k+1}^{2k}h_r^{(ab)}p_r\right)},j=0,1,2,..,k-1$$

and

$$p(S_m=k)=\frac{\sum_{(a,b)}h_k^{(ab)}p_k}{\sum_{(a,b)}\left(\sum_{r=0}^{k-1}h_r^{(ab)}p_r+h_k^{(ab)}p_k+\sum_{r=k+1}^{2k}h_r^{(ab)}p_r\right)} \quad (11)$$

where $h_j^{(ab)}$is the function h_j calculated assuming a is derived and b is ancestral, and the sum is over all ordered pairs (a, b) $\in B^2$. There is here an implicit assumption of equal weighting of all possible alleles as ancestral and derived. The method could possibly be improved by using a more careful weighting using empirically derived proportions of segregating nucleotide pairs.

The expression given above can be used directly for SNP calling using a fixed cut-off for$p(S_m=0|X)$, such as $p(S_m=0|X)<0.05$ or some lower value depending on how conservative one wants to be in calling SNPs.

If SNP calling has already been performed based on the same data, so that only sites expected to be variable are included in the analysis, estimation of allele frequencies should proceed by conditioning on the site being variable in the sample, by modifying the denominator in the expression above to reflex that zero probability is assigned to the event $S_m=0$.or $S_m=2k$. For example, Equation (11) becomes

$$p(S_m=j|X)=\frac{\sum_{(a,b)}\left(h_j^{(ab)}p_j+h_{2k-j}^{(ab)}P_{2k-j}\right)}{\sum_{(a,b)}\left(\sum_{r=1}^{k-1}h_r^{(ab)}p_r+h_k^{(ab)}p_k+\sum_{r=k+1}^{2k-1}h_r^{(ab)}p_r\right)},j=1,2,...k-1 \text{ and}$$

$$p(S_m=k|X)=\frac{\sum_{(a,b)}h_k{}^{(ab)}p_k}{\sum_{(a,b)}\left(\sum_{r=1}^{k-1}h_r{}^{(ab)}p_r+h_k{}^{(ab)}p_k+\sum_{r=k+1}^{2k-1}h_r{}^{(ab)}p_r\right)} \quad (12)$$

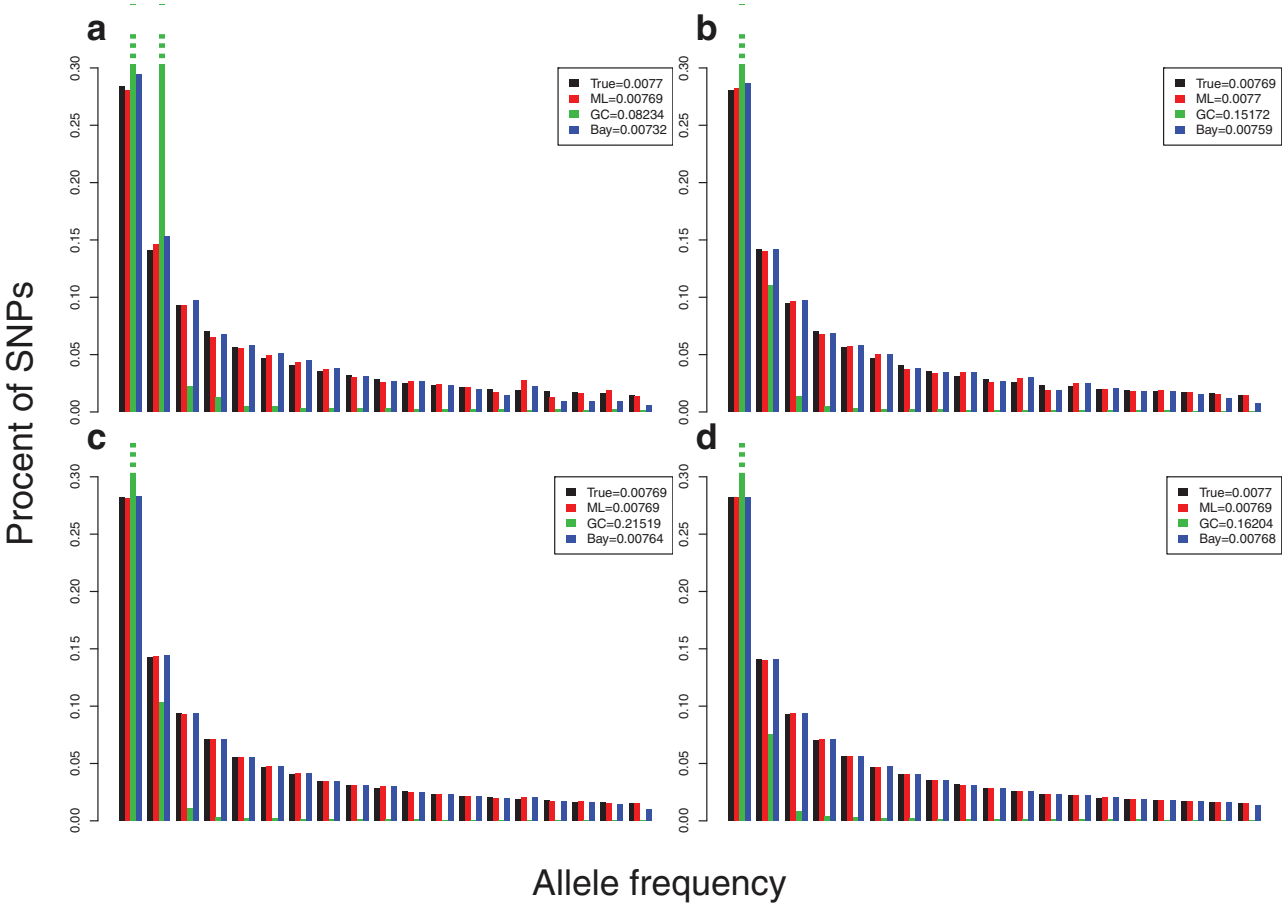

Figure 1. The distribution of true (True) and estimated unfolded SFS using the Maximum Likelihood method (ML) presented in the paper, genotype calling based on choosing the genotype with highest posterior probability (GC), and using the Bayesian procedure described in the text (Bay) in a sample from 50 MB 10 diploid individuals, where 2% of all SNPs are variable in the population and follow a distribution of allele frequencies, *p*, proportional to 1/*p*. An error rate of 0.5% is assumed. The mean sequencing depths are 1X (a), 3X (b), 5X (c), and 10X (d). The values presented in the figure legend box are the estimates of the proportion of sites that are variable in the sample.
doi:10.1371/journal.pone.0037558.g001

Genotype probabilities

The framework derived above for allele frequency estimation and SNP calling can also be used for estimating individual genotype probabilities, leveraging information from all other individuals in the genotype call for a single individual. We will assume that the site has already been called to be variable with a SNP of a specific type with nucleotides h and g.

The posterior probability for a genotype for an individual, d, then becomes

$$p(G_d(g,h)=j \text{ or } G_d(h,g)=2-j|X)$$

$$=\frac{p(X,G_d(g,h)=j)+p(X,G_d(h,g)=2-j)}{p(X)} \qquad (13)$$

$$=\frac{p(X_d|G_d(g,h)=j)\sum_{r=j}^{2k-2+j}\left[c_{r,j}p_r h_{r-j,d}^{*(g,h)}\right]+p(X_d|G_d(h,g)=2-j)\sum_{r=2-j}^{2k-j}\left[c_{r,2-j}h_{r-2+j,d}^{*(h,g)}p_r\right]}{\sum_{r=0}^{k2}\left[p_r\left(h_r^{(g,h)}+h_r^{(h,g)}\right)\right]}$$

and

$$c_{r,j}=\begin{cases}\dfrac{\binom{r}{j}\binom{2k-r}{2-j}}{\binom{2k}{2}} & \text{if } j\le r\\ 0 \text{ otherwise}\end{cases}$$

Here, the event $G_d(g,h)=j$ indicates that individual d has genotype $j\in\{0,1,2\}$, j indicating the number of derived allele, with g as the derived and h as the ancestral allele. $h_{r,d}^{*(g,h)}$ is the value of $h_r^{(g,h)}$ calculated for individuals (1, 2,..., d-1, d+1,.., k). This algorithm for estimation of genotype probabilities, therefore requires recalculation of the h_j functions for all k possible subsets found by excluding one individual from the data.

We notice that $p(X,G_d(g,h)=j)=p(X,G_d(h,g)=2-j)$. Furthermore, assuming symmetry in the probability of being ancestral and derived among nucleotides, $h_r^{(g,h)}=h_{2k-r}^{(h,g)}$ and $h_{r,d}^{*(g,h)}=h_{2k-2-r,d}^{*(h,g)}$, and, therefore the denominator can be calculated faster as

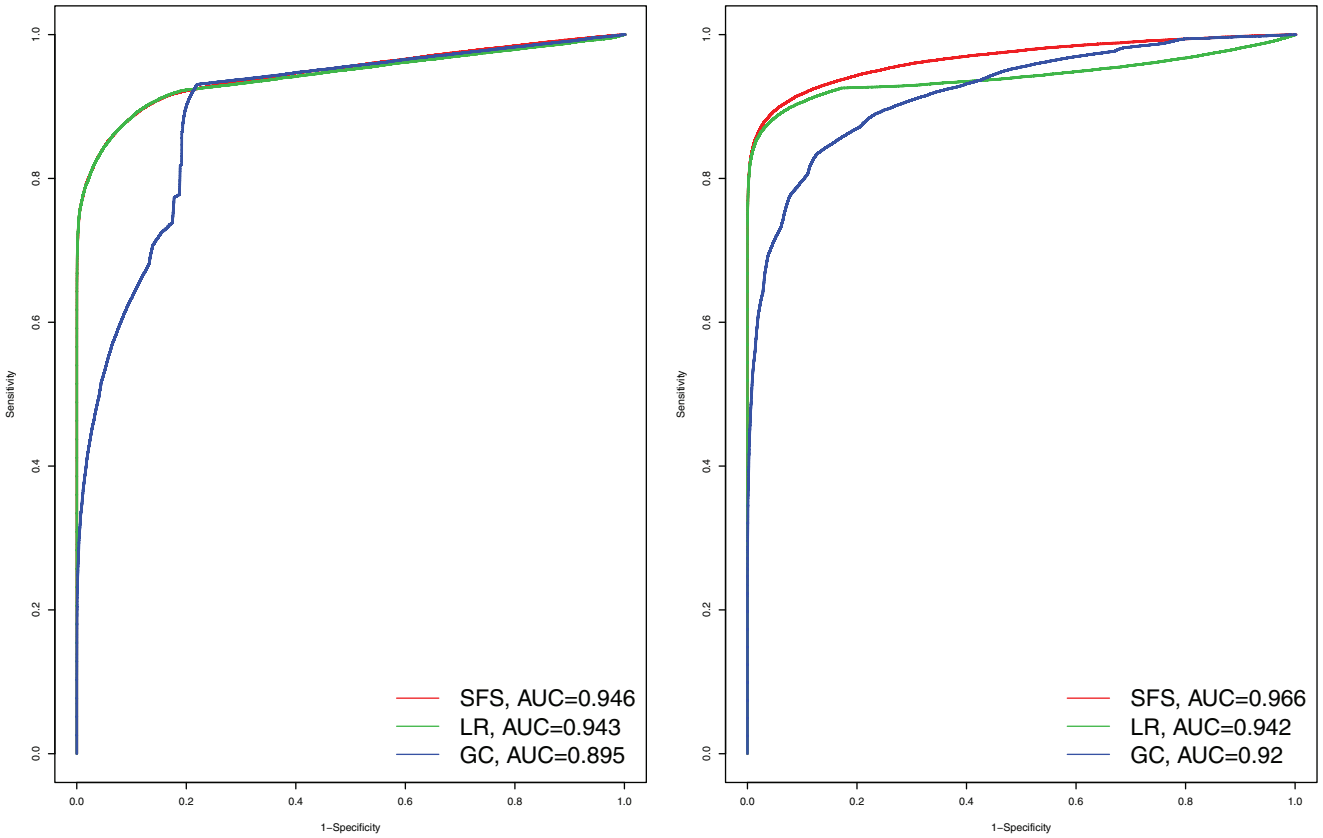

Figure 2. ROC curves for different SNP callers. Data for 10 individuals were simulated assuming a sequencing depth of 2 and a raw sequencing error rate of 1% (A) and (B) a depth of 5 and a raw sequencing error rate of 5%. The SFS method is the main method described in the text. The GC method is based on genotype calling using the genotype with the highest posterior probability. The LR method is based on a likelihood ratio test of the hypothesis that the allele frequency is zero. The SFS based method and the LR method have similar performance except for very high error rates, where the SFS tends to be somewhat better. Both methods in general perform much better than the GC method. The difference would even larger in larger panels of individuals. Simulations under other conditions can be found in Figure S1.
doi:10.1371/journal.pone.0037558.g002

$$2\left(\sum_{r=0}^{k-1}\left[h_r^{(g,h)}(p_r+p_{2k-r})\right]+h_k^{(g,h)}p_k\right) \tag{14}$$

Likewise, the numerator becomes

$$2p(X_d|G_d(g,h)=j)\left(\sum_{r=0}^{k-1}\left[c_{r,j}h_{r-j,d}^{*(g,h)}(p_r+p_{2k-r})\right]+c_{k,j}h_{k-j,d}^{*(g,h)}p_k\right) \tag{15}$$

In cases where SNP calling precedes genotype calling, the summations in the numerator and denominator should be modified to appropriately condition on variability.

Again, specific weighting schemes for the pairs (a, b) could possibly be used to improve the estimates. Finally, we note that these expressions assume Hardy-Weinberg equilibrium.

Incorporating external information regarding allele frequency

The algorithms described above have been developed assuming that no external information exists regarding allele frequencies. When that is not true, the algorithm can be modified to incorporate external estimates of the allele frequency.

Assume that we know the population allele frequency of the major allele, f, in the site. Then, assuming Hardy-Weinberg equilibrium, the marginal posterior for a particular genotype is

$$p(G_d=j|X_d)=\frac{p(X_d|G_d=j)p(G_d=j|f)}{\sum_{i=0}^{2}p(X_d|G_d=i)p(G_d=i|f)}, \tag{16}$$

where, assuming Hardy-Weinberg Equilibrium.

$p(G_d=0\,|f)=f^2, p(G_d=1\,|f)=2f(1-f), p(G_d=2\,|f)=(1-f)^2$.

The allele frequency, f, will typically be based on estimates obtained from a larger set of sites. We can consider this another type of Empirical Bayes (EB) procedure in that a parameter of the prior for each individual is estimated jointly based on all individuals (and possibly other external data). We will use the maximum likelihood estimator of population allele frequency (not to be confused with sample allele frequency) described by [22] in any data applications in this paper. This approach may not work well when there are only very few individuals for which to estimate f. In such cases, it might work better to obtain joint ML estimates of genotypes from all individuals using an EM algorithm with f as the latent variable, to use a full Bayesian approach integrating the joint likelihood function all individuals over f, or to revert to the previously discussed methods which does not rely on estimation of f. When k is relatively large (e.g.,

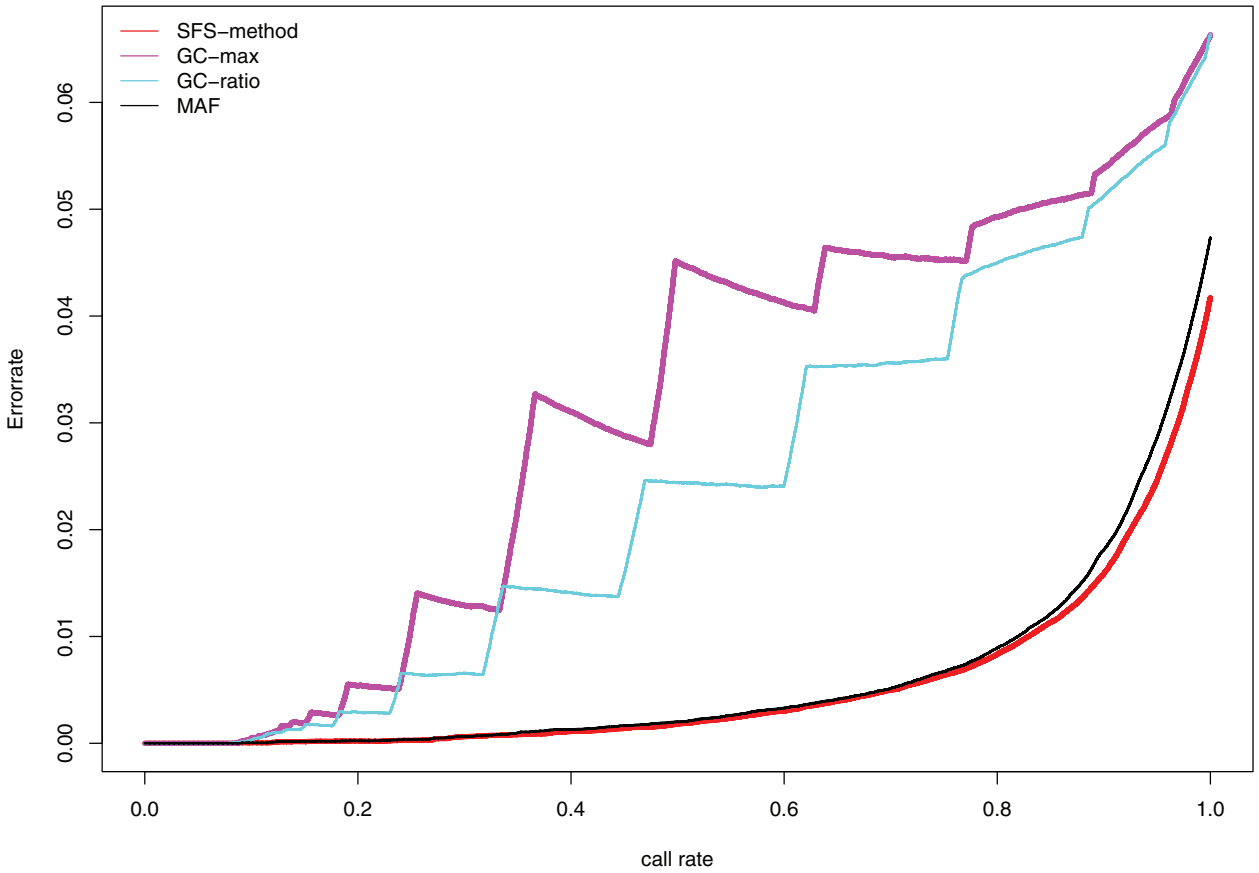

Figure 3. The error rate of different genotype callers for different call rates. The SFS-method is the method described in the main text. The MAF method is based on first obtaining a maximum likelihood estimate of the allele frequency, and then use the estimated allele frequency to define priors for genotype calling. The GC-max method is based on calling genotypes with highest posterior probability. The GC-ratio method is based on calling genotypes depending on the ratio of the likelihood for the most likely to second most likely genotype. The jagged behavior of some of the curves is a consequence of the discrete nature of the data, i.e. an individual contains a discrete number of copies of the minor allele. 10 individuals are simulated for 50,000 variable sites with a distribution of allele frequencies (p), proportional to $1/p$ with an error rate of 0.5%. Results for other error rates are shown in Figure S2.
doi:10.1371/journal.pone.0037558.g003

>20), the EB procedure should provide marginal posteriors for the genotypes from each individual close to the ones that would have been obtained using a full Bayesian approach.

Similarly, we will use an estimator of f obtained for each site independently, but jointly for all individuals: the maximum likelihood estimator described by [8,22]. For simulation purposes we will occasionally also use the estimator by [8], which is faster to calculate but may not be as accurate as the ML estimator. Determination of status as major or minor will be defined based on these estimates. Because of this we can also safely ignore the possibility that a site is invariable because the minor allele is fixed, and equate invariability to $0 < S_m < 2k$.

We are then interested in obtaining

$$p(S_m=j|X) = \frac{p(X|S_m=j)p(S_m=j|\mathrm{var})p_{\mathrm{var}}}{\sum_{i=1}^{2k-1} p(X|S_m=i)p(S_m=i|\mathrm{var})p_{\mathrm{var}} + (1-p_{\mathrm{var}})p(X|0<S_m<2k)}, j=1,2,..,2k-1 \quad (17)$$

and $X=(X_1,\ldots, X_k)$ now is the vector of read data for all individuals. The variable 'var' indicates the event that the site is a variant, i.e. $0 < S_m < 2k$. $p(S_m=j|X)$ can then be estimated, using the same algorithm as described for calculation of the likelihood function, but with the following *Termination* step

Termination

Set $h_j = h_j f^{2k-j}(1-f)^j \Big/ \left(1-f^{2k}-(1-f)^{2k}\right)$ for $j=1,2,\ldots,2k$-1.

The posterior probabilities are then given by

$$p(S_m=j|X) = \frac{h_j p_{\mathrm{var}}}{p_{\mathrm{var}} \sum_{r=1}^{2k} h_r + (1-p_{\mathrm{var}})(h_0+h_{2k})}, j=1,2,..2k-1, \quad (18)$$

After completion of the algorithm, status of major and minor allele might then appropriately be re-assigned if this is used in the downstream inferences and if $p(S_m > k \mid X) > 0.5$. Alternatively, the results can be polarized with respect to ancestral and derived allele or be folded. The allele frequency can then be estimated as the value of j that maximizes $p(S_m=j|X)$, or inferences can, in most cases, more appropriately be made by summing over the posterior distribution of S_m.

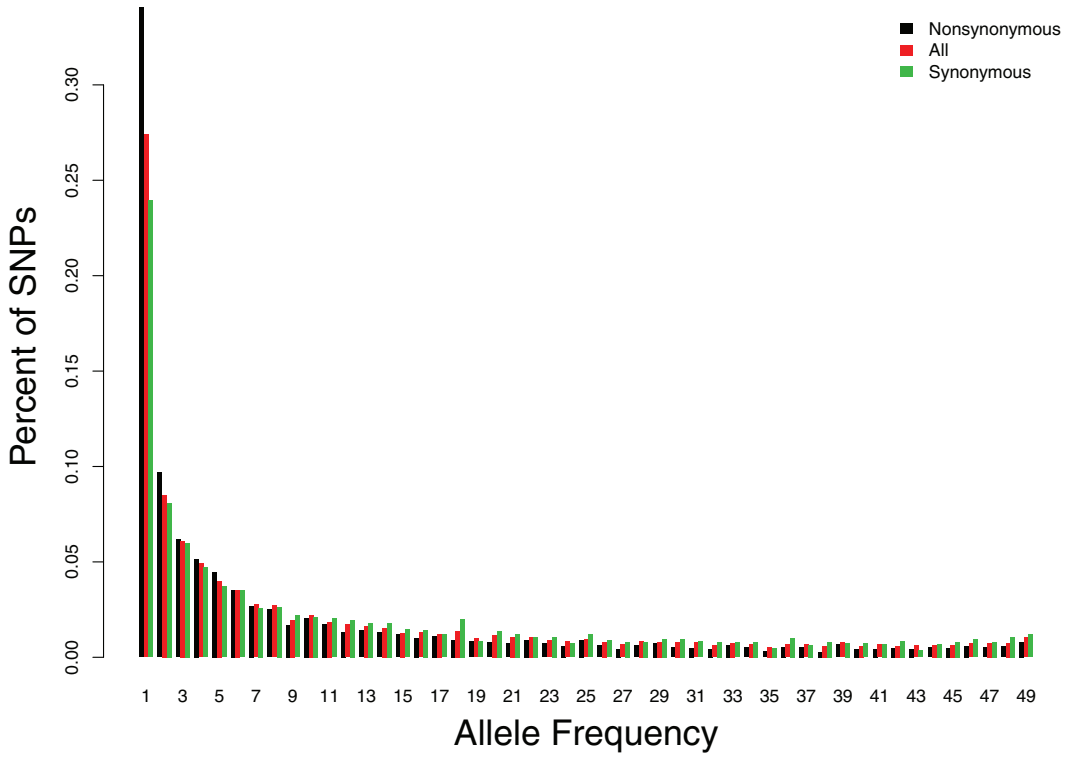

Figure 4. The unfolded site frequency spectrum from 25 Danish indivuduals. The data were previously analyzed in Yi *et al.* 2010.
doi:10.1371/journal.pone.0037558.g004

Incorporating deviations from Hardy-Weinberg Equilibrium (HWE)

The EB estimator of allele frequency can also be modified to incorporate deviations from HWE. Assume that an inbreeding coefficient, F_d, has been estimated for individual d, $d = 1, 2, \ldots, k$. F_d can take on both positive and negative values. Let $m_{d0} = f^2 + F_d f(1-f)$, $m_{d1} = (1-F_d)2f(1-f)$, and $m_{d2} = (1-f)^2 + F_d f(1-f)$. Then the following algorithm calculates the likelihood used in the EB estimation:

Initialization:

Set $h_0 = m_{10} p(X_1|G_1 = 0), h_1 = m_{11} p(X_1|G_1 = 1)$, $h_2 = m_{12} p(X_1|G_1 = 2)$, and $h_j = 0$ for $j = 3,4,\ldots,2k$.

Recursion

For $d = 2, 3,\ldots, k$:

For $j = 2d, 2d\text{-}1,\ldots,2$:

Set

$$h_j = m_{d2} p(X_d|G_d = 2) h_{j-2} + m_{d1} p(X_d|G_d = 1) h_{j-1} + m_{d0} p(X_d|G_d = 0) h_j$$

Set h_1 $m_{d0} p(X_d \mid G_d = 0) h_1 + m_{d1} p(X_d \mid G_d = 1) h_0$

Set $h_0 = m_{d0} p(X_d \mid G_d = 0) h_0$

Termination

Set $h_j = h_j / \left(1 - \prod_{d=1}^{k} m_{d0} - \prod_{d=1}^{k} m_{d2}\right)$ for $j = 1,2,\ldots,2k\text{-}1$.

The posterior probabilities can then be evaluated as before.

Simulations

To compare methods we conducted simulations under simplified assumptions. In all simulations, except if otherwise stated, we simulated data by allowing a Poisson distributed number of reads for each individual in each site independently of each other. The distribution of allele frequency (x) was assumed to be proportional to $1/x$ in the population. Each site is assumed to be variable with probability p_{var}. Errors are introduced randomly an symmetrically among all bases. Genotype probabilities are calculated according to the model assuming known error rates. We also compared methods by examining their performance on real data. This was done using HapMap data with known genotypes (the reported genotype error rate is $<0.1\%$).

Results

In the Methods section, we described a likelihood function for **P**, i.e. we derived $\Pr(X \mid \mathbf{P})$, where X is all the sequencing data from multiple individuals and multiple sites. This is the likelihood function underlying the EM algorithm for estimating the SFS presented in [24]. [25] also developed a similar method, but could only analyze small sample sizes due to computational constraints. As shown in the Methods section, the likelihood function can be evaluated directly, using a dynamic programming algorithm, with computational time that is linear in the number of sites, and quadratic in the number of individuals. The function can be optimized using standard optimization algorithms, using analytical derivatives, to provide a maximum likelihood estimate of the SFS. This method provides a computational alternative to the method of [24] for obtaining maximum likelihood estimates of the SFS.

To evaluate the method, we simulate data with known error rates (Fig. 1), mimicking the variation in sequencing depths observed in real data. The number of data points needed to provide good estimates depend both on the number of sites/SNPs analyzed, the number of individuals and on the sequencing depth. For example, 50 MB of data with 1% variable sites is sufficient to provide reasonable estimates even if the average sequencing depth

is only 1X per individual (0.5X per chromosome). However, with only 10 MB, higher depth is needed and good estimates are first obtained with a depth of 3–5X.

To illustrate the difference between the new method and methods based directly on genotype calling, we compared with the case where the most likely genotype is chosen, with and without filtering of genotypes with low confidence (Figure 1). Clearly, simple genotype calling leads to an excess of singletons when no filtering is done. This problem can be partly corrected by using more conservative SNP calling procedures. But even in such cases, the SFS estimates tend to be poor from low frequency data. The effect is very similar to the one described [15,16] in which they show that no simple cut-off method leads to unbiased estimates of the population genetic parameter θ ($=4N\mu$ where N is the population size and μ is the mutation rate) when using low or moderate coverage shotgun sequencing data. The same effect is observed for estimation of the SFS. Using filters in which only high confidence genotypes are called leads to new biases because it is easier to call homozygous than heterozygous individuals. This bias will affect different allele frequency categories differentially and lead to biases in the estimate of the SFS. [15,16] argue that methods for estimating θ should instead take the inherent uncertainty in the data into account. The method developed here is a conceptual extension of this concept to the SFS.

Inferences for individual sites

The method used for inferences of the SFS for a whole genome or for a large set of sites, can also be modified to make inferences for a single site [24]. The estimated SFS can be used as a prior for the allele frequency, and inferences regarding a particular site can then proceed using classical Bayesian procedures. The algorithmic details are provided in the Methods section. This method can be considered a Empirical Bayes method (e.g., [28]) as a large set of data points is being used to define a prior that subsequently can be applied to each data point. Figure 1 shows that, in average, the use of this procedure provides a distribution of allele frequencies that accurately reflects the true distribution of allele frequencies.

SNP calling

An algorithm similar to the one used for estimating the SFS can be used to make inferences for individual sites, and is described in the Methods section. The method proceeds by first estimating the SFS. The estimated distribution of allele frequencies, and the total frequency of SNPs in the sample ($1-G_0^*$), then provides priors in a Bayesian SNP caller similar to the one used in the 1000 Genomes project [7]. This is the approach outlined for SNP calling in [24]. We compare this type of SNP calling to two other methods: (1) traditional SNP calling based on observing at least X high quality reads of the minor allele, and (2) a likelihood ratio test based on testing the null hypothesis that the minor allele frequency is zero (Figure 2; Figure S1). The latter method is based on the likelihood function described in [11,22,23,29].

We see that the Bayesian SNP caller is substantially better than traditional methods, but does only marginally better than the likelihood ratio tests (Figure 2; Figure S1). The use of prior information regarding allele frequencies only provide a marginal improvement. However, in the analyses of human data, or other data where large reference data sets are available, optimal SNP callers will include prior information from the reference data, possibly using methods related to imputation (e.g., [30,31,32,33,34]) as in the 1000 Genomes project [7]. For such methods, an important initial step is calculation of posterior probabilities for each SNP. Depending on the specifics of the implementation of the imputation methods, the methods described here for estimating sample allele frequencies may also be useful in the application of some imputation methods.

Genotype calling

The Methods section described a Bayesian method for genotype calling using the estimated SFS as a prior. In brief, it calculates the posterior probability of the genotype in each individual conditional on all data from both the focal individual and the other individuals in the sample. Information from other individuals can substantially improve genotype calling. This is illustrated using simulations in Figure 3 (see also Figure S2). Notice that the genotype calling accuracy is greatly improved compared to the case of just choosing the most likely genotype.

Again, in human data, and other data for which large reference data sets are available, these data should be incorporated for genotype calling. In fact, imputation based genotype calling will lead to a substantial increase in accuracy over other methods [7].

Applications to data from 25 exomes

To illustrate the use of these methods on real data, we analyzed previously published data from 25 Danish exomes [6]. The resulting frequency spectrum is depicted in Figure 4. As in [6] we find that nonsynonymous mutations show an excess of rare alleles compared to synonymous mutations, presumably due to slightly and weakly deleterious alleles.

The Methods section also describes a method for incorporating prior information regarding allele frequencies and for incorporating deviations from Hardy-Weinberg Equilibrium when estimation allele frequencies.

Discussion

We have here developed a method for estimating the SFS that can be used for population genetic inferences. This method may also be used to define priors used in SNP calling and genotype calling leading to improved analyses of next-generation sequencing data.

The methods rely on accurate estimation of genotype likelihoods. Much research has been devoted to this (e.g. [10,11]), and there is some hope that reasonably accurate genotype likelihoods eventually can be calculated for most sequencing platforms. However, it is worth emphasizing that inaccurate genotype likelihoods can lead to false inferences when applied in the present context. In real data, it can often be difficult to determine if genotype likelihoods have been calculated correctly. However, the improvements observed over simpler method when applied to real data, suggests that genotype likelihoods, as calculated by, for example, the SOAPsnp program ([11]) used here, provides sufficiently accurate genotype likelihoods to make the application of the new methods worthwhile.

Several of the methods presented here are similar to methods developed in parallel and recently published by [24] Li (2011). In particular, [24] provided an EM algorithm for estimating the SFS under the same model and [25] developed a method applicable to smaller samples. Our approach differs from these approaches by the use of a dynamic programming algorithm that makes the likelihood function accessible to direct fast evaluation and numerical optimization. Similar dynamic programming algorithm has previously been used in [8] and [26] for single site inferences. In addition, we show how to use the resulting estimated SFS for genotype calling. Our genotype caller differs from previous genotype callers by explicitly calculating the posterior probability of a genotype conditional on the data obtained from all individuals in the sample under a joint prior for the sample allele frequency. [26] presented closely related genotype callers based on inferences on single sites,

also using a dynamic programming algorithm allowing calculation of joint allele frequencies. The SNP calling algorithm we use is identical to the one in [24]. We also present additional results on how to incorporate deviations from Hardy-Weinberg equilibrium when estimating allele frequencies, how to address issues regarding the folding of the frequency spectrum and how to incorporate external information regarding allele frequencies. In addition, we provide some simulation results evaluating the performance of the SNP callers, Genotype callers and SFS estimators.

A number of different methods have been proposed for estimating allele frequencies and the SFS from NGS data. In this paper we discuss the use of joint maximum likelihood estimates from multiple sites. This was also the approach taken by [24] and [25]. As illustrated in Figure 1, this approach will recover the true frequency spectrum when the modeling assumptions are correct. Methods based on estimating the allele frequency separately in each site will not generally have this property. [35] provided an alternative approach. The idea in this approach is to compare the inferred SFS based on genotype calling to the SFS obtained in other data that can be assumed not to have the types of biases introduced in NGS data. The extent of bias can then be quantified statistically, and used to correct SFS based on genotype calling in a larger data set. This approach may be preferable when the error structure is difficult to model, because it does not rely on such modeling. However, it requires the availability of accurate genotype calls from a large representative panel.

Supporting Information

Figure S1 ROC curves for different SNP callers. Data for 10 individuals were simulated for different depths and error rates (d indicates depth and e is the reror rate). The SFS method is the main method described in the text. The GC method is based on genotype calling using the genotype with the highest posterior probability. The LR method is based on a likelihood ratio test of the hypothesis that the allele frequency is zero. larger panels of individuals.
(DOC)

Figure S2 The error rate of different genotype callers for different call rates. The SFS-method is the method described in the main text. The MAF method is based on first obtaining a maximum likelihood estimate of the allele frequency, and then use the estimated allele frequency to define priors for genotype calling. The GC-max method is based on calling genotypes with highest posterior probability. The GC-ratio method is based on calling genotypes depending on the ratio of the likelihood for the most likely to second most likely genotype. The jagged behavior of some of the curves is a consequence of the discrete nature of the data, i.e. an individual contains a discrete number of copies of the minor allele. 10 individuals are simulated for 50,000 variable sites with a distribution of allele frequencies (p), proportional to $1/p$ and with a varying error rate.
(DOC)

Author Contributions

Conceived and designed the experiments: RN TK AA YL JW. Performed the experiments: RN TK AA. Analyzed the data: RN TK AA. Contributed reagents/materials/analysis tools: YL WJ. Wrote the paper: RN.

References

1. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nature Methods 5: 16–18.
2. Li RQ, Fan W, Tian G, Zhu HM, He L, et al. (2010) The sequence and de novo assembly of the giant panda genome. Nature 463: 311–317.
3. Tucker T, Marra M, Friedman JM (2009) Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. American Journal of Human Genetics 85: 142–154.
4. Voelkerding KV, Dames SA, Durtschi JD (2009) Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry 55: 641–658.
5. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.
6. Li YR, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nature Genetics 42: 969–U982.
7. Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
8. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, et al. (2010) Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science 329: 75–78.
9. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, et al. (2009) Benchmarking Next-Generation Transcriptome Sequencing for Functional and Evolutionary Genomics. Molecular Biology and Evolution 26: 2731–2744.
10. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18: 1851–1858.
11. Li RQ, Li YR, Fang XD, Yang HM, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. Genome Research 19: 1124–1132.
12. Harismendy O, Ng PC, Strausberg RL, Wang XY, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology 10.
13. Hedges D, Burges D, Powell E, Almonte C, Huang J, et al. (2009) Exome Sequencing of a Multigenerational Human Pedigree. Plos One 4.
14. Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, et al. (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res 18: 1020–1029.
15. Johnson PLF, Slatkin M (2006) Inference of population genetic parameters in metagenomics: A clean look at messy data. Genome Research 16: 1320–1327.
16. Johnson PLF, Slatkin M (2008) Accounting for bias from sequencing error in population genetic estimates. Molecular Biology and Evolution 25: 199–206.
17. Lynch M (2008) Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects. Molecular Biology and Evolution 25: 2409–2419.
18. Watterson GA (1975) On the number of segregation sites. Theor Pop Biol 7: 256–276.
19. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
20. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. Plos Genetics 5.
21. Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38: 1358–1370.
22. Kim SY, Li YR, Guo YR, Li RQ, Holmkvist J, et al. (2010) Design of Association Studies with Pooled or Un-pooled Next-Generation Sequencing Data. Genetic Epidemiology 34: 479–491.
23. Kim SY, Lohmueller KE, Albrechtsen A, Li YR, Korneliussen T, et al. (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. Bmc Bioinformatics 12.
24. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.
25. 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–U295.
26. Le SQ, Durbin R (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Research 21: 952–960.
27. Press WH (2000) Numerical recipes in C; the art of scientific computing. 2nd ed. Cambridge; New York: Cambridge University Press.
28. Casella G (1985) An Introduction to Empirical Bayes Data-Analysis. American Statistician 39: 83–87.
29. Martin ER, Kinnamon DD, Schmidt MA, Powell EH, Zuchner S, et al. (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 26: 2803–2810.
30. Dai JY, Ruczinski I, LeBlanc M, Kooperberg C (2006) Imputation methods to improve inference in SNP association studies. Genetic Epidemiology 30: 690–702.
31. Minichiello MJ, Durbin R (2006) Mapping trait loci by use of inferred ancestral recombination graphs. American Journal of Human Genetics 79: 910–922.

32. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81: 1084–1097.
33. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39: 906–913.
34. Howie BN, Donnelly P, Marchini J (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Plos Genetics 5.
35. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, et al. (2011) Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America 108: 11983–11988.