

Genotype-Frequency Estimation from High-Throughput Sequencing Data

Takahiro Maruki¹ and Michael Lynch

Department of Biology, Indiana University, Bloomington, Indiana 47405

ABSTRACT Rapidly improving high-throughput sequencing technologies provide unprecedented opportunities for carrying out population-genomic studies with various organisms. To take full advantage of these methods, it is essential to correctly estimate allele and genotype frequencies, and here we present a maximum-likelihood method that accomplishes these tasks. The proposed method fully accounts for uncertainties resulting from sequencing errors and biparental chromosome sampling and yields essentially unbiased estimates with minimal sampling variances with moderately high depths of coverage regardless of a mating system and structure of the population. Moreover, we have developed statistical tests for examining the significance of polymorphisms and their genotypic deviations from Hardy–Weinberg equilibrium. We examine the performance of the proposed method by computer simulations and apply it to low-coverage human data generated by high-throughput sequencing. The results show that the proposed method improves our ability to carry out population-genomic analyses in important ways. The software package of the proposed method is freely available from <https://github.com/Takahiro-Maruki/Package-GFE>.

KEYWORDS genotype frequency; Hardy–Weinberg test; inbreeding coefficient; polymorphism detection; population genomics

THE estimation of allele and genotype frequencies is fundamental in population-genetic studies. Most evolutionary inferences in population genetics, including those concerned with population demography and natural selection, start with this sort of information. When we study the relationship between genotypes and phenotypes in a population, proper inferences on genotype frequencies are also essential. Therefore, it is crucial to correctly estimate allele and genotype frequencies in population-genetic studies.

High-throughput sequencing technologies enable the extension of population-genomic analyses to a wide variety of organisms, which will improve our ability to draw evolutionary inference in several ways. First, rapidly declining sequencing costs enable researchers to sequence many individuals in a population. One of the major findings of recent population-genomic studies is the discovery of polymorphic sites harboring rare alleles and their importance in genotype–phenotype relationships. For example, Nelson *et al.* (2012) sequenced

202 drug-target genes in a sample of 14,002 human individuals and found that rare alleles likely associated with disease are abundant. Second, genome-wide analyses of polymorphisms provide a basis for much more accurate inferences of natural selection and population demography than can be obtained with more limited data (Luikart *et al.* 2003). Finally, because most models for inferring population demography assume neutral evolution, exclusion of sites that are targets of natural selection is desirable, and this is substantially easier to accomplish with whole-genome sequence data.

Despite the promise, high-throughput sequencing technologies have two disadvantages: high sequence error rates, which typically range from 0.001 to 0.01 with commonly used sequencing platforms (Glenn 2011; Quail *et al.* 2012); and genotypic uncertainties resulting from random chromosome sequencing. Because sequencing occurs randomly among sites, individuals, and chromosomes within diploid individuals, depths of coverage vary among them, and this can introduce biases in subsequent population-genetic analyses unless accounted for in a proper statistical framework (Pool *et al.* 2010).

To overcome the above difficulties in estimating allele and genotype frequencies from high-throughput sequencing data, several statistical methods have been recently developed

Copyright © 2015 by the Genetics Society of America

doi: 10.1534/genetics.115.179077

Manuscript received June 5, 2015; accepted for publication July 26, 2015; published Early Online July 29, 2015.

Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179077/-/DC1.

¹Corresponding author: Department of Biology, Indiana University, Bloomington, IN 47405. E-mail: tmaruki@indiana.edu

(Hellmann *et al.* 2008; Johnson and Slatkin 2008; Jiang *et al.* 2009; Li *et al.* 2009b; Lynch 2009; Hohenlohe *et al.* 2010; McKenna *et al.* 2010; DePristo *et al.* 2011; Kim *et al.* 2011; Keightley and Halligan 2011; Li 2011; Le and Durbin 2011; Nielsen *et al.* 2012; Vieira *et al.* 2013; Lynch *et al.* 2014). Among these, Lynch (2009) developed a maximum-likelihood (ML) method for estimating site-specific allele frequencies and error rates from high-throughput sequencing data, assuming Hardy–Weinberg equilibrium (HWE).

In this study, we generalize the ML allele-frequency estimator to allow application to diploid organisms with an arbitrary mating system and population structure. Specifically, we relax the assumption of HWE by adding to the model parameters the site-specific disequilibrium coefficient (Weir 1996), which measures the deviation of genotype frequencies from their HWE expectations. Furthermore, we develop statistical tests for the significance of polymorphisms and deviations from HWE. Examination of the performance of the proposed method with computer simulations reveals the generation of essentially unbiased estimates of allele and genotype frequencies with moderately high depths of coverage under general genetic conditions. As an example application to empirical data, we apply the proposed method to low-coverage, high-throughput sequencing data of 81 individuals from the CEU (Utah residents with ancestry from Northern and Western Europe) population generated in the human 1000 Genomes Project (1000 Genomes Project Consortium *et al.* 2012).

Methods

In the following, a maximum-likelihood (ML) method is developed for estimating the genotype frequencies at a site, using high-throughput sequencing data from a diploid population, which need not be in Hardy–Weinberg equilibrium. This is achieved by estimating three parameters from a population sample of observed site-specific sequence read data: the major-allele frequency p , disequilibrium coefficient D_A , and the error rate per read per site ϵ . The proposed method assumes that the attribution of sequence reads to specific individuals is known, *e.g.*, by uniquely tagging the DNA from each individual prior to sequencing. The two most abundant nucleotides in the population sample are considered to be candidates for alleles at the site.

For each individual i , the log-likelihood of the observed set of site-specific sequence reads is

$$\ln L_i = \ln \left[\sum_{g=1}^3 \gamma_g P_g(n_{iM}, n_{im}, n_{ie_1}, n_{ie_2}) \right], \quad (1)$$

where γ_g denotes genotype frequencies, with $\gamma_1 = p^2 + D_A$, $\gamma_2 = 2\{p(1-p) - D_A\}$, and $\gamma_3 = (1-p)^2 + D_A$, and n_{iM} , n_{im} , n_{ie_1} , n_{ie_2} are the observed number of reads of the most abundant nucleotide (major allele) M (*e.g.*, C), the second-most

abundant nucleotide (minor allele) m (*e.g.*, T), and other nucleotides e_1 and e_2 (*e.g.*, in this case A and G), respectively. $P_g(n_{iM}, n_{im}, n_{ie_1}, n_{ie_2})$ is the probability of the specific observed set of nucleotide reads given genotype g . Given a total depth of sequence coverage at the site in individual i , $n_{iT} = n_{iM} + n_{im} + n_{ie_1} + n_{ie_2}$, $P_g(n_{iM}, n_{im}, n_{ie_1}, n_{ie_2})$ is calculated using the following formula for the multinomial distribution,

$$P_g(n_{iM}, n_{im}, n_{ie_1}, n_{ie_2}) = \frac{n_{iT}!}{n_{iM}! n_{im}! n_{ie_1}! n_{ie_2}!} P_g(M)^{n_{iM}} P_g(m)^{n_{im}} P_g(e_1)^{n_{ie_1}} P_g(e_2)^{n_{ie_2}}, \quad (2)$$

where $P_g(M)$ is the probability of observed nucleotide read M with genotype g . $P_g(M)$ is a function of ϵ and is obtained by summing conditional probabilities of observed nucleotide read M , given the true nucleotide on the sequenced chromosome chosen from the pair (Table 1). For example, when $g = 2$ (Mm), the probability of nucleotide read M is $P_2(M) = (1/2)(1 - \epsilon) + (1/2)(\epsilon/3)$, the second term arising because we assume a random distribution of sequencing error types. In practice, because the multinomial coefficient in Equation 2 is constant regardless of the values of the parameters to be estimated, for computational efficiency, we reduce the preceding expression to

$$P_g(n_{iM}, n_{im}, n_{ie_1}, n_{ie_2}) = P_g(M)^{n_{iM}} P_g(m)^{n_{im}} P_g(e_1)^{n_{ie_1}} P_g(e_2)^{n_{ie_2}}. \quad (3)$$

The ML estimates of the major-allele frequency, disequilibrium coefficient, and error rate are found by maximizing the log-likelihood of the observed site-specific reads in the entire population sample, which is calculated by summing the log-likelihood (Equation 1) over N individuals:

$$\ln L = \sum_{i=1}^N \ln L_i. \quad (4)$$

To accurately and rapidly estimate the parameters, we take two steps. We first analytically estimate the allele frequencies and error rate and then conditioned on these approximations estimate the disequilibrium coefficient by a grid search. In the second step, we refine the preliminary ML estimates, using Equation 4. The preliminary estimates of the major-allele frequency and error rate are

$$\hat{p} = \frac{2n_M - n_e}{2(n_M + n_m - n_e)}, \quad (5)$$

$$\hat{\epsilon} = \frac{3}{2} \cdot \frac{n_e}{n_T}, \quad (6)$$

(Appendix), where n_M , n_m , n_e , and n_T are the read counts of the candidate major allele, minor allele, other nucleotides, and their sum, respectively, in the population sample.

Table 1 Probability of an observed nucleotide read as a function of the individual genotype and error rate ϵ

Genotype	Nucleotide on a sequence read			
	M	M	e_1	e_2
1 (MM)	$1 - \epsilon$	$\frac{\epsilon}{3}$	$\frac{\epsilon}{3}$	$\frac{\epsilon}{3}$
2 (Mm)	$\frac{1}{2} \cdot (1 - \epsilon) + \frac{1}{2} \cdot \frac{\epsilon}{3}$	$\frac{1}{2} \cdot \frac{\epsilon}{3} + \frac{1}{2} \cdot (1 - \epsilon)$	$\frac{\epsilon}{3}$	$\frac{\epsilon}{3}$
3 (mm)	$\frac{\epsilon}{3}$	$1 - \epsilon$	$\frac{\epsilon}{3}$	$\frac{\epsilon}{3}$

M and m denote candidate alleles (the two most abundant nucleotide reads in the population sample, e.g., C and T) and e_1 and e_2 denote other nucleotide reads (e.g., in this case A and G).

A preliminary estimate of the disequilibrium coefficient $\hat{D}_A$ is obtained by substituting $\hat{p}$ and $\hat{\epsilon}$ into Equation 4 and finding the value of D_A that maximizes the likelihood of the observed data. The minimum and maximum possible values of D_A , $D_{A_{\min}}$ and $D_{A_{\max}}$, are functions of the allele-frequency estimates and can be derived by noting that all of the three genotype frequencies need to be between zero and one,

$$D_{A_{\min}} = \max[-\hat{p}^2, -(1-\hat{p})^2] \quad (7)$$

$$D_{A_{\max}} = \hat{p}(1-\hat{p}) \quad (8)$$

(Weir 1996). Because the maximization is now reduced to a one-dimensional problem, $\hat{D}_A$ is rapidly found by a search over the span of possible D_A with interval size $1/N$.

Although the preliminary genotype-frequency estimates are reasonable for many cases, in situations involving small numbers of individuals and/or high variance in coverage among individuals, unnecessarily high estimation variance can arise. To overcome this limitation, we iteratively adjust the parameter estimates by examining whether the set of parameter estimates yields a local maximum in the likelihood surface, iterating by a localized grid search. Specifically, for each current estimate of major- and minor-homozygote frequency estimates $\hat{\gamma}_1$ and $\hat{\gamma}_3$, we evaluate the likelihood of all adjacent pairs of estimates deviating by $1/N$, where N is the sample size. This procedure is repeated until no deviations from the current ML estimate yield further increases in the likelihood.

If the final ML estimate of the major-allele frequency $\hat{p}$ is less than one, the ML disequilibrium coefficient estimate $\hat{D}_A$ and inbreeding coefficient estimate $\hat{f}$ are calculated as

$$\hat{D}_A = \hat{\gamma}_1 - \hat{p}^2 \quad (9)$$

$$\hat{f} = \frac{2 \cdot \hat{p}(1-\hat{p}) - \hat{\gamma}_2}{2 \cdot \hat{p}(1-\hat{p})}, \quad (10)$$

where $\hat{\gamma}_2$ is the ML estimate of the heterozygote frequency.

Statistical test of candidate polymorphisms

To avoid false-positive polymorphisms, we statistically test the significance of candidate polymorphisms by a likelihood-ratio test (Kendall and Stuart 1979). Letting LL_p and LL_m denote the maximum log-likelihood of the observed site-specific data under the assumptions of polymorphism and monomorphism, respectively, the likelihood-ratio test statistic LRT_p is

$$LRT_p = 2(LL_p - LL_m), \quad (11)$$

where LL_p is the maximum value of the log-likelihood given by Equation 4. LL_m is calculated as

$$LL_m = \sum_{i=1}^N \ln L_{im}, \quad (12)$$

where the log-likelihood of the observed data for individual i under the assumption that the population is fixed for the major allele is

$$L_{im} = (1 - \hat{\epsilon}_m)^{n_{im}} \left(\frac{\hat{\epsilon}_m}{3} \right)^{n_{iT} - n_{im}}, \quad (13)$$

where $\hat{\epsilon}_m$ is the ML estimate of the error rate, and the multinomial coefficient is again ignored. $\hat{\epsilon}_m$ is analytically found by taking the derivative of the likelihood function with respect to ϵ_m and setting it equal to zero, yielding

$$\hat{\epsilon}_m = \frac{n_T - n_M}{n_T}, \quad (14)$$

where n_T and n_M are the total number of nucleotide reads and number of the most abundant nucleotide read, respectively, in the population sample. LRT_p is expected to be asymptotically χ^2 distributed with 2 d.f.

Statistical test of Hardy-Weinberg equilibrium deviation

When a site is considered to be polymorphic by the preceding test, the deviation from HWE can also be statistically evaluated with a likelihood-ratio test, in this case using

$$LRT_{HWE} = 2(LL - LL_{HWE}), \quad (15)$$

where LL is the maximum log-likelihood of the observed site-specific data under the full model, calculated using Equation 4. LL_{HWE} is the corresponding maximum log-likelihood of the observed site-specific data assuming HWE, calculated by substituting $\hat{p}$ and $\hat{\epsilon}$ into Equation 4 with $\hat{D}_A$ set to zero. LRT_{HWE} is expected to be asymptotically χ^2 distributed with 1 d.f.

Sampling variance of the genotype-frequency estimates

If genotypes could be inferred from all individuals without error, the sampling variances of the ML estimates of the major-allele frequency $\hat{p}$, major-homozygote frequency $\hat{\gamma}_1$, and minor-homozygote frequency $\hat{\gamma}_3$ would be

$$\text{Var}(\hat{p}) = \frac{1}{2N} (p + \gamma_1 - 2p^2) \quad (16)$$

$$\text{Var}(\hat{\gamma}_1) = \frac{1}{N} \gamma_1 (1 - \gamma_1) \quad (17)$$

$$\text{Var}(\hat{\gamma}_3) = \frac{1}{N} \gamma_3 (1 - \gamma_3) \quad (18)$$

(Weir 1996). The sampling variances of the ML estimates by the proposed method are expected to asymptotically approach these values with high depths of coverage.

However, in high-throughput sequencing data, depths of coverage vary among sites, individuals, and chromosomes within individuals. Therefore, Equations 16–18 need to be modified. We do so by assuming that the depth of coverage at each site per individual is Poisson distributed with mean μ and considering the effective numbers of sampled chromosomes and individuals (Maruki and Lynch 2014), defined here at a single locus, and substituting them for $2N$ and N in Equations 16–18. The effective number of sampled chromosomes N_c is equivalent to the expected number of sampled chromosomes covered by at least one sequence read, and that of sampled individuals N_i is equivalent to the expected number of sampled individuals for which both alleles are covered by at least one sequence read. These are given by

$$N_c = 2N(1 - e^{-\mu/2}) \quad (19)$$

$$N_i = N \left[1 - (2e^{\mu/2} - 1)e^{-\mu} \right]. \quad (20)$$

We substitute N_c for $2N$ in Equation 16 and N_i for N in Equations 17 and 18 to obtain the expected asymptotic sampling-variance formulas.

Generation of high-throughput sequencing data by computer simulations

To examine the performance of the proposed genotype-frequency estimator, we generated high-throughput sequencing data for N diploid individuals by computer simulations and applied the proposed method to the simulated data. In the simulations, the probability of sampling an individual with a particular genotype was equal to its relative frequency. The frequencies of major and minor homozygotes were specified by γ_1 and γ_3 , respectively. The depths of coverage were assumed to be Poisson distributed with mean μ among the individuals and were specified as

$$c(X, \mu) = \frac{(\mu)^X e^{-\mu}}{X!}, \quad (21)$$

where X is a particular value of the coverage for an individual and c is a probability mass function of X . The sequences from each individual were randomly chosen from the pair of alleles. Sequence errors were randomly introduced at rate ε from the true nucleotide to one of the other three nucleotides.

A C++ program for simulated data analysis (Supporting Information, File S1) and its README (File S2) are available.

Application to empirical data

To further examine the performance of the proposed method when applied to actual data, we analyzed low-coverage (mean $4 \times$) phase I data of the 1000 Genomes Project Consortium (1000 Genomes Project Consortium *et al.* 2012). Specifically, we analyzed chromosome 6 data of individuals from the CEU population. We downloaded BAM files of the Illumina-sequencing read data from the Web site <ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/data/>. There were a total of 81 such files (File S3). We also downloaded the corresponding reference genome (GRCh37) used for mapping the sequence-read data from the Web site <ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/>.

Then, we generated mpileup files of the position-based sequence data, using SAMtools (Li *et al.* 2009a). The quartets of nucleotide read counts at each position necessary for the analyses were generated from the mpileup files, using sam2pro (http://guanine.evolbio.mpg.de/mlRho/sam2pro_0.6.tgz). To minimize mismapping of sequence reads, we excluded sites with the total depth of coverage in the population sample (sum of the coverage over the individuals) greater than twice the mean before subsequent analyses. Furthermore, we excluded sites involved in putative repetitive sequences predicted by RepeatMasker (<http://www.repeatmasker.org/>) before subsequent analyses. We used the repeat-masked GRCh37 reference genome downloaded from the Ensembl Web site, ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/ for this purpose. To avoid estimating the parameters from only a few individuals, we required that the total depth of coverage in the population sample be at least 81.

To examine the spatial patterns of polymorphisms characterized by the proposed method, we carried out sliding-window analyses of the per-site heterozygosity estimates $\hat{h}$ [using $2\hat{p}(1 - \hat{p})$, where $\hat{p}$ is the major-allele frequency estimate] and inbreeding coefficient estimates $\hat{f}$. To describe fine spatial patterns of polymorphisms, we calculated the weighted mean of $\hat{h}$ or $\hat{f}$ in each window (Hohenlohe *et al.* 2010). The sliding windows have a center position x (in base pairs) and width $6w$ (in base pairs) and the center position moves by a step of size s (in base pairs) ($x = i \cdot s$, $i = 1, 2, \dots$). The weight was given by $\exp[-(y-x)^2/2w^2]$, where y is a position of each site (in base pairs). In addition, to examine the potential confounding effect of misassembly and mismapping on the spatial patterns of polymorphisms, we also calculated the weighted means of the depths of coverage and error rate estimates in the same way. We used $s = 100,000$ and $w = 150,000$ in the analyses. To examine the annotations in regions of interest, we downloaded a GTF file on GRCh37 from the Ensembl Web site, ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/.

Comparison with other recently developed methods

To compare the performance of the proposed method with that of other recently developed methods, we analyzed the chromosome 6 data of 81 CEU individuals, using ANGSD (Korneliussen *et al.* 2014). To make fair comparisons, we made a list of sites analyzed by the proposed method and supplied it to ANGSD so that both approaches were applied to the same set of sites.

We estimated the allele frequencies from the BAM files of the 81 individuals, using the method by Kim *et al.* (2011) and Samtools (Li 2011) or GATK (McKenna *et al.* 2010) genotype likelihoods, and compared their allele-frequency estimates to those by the proposed method.

We estimated the folded site-frequency spectrum from the BAM files of the 81 individuals, using the method by Nielsen *et al.* (2012), and compared this to the results using the proposed method. We calculated the site-allele-frequency likelihood, assuming Hardy–Weinberg equilibrium, using Samtools or GATK genotype likelihoods.

We estimated per-site inbreeding coefficients from the BAM files of 81 individuals, using the method by Vieira *et al.* (2013) and Samtools genotype likelihoods. We examined the statistical significance of polymorphisms by the likelihood-ratio test described in Kim *et al.* (2011). Using per-site inbreeding coefficient estimates conditioned on significant polymorphisms at the 5% level, we carried out their sliding-window analysis in the way described above and compared the results with those using the proposed method.

Data availability

The software package of the proposed method is available from <https://github.com/Takahiro-Maruki/Package-GFE>. File S1 contains a C++ program for simulated data analysis. File S2 contains the README of the program. File S3 contains names of the analyzed BAM files of Illumina-sequencing read data of 81 individuals from the CEU population, which are available from the Web site <ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/data/>.

Results

The performance of the proposed genotype-frequency estimator was examined using computer simulations described above. To examine the results for the worst situations, we used 0.01 as the error rate, which is typically the upper bound in commonly used sequencing platforms. We evaluated the behavior of the proposed method under HWE and two extreme conditions, where the inbreeding coefficient f is minimized or maximized given a minor-allele frequency q . Specifically, the frequencies of major homozygotes, heterozygotes, and minor homozygotes are $1 - 2q$, $2q$, and 0, respectively, when f is minimized. When f is maximized, the frequencies of major homozygotes, heterozygotes, and minor homozygotes are $1 - q$, 0, and q , respectively. The means of the ML estimates of the allele frequencies were essentially

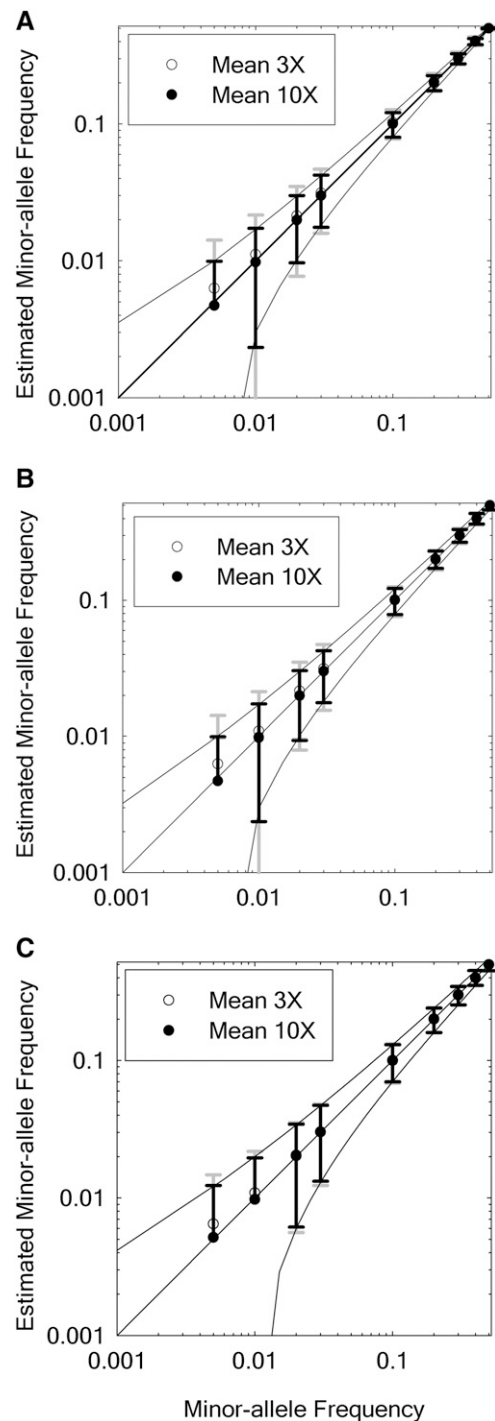


Figure 1 ML estimate of the minor-allele frequency as a function of its true value. The inbreeding coefficient f is (A) minimized, (B) equal to zero (Hardy–Weinberg equilibrium), and (C) maximized, given a minor-allele frequency (MAF). The mean and standard deviation of the estimated MAF are shown by the symbols and bars (shaded for mean $3\times$ and solid for mean $10\times$), respectively. The diagonal line and the curves surrounding the line represent the ideal situation where the ML estimate is equal to the true value and theoretical asymptotic sampling standard deviation (calculated as the square root of Equation 16), respectively. Number of sampled individuals $N = 100$, error rate $\varepsilon = 0.01$. A total of 10,000 simulation replications were run for each set of parameter values.

unbiased under all three examined conditions, although frequencies of rare alleles were slightly overestimated with low depths of coverage, due to sequence errors (Figure 1). The sampling standard deviations of the estimates became lower and approached the theoretical minimum values with higher depths of coverage under all conditions.

The means of the ML estimates of the inbreeding coefficient were somewhat biased when the mean depth of coverage μ was low and allele frequencies were extreme ($q < 0.1$) (Figure 2). Specifically, they were upwardly biased when f was minimized or equal to zero, due to sequencing just one of the two alleles from heterozygous individuals. On the other hand, they were downwardly biased when f was maximized, due to sequence errors. However, they were nearly unbiased when $q \geq 0.1$ even when μ was low (3), and when μ was moderately high (10), they were essentially unbiased under all of the examined conditions. We note that the biases in the inbreeding coefficient estimates with low depths of coverage are relatively large because they are ratios measuring relative deviation of genotype-frequency estimates from their HWE expectations. In fact, the corresponding biases in the disequilibrium coefficient estimates are small (Figure S1).

The means of the ML estimates of major-homozygote frequencies were essentially unbiased even with low depths of coverage (Figure S2, A, C, and E). On the other hand, the means of the ML estimates of minor-homozygote frequencies were slightly biased when μ was low (3) (Figure S2, B, D, and F). Specifically, they were slightly overestimated when f was minimized or equal to zero. When f was maximized, they were slightly underestimated. These biases in minor-homozygote frequency estimates can be understood by considering the interplay among the parameter estimates (Table 2). In particular, they are closely related to the biases in the inbreeding coefficient estimates. The sampling standard deviation of the ML estimates of genotype frequencies became lower and approached the theoretical minimum values with higher depths of coverage under all of the examined conditions.

We examined the power of the proposed method for detecting true polymorphisms as a function of the mean depth of coverage μ , rejecting the null hypothesis of monomorphism at a site at the 5% significance level (when LRT_p in Equation 11 was >5.991). Overall, the false-positive rate of polymorphism detection was low and decreased with higher μ (Figure 3A).

Because the statistical power of polymorphism detection should be dependent on the minor-allele frequency q at a site, we also examined the false-negative rate of polymorphism detection as a function of q (Figure 3, B–D). When q was low, the false-negative rate was high, especially with low depths of coverage. We note that this is because of the inherent limitations resulting from sampling only one of the alleles and not because of faults of the proposed method. The false-negative rate declined with higher q and μ and became zero when $q \geq 0.1$ even when $\mu = 3$ under all of

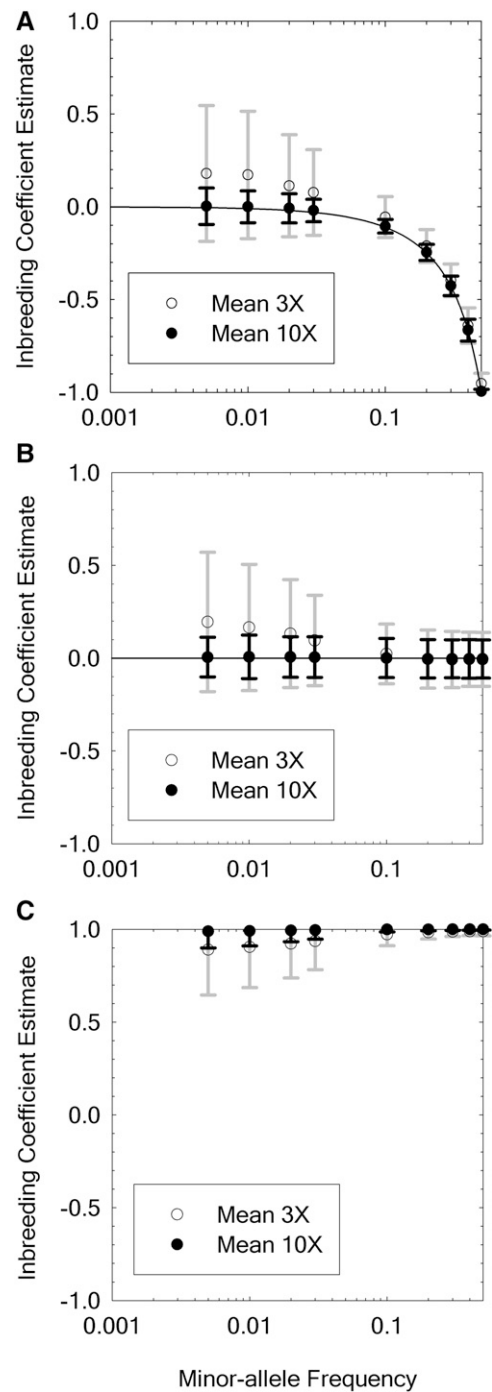


Figure 2 ML estimate of the inbreeding coefficient. The ML estimate of the inbreeding coefficient f as a function of the minor-allele frequency is shown when f is (A) minimized, (B) equal to zero (Hardy–Weinberg equilibrium), or (C) maximized. The results are conditioned on significant polymorphism at the 5% level. The mean and standard deviation of the estimated f are shown by the symbols and bars (shaded for mean 3 $\times$ and solid for mean 10 $\times$), respectively. The curve in A represents the ideal situation where the ML estimate is equal to the true value. The true value of f is zero (shown by the line) and one in B and C, respectively. Number of sampled individuals $N = 100$, error rate $\varepsilon = 0.01$. A total of 10,000 simulation replications were run for each set of parameter values.

Table 2 Interplay among the parameter estimates as a function of the inbreeding coefficient f when the minor-allele frequency q is low ($q < 0.1$) and depth of coverage is low (e.g., mean $3\times$)

f	$\hat{q}$	$\hat{f}$	$\hat{Q}$
Minimized	Slightly overestimated	Somewhat overestimated	Slightly overestimated
Zero (HWE)	Slightly overestimated	Somewhat overestimated	Slightly overestimated
Maximized	Slightly overestimated	Somewhat underestimated	Slightly underestimated

Q denotes minor-homozygote frequency, and the circumflex above a symbol denotes an estimate.

the conditions. These results indicate that the proposed method both is conservative and has reasonably high power for detecting polymorphisms with minor-allele frequencies ≥ 0.1 .

In carrying out power analysis of the HWE-deviation detection by the proposed method as a function of μ , we rejected the null hypothesis of HWE at a site at the 5% significance level (when LRT_{HWE} in Equation 15 was >3.841). Because we carry out the statistical test of HWE deviation only when the site is considered to be polymorphic, and the power for detecting polymorphisms depends on minor-allele frequencies, we examined false-positive and -negative rates of HWE-deviation detection as a function of q (Figure 4). Overall, the false-positive rate was close to or below the specified significance level (0.05) with examined parameter values (Figure 4A). When μ was low (3), the false-positive rate had its minimum when $q = 0.1$. This pattern can be understood by considering the sampling effects. When $q < 0.1$, the frequency of minor homozygotes is low (< 0.01), and most minor alleles exist in heterozygotes. When μ is low, only one of the alleles can be sequenced from heterozygotes, leading to excess of homozygotes and false deviation from HWE. When $q > 0.1$, the frequency of heterozygotes becomes relatively high, and they can be sampled more often than expected, leading to excess of heterozygotes and false deviation from HWE.

The false-negative rate of the HWE-deviation detection decreased with increased q and μ both when f was minimized and when f was maximized (Figure 4, B and C). The false-negative rate was high when f was minimized and q was low ($q \leq 0.1$) (Figure 4B). However, when $q \geq 0.3$, the false-negative rate was reasonably low both when $\mu = 3$ and $\mu = 10$. On the other hand, when f was maximized, the false-negative rate was reasonably low (Figure 4C). In particular, when $q \geq 0.1$, the false-negative rate was essentially zero even when $\mu = 3$. The much higher false-negative rate with minimum f compared to that with maximum f , especially when q is low, is not due to faults of the proposed method but due to the following factors. First, the deviation of genotype frequencies from HWE, which can be measured as the absolute value of f , is small with low q when f is minimized compared to that when f is maximized. When f is minimized, $f = q/(q - 1)$ (Weir 1996) and $|f| = q/(1 - q)$, which is small when q is low. On the other hand, when f is maximized, $|f|$ is always equal to one, regardless of q . Second, only one of the alleles from a heterozygote can be sampled when μ is low. These make detecting the

excess of heterozygotes much more difficult compared to detecting the deficit of heterozygotes.

Our evaluation of the performance of the proposed method with the low-coverage (mean $4\times$) sequencing data on chromosome 6 for 81 individuals revealed 8,528,190 sites with minor-allele frequency (MAF) estimates > 0 , implying potential polymorphisms. The vast majority (92%) of these were singletons or doubletons, with the number of sites decreasing with increasing minor-allele count estimates (Figure 5A). Of these potential polymorphisms, 452,577 were significant at the 5% level, and 33% of these were singletons or doubletons. These numbers are much less than the corresponding values for the entire pool of sites with MAF estimates > 0 , reflecting both the difficulty with distinguishing rare alleles from sequence errors and the limited power of the statistical test for finding significant polymorphisms involving rare alleles with low depths of coverage. However, the site-frequency distribution at significantly polymorphic sites was nearly identical to that for the entire pool of sites with MAF estimates > 0 , provided the minor-allele count estimate was > 16 (MAF estimates > 0.1). In fact, the fraction of significantly polymorphic sites is $\geq 95\%$ when MAF estimates are > 0.1 (Figure 5B), which is consistent with the simulation results showing that the power of the method is high with MAF > 0.1 even when depths of coverage are low.

We compared the performance of our allele-frequency estimation with that by ANGSD (Korneliussen *et al.* 2014). We estimated allele frequencies at each site, using the method by Kim *et al.* (2011), and compared their allele-frequency estimates with ours (Figure S3). Their allele-frequency estimates with both Samtools and GATK genotype likelihoods were similar to ours except when they were close to zero or one. When allele-frequency estimates were close to zero or one, their estimates were closer to zero or one than ours.

We compared the site-frequency spectrum estimated by the proposed method with that by Nielsen *et al.* (2012). Of the sites analyzed, 0.54% were significantly polymorphic at the 5% level by the proposed method. The fraction of polymorphic sites by Nielsen *et al.*'s method was similar to ours but varied, depending on the genotype-likelihood model used, 0.39% with the Samtools model and 0.51% with the GATK model.

Consistent with the findings by Korneliussen *et al.* (2014), the site-frequency spectrum (SFS) by ANGSD also varied, depending on the genotype-likelihood model used (Figure S4). Specifically, the SFS with the GATK model showed more

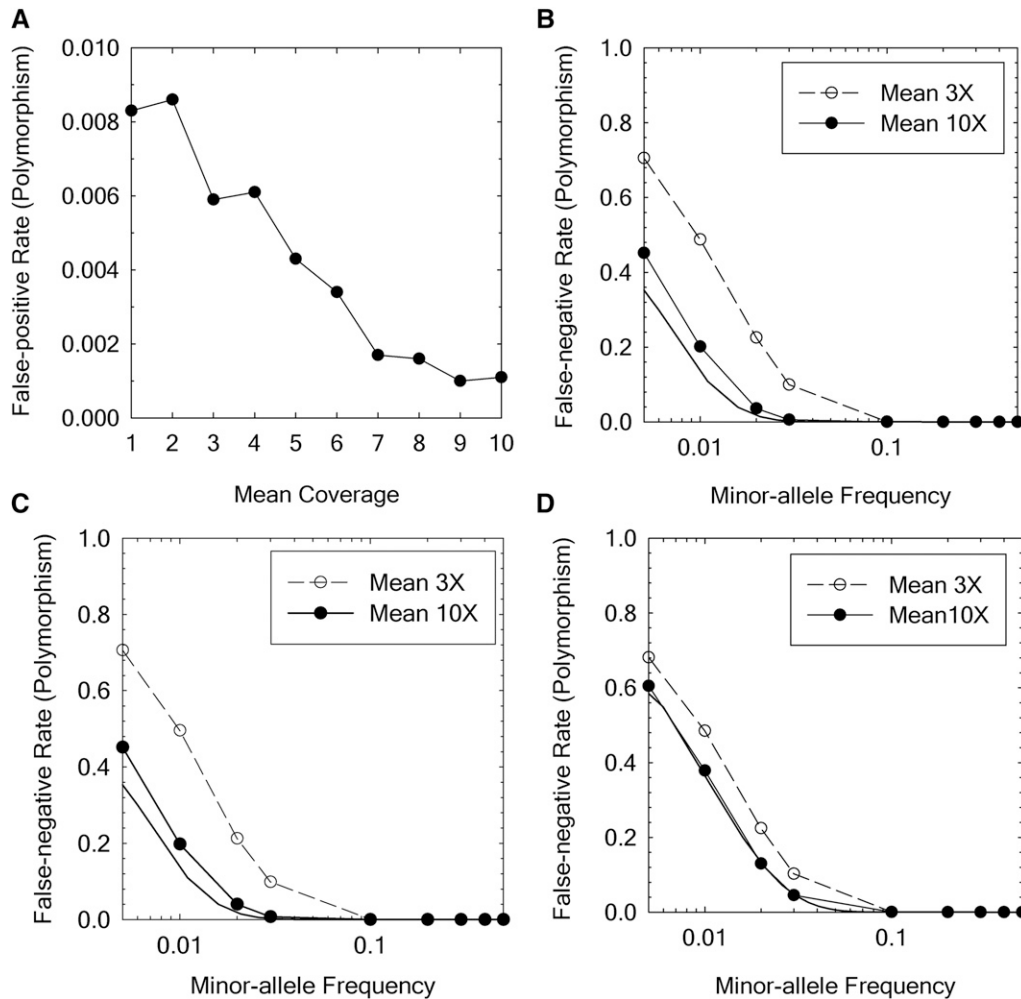


Figure 3 Power analysis of polymorphism detection by the proposed method. The significance of the polymorphism test is set at the 5% level. (A) The false-positive rate of polymorphism detection as a function of the mean depth of coverage is shown. The false-negative rate of polymorphism detection as a function of the minor-allele frequency q is shown when the inbreeding coefficient is (B) minimized, (C) equal to zero (Hardy-Weinberg equilibrium), or (D) maximized. The solid curve in B–D represents the theoretical minimum of the false-negative rate, which is due to sampling only one of the alleles and calculated as $(1 - 2q)^N$, $(1 - q)^{2N} + q^{2N}$, and $(1 - q)^N + q^N$, respectively, where N denotes the number of sampled individuals. $N = 100$, error rate $\varepsilon = 0.01$. A total of 10,000 simulation replications were run for each set of parameter values.

polymorphic sites with low minor-allele frequencies than that with the Samtools model. The corresponding SFS by the proposed method was more similar to Nielsen *et al.*'s (2012) SFS with the GATK model than to that with the Samtools model. Interestingly, our SFS showed more singletons and doubletons and was smoother than either of SFSs estimated by ANGSD. Furthermore, our SFS showed an excess of singletons and doubletons compared to the theoretical prediction in a population with constant size (Ewens 2004), which is consistent with previous results reporting that many human populations show an excess of rare variants as a result of recent population growth (e.g., Coventry *et al.* 2010; Keinan and Clark 2012). These results indicate that the proposed method may enable more accurate description of the SFS and powerful detection of rare variants than the currently widely used methods.

To evaluate the deviation of genotype frequencies from HWE, we examined the inbreeding coefficient estimates as a function of MAF estimates at the significantly polymorphic sites (Figure 6A). The majority of the sites had inbreeding coefficient estimates close to zero, and the mean estimate (0.038) was also close to zero. Of the 452,577 significant SNPs, only 18,595 (4%) significantly deviated from HWE.

These results are consistent with previous studies estimating small amounts of inbreeding in human populations (e.g., Weir *et al.* 2004). When MAF estimates were low (<0.05), the inbreeding coefficient estimates were relatively high, presumably due to the upward biases when MAFs are extreme and depths of coverage are low. In fact, only a small fraction (2%) of significant SNPs with MAF estimates <0.05 significantly deviated from HWE (Figure 6B). When MAF estimates were intermediate (between 0.1 and 0.4), the mean of the inbreeding coefficient estimates fluctuated around -0.01 , which is close to the intrinsic downward bias of the inbreeding coefficient of $1/(2N - 1)$ when ML methods are applied to populations in HWE (Weir 1996).

Interestingly, when MAF estimates approached 0.5, the inbreeding coefficient estimates suddenly dropped, and a substantial fraction (21%) of significant SNPs significantly deviated from HWE. Because our inbreeding coefficient estimates are essentially unbiased when MAFs are high, this is not due to an artifact of the proposed method. To examine the cause of this observation, we examined the spatial patterns of polymorphisms on chromosome 6 by sliding-window analyses (Figure 7). Consistent with previous studies (Garrigan and Hedrick 2003; Solberg *et al.* 2008), heterozygosity was highly

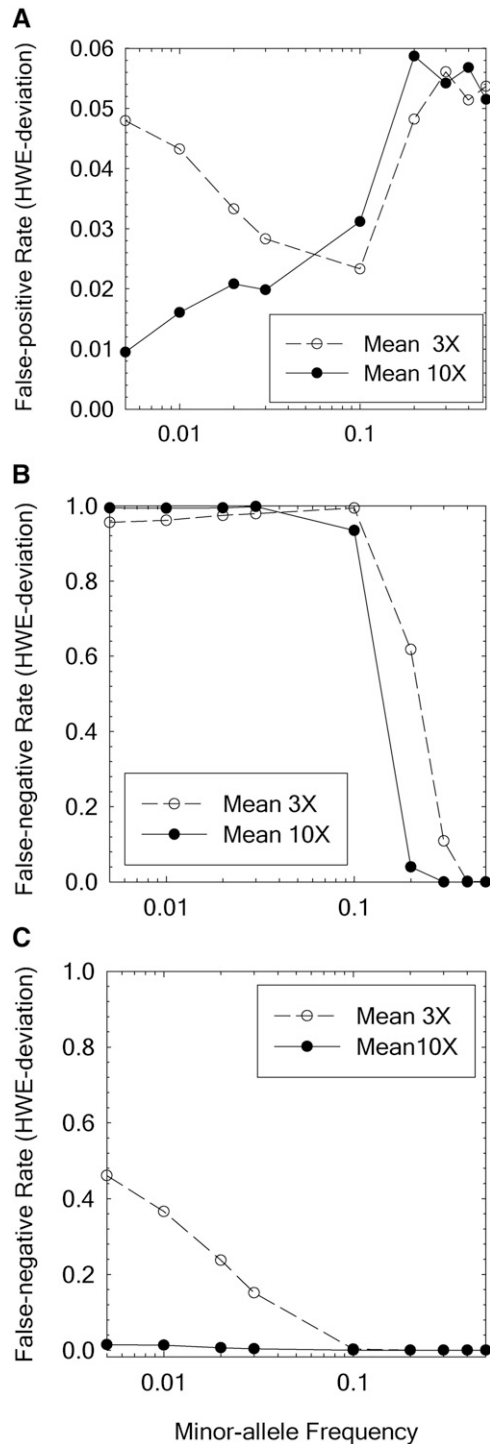


Figure 4 Power analysis of the detection of deviation from Hardy-Weinberg equilibrium by the proposed method. The results are conditioned on significant polymorphism at the 5% level. The significance of the Hardy-Weinberg equilibrium (HWE) deviation test is also set at the 5% level. (A) The false-positive rate of HWE-deviation detection as a function of the minor-allele frequency (MAF) is shown. (B and C) The false-negative rate of HWE-deviation detection as a function of MAF is shown when the inbreeding coefficient is (B) minimized or (C) maximized. Number of sampled individuals $N = 100$, error rate $\varepsilon = 0.01$. A total of 10,000 simulation replications were run for each set of parameter values.

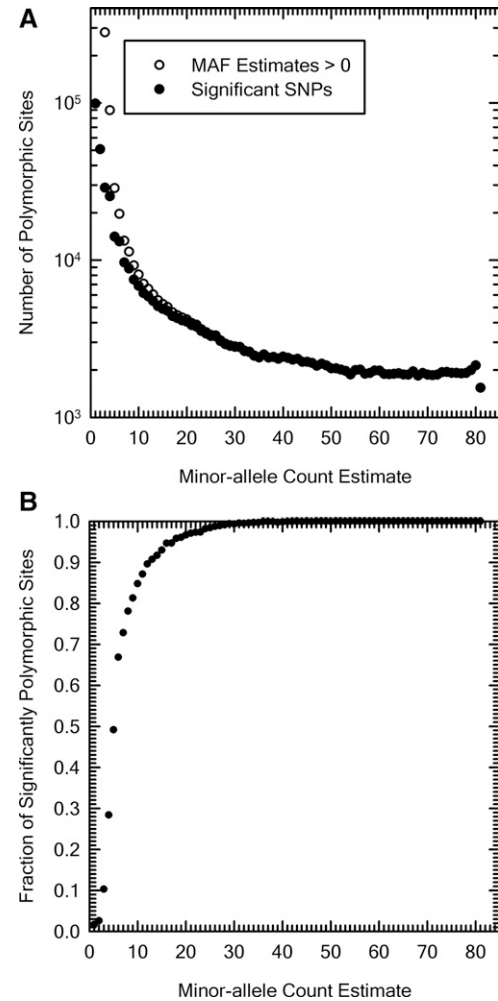


Figure 5 Polymorphisms on human chromosome 6 characterized by the proposed method. (A) The distribution of polymorphic sites in terms of their minor-allele count estimates. The vast majority of sites with nonzero MAF estimates are singletons and doubletons (5,923,034 and 1,951,233 sites, respectively) and these results are not shown. All results are shown for significant (at the 5% level) SNPs. (B) The fraction of significantly polymorphic sites as a function of their minor-allele count estimates.

elevated in the region around 3×10^7 bp, where many human leukocyte antigen (HLA) genes are clustered (Figure 7A). However, inbreeding coefficient estimates in the HLA region (28,477,797–33,448,354 bp) were not very different from those in the other regions (Figure 7C).

The heterozygosity was also highly elevated compared to the others in the region around 5.74×10^7 , but here the inbreeding coefficient estimates were highly negative compared to those in the other regions. The unusual excess of heterozygote frequency estimates in this region was also seen by sliding-window analysis of the disequilibrium coefficients, which were highly negative in this region but close to zero in the majority of the other analyzed regions (Figure S5A). Contrary to our analyses, the corresponding inbreeding coefficient estimates by the method of Vieira *et al.* (2013) were close to zero in this region, implying HWE (Figure S5B). In fact, Vieira *et al.*'s method estimated no negative inbreeding

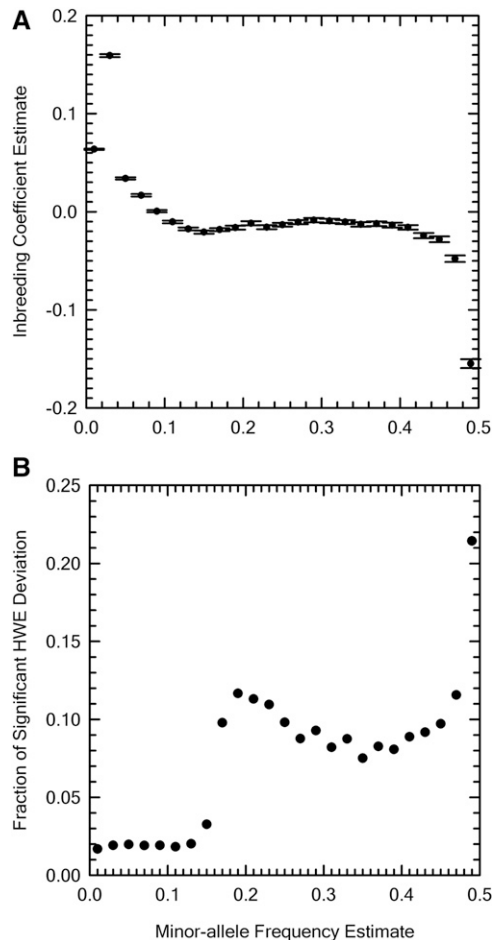


Figure 6 Inbreeding coefficient estimates on human chromosome 6. Results are conditioned on significant polymorphism at the 5% level. (A) The midpoint value of the minor-allele frequency (MAF) estimate and mean (symbols) $\pm$ SE (bars) of inbreeding coefficient estimates are shown in each of the MAF bins with equal size of 0.02. The mean and standard error of the inbreeding coefficient estimates among all significant SNPs are 0.038 and 0.00037, respectively. (B) The fraction of significant SNPs significantly (at the 5% level) deviating from Hardy–Weinberg equilibrium (HWE), as a function of their MAF estimates.

coefficients. Although the error rate estimates in this region were not very different from those in the other regions (Figure S5C), depths of coverage in this region were somewhat higher compared to those in the others (Figure 7, B and D), raising the possibility of misassembly and subsequent mismapping. In addition, the overall error rate estimates were positively correlated with depths of coverage outside this region, which is consistent with recent findings that error rates in genotype calling increased with higher depths of coverage with Illumina sequencing data (Wall *et al.* 2014).

The peak of negative inbreeding coefficients was at the *PRIM2* gene, which has been suggested to be a putative target of balancing selection by some researchers (*e.g.*, Hodgkinson and Eyre-Walker 2010; DeGiorgio *et al.* 2014). Because the elevation of the coverage in the region around the *PRIM2* gene is not very extreme and the gene is functionally important, distinguishing the two hypotheses (balancing selection

vs. misassembly) for the unusual polymorphism estimates around the gene is not easy. However, a recent study (Genovese *et al.* 2013) found evidence of misassembly in the region around the *PRIM2* gene and the existence of “cryptic” segmental duplicates missing from the reference genome.

Discussion

The rapidly declining cost of high-throughput sequencing provides unprecedented opportunities for carrying out population-genomic analyses in various organisms. To take full advantage of these opportunities, it is essential to accurately estimate allele and genotype frequencies without assuming HWE. Although HWE is often assumed in previous allele-frequency estimators (*e.g.*, Lynch 2009; Keightley and Halligan 2011; Kim *et al.* 2011), this condition is certainly violated in some organisms and environmental/sampling settings. Some loci deviate from HWE even in randomly mating populations due to, for example, natural selection.

Our ML method relaxes the assumption of HWE in the allele-frequency estimator of Lynch (2009) and estimates site-specific allele frequencies and error rates directly from sequence-read data. Because sequencing error rates are known to vary among sites (Nakamura *et al.* 2011), estimating site-specific error rates is useful for factoring out confounding effects on allele-frequency estimation. By estimating error rates from the data themselves, and not relying on *ad hoc* estimates, our method factors out errors beyond those resulting from sequencing processes, *e.g.*, those introduced during sample preparation and others that cannot be inferred from read quality scores alone. Because there is growing interest in carrying out population-genomic analyses using time-series data (*e.g.*, Franssen *et al.* 2015), where errors may be introduced during long storage periods or error rates may change owing to technological changes, our method should have an advantage over others that rely simply on read quality scores (*e.g.*, Li *et al.* 2009b; DePristo *et al.* 2011; Li 2011).

Another advantage of our method is the incorporation into the likelihood function of uncertainties in individual genotypes and sampling of parental chromosomes in diploid organisms. Because of the random sequencing of parental chromosomes, confident inference of genotypes of diploid individuals is difficult to achieve when the depth of coverage is low, and many other methods use an arbitrary coverage cutoff to avoid this problem. However, the latter procedure can lead to the loss of substantial information, and recent studies have shown that allele frequencies directly estimated from mapped sequence reads using ML methods are unbiased, whereas those obtained via genotype calling are biased when the depth of coverage is low (*e.g.*, Kim *et al.* 2011; Han *et al.* 2014).

In addition to yielding allele-frequency estimates, the proposed method enables estimation of genotype frequencies, which is necessary for examining HWE deviations, and hence can be applied to populations with arbitrary mating systems

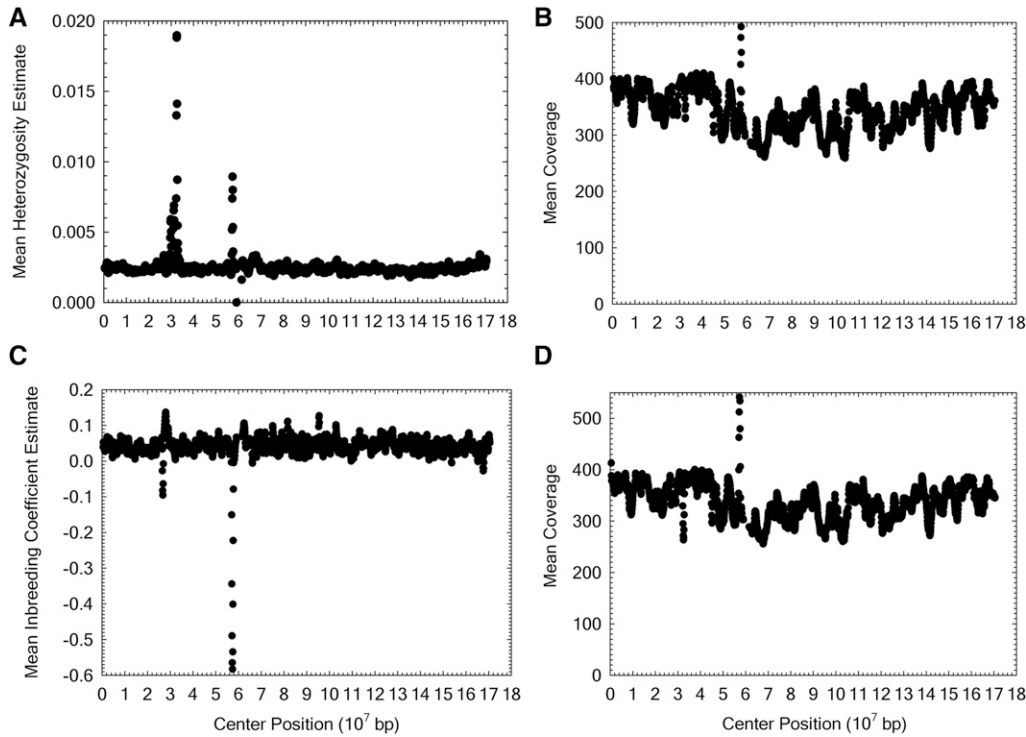


Figure 7 Sliding-window analyses of polymorphisms on human chromosome 6. (A and B) Weighted means of the per-site heterozygosity estimates and depths of coverage in the population sample among sites with data in each window. (C and D) Weighted means of the inbreeding coefficient estimates and depths of coverage in the population sample among significantly (at the 5% level) polymorphic sites in each window. $s = 100,000$ (bp) and $w = 150,000$ (bp), used for defining the step size and width of the windows, respectively.

and/or internal population structure. Using computer simulations, we showed that our method yields essentially unbiased estimates of allele and genotype frequencies with moderately high depths of coverage. Furthermore, the sampling variances of our estimates asymptotically approach the theoretically minimum values at high depths of coverage, indicating that the method maximally exploits the information in sampling data. Finally, we provide systematic statistical tests, which are conservative and have reasonably high power for testing the significance of polymorphisms and HWE deviations. Although we focused on analyzing diploid populations in this study, the allele-frequency estimation and significance test of polymorphisms can be similarly made for haploid populations.

We examined the performance of the proposed method by applying it to low-coverage (mean $4\times$), high-throughput sequencing data of 81 humans. The abundance of rare alleles in our site-frequency spectrum of significant SNPs indicates that the proposed method has reasonably high power for detecting polymorphic sites harboring rare alleles even with low depths of coverage. Although the vast majority of the examined SNPs did not significantly deviate from HWE, by examining the distribution of the estimated inbreeding coefficients of the SNPs and statistically testing the significance of their deviations from HWE, we found a region most likely to be misassembled in the reference genome (GRCh37). Our statistical framework for identifying SNPs deviating from HWE is especially useful for finding unreliable regions due to misassembly and also carrying out subsequent population-genetic analyses identifying putative targets of natural selection. To promote the use of our method with other data sets,

we uploaded its software package at <https://github.com/Takahiro-Maruki/Package-GFE>.

To find the optimal sequencing strategy for estimating genotype frequencies under limited research budgets, we examined the root mean-square deviation (RMSD) of the inbreeding coefficient estimates as a function of the mean depth of coverage μ , fixing the product of μ and N (the number of sampled individuals) at 1000, using computer simulations (Figure S6). Unless the focus is on genotypes at sites with extreme frequencies (minor-allele frequency <0.05), the smallest RMSD is generally achieved when the mean depth of coverage is somewhere between $4\times$ and $10\times$. Because the variability of the depth of coverage in real data is higher than the Poisson expectation (Quail *et al.* 2012), which we relied upon in simulations, we recommend a mean of $\sim 10\times$ as the optimal depth of coverage for population-genomic studies examining genotype frequencies.

The proposed method greatly enhances our ability to carry out subsequent population-genomic analyses. For example, by incorporating genotype-frequency estimates in genotype calling using Bayes' theorem, the accuracy of genotype calls can be improved (*e.g.*, Martin *et al.* 2010; Nielsen *et al.* 2012). Another useful application involves the estimation of Wright's fixation indexes. The proposed method enables accurate and rapid estimation of allele frequencies in a population regardless of its mating system and population structure, which is necessary for estimating F_{ST} . Genotype-frequency estimates not only enable better estimation of F_{ST} but also enable estimation of F_{IS} (Nei and Chesser 1983; Weir 1996), which is important for improving our ability to identify putative targets of natural selection (Black *et al.* 2001). Finally, the accurate

and rapid estimation of allele frequencies by the proposed method greatly reduces the computational demands in estimating linkage disequilibrium at the population level (Maruki and Lynch 2014). In all of these applications, our systematic statistical tests can play useful roles in avoiding the downstream analysis of false polymorphisms.

Acknowledgments

We thank Sen Xu for his help in designing the input file of our program for data analysis. We also thank Ken Spitze for his feedback on using the program. In addition, we thank Matthew Ackerman for his technical assistance. This work was supported by National Science Foundation (NSF) grant DEB-1257806 and National Institutes of Health (NIH) grant NIH-National Institute of General Medical Sciences 1R01GM101672-01A1. It was also supported by the National Center for Genome Analysis Support, funded by NSF grant ABI-1062432 to Indiana University, and by Indiana University Research Technology's computational resources.

Literature Cited

- Black, IV, W. C., C. F. Baer, M. F. Antolin, and N. M. DuTeau, 2001 Population genomics: genome-wide sampling of insect populations. *Annu. Rev. Entomol.* 46: 441–469.
- Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell *et al.*, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* 1: 131.
- DeGiorgio, M., K. E. Lohmueller, and R. Nielsen, 2014 A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet.* 10: e1004561.
- DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43: 491–498.
- Ewens, W. J., 2004 *Mathematical Population Genetics*. Springer-Verlag, New York.
- Franssen, S. U., V. Nolte, R. Tobler, and C. Schlotterer, 2015 Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental *Drosophila melanogaster* populations. *Mol. Biol. Evol.* 32: 495–509.
- Garrigan, D., and P. W. Hedrick, 2003 Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution* 57: 1707–1722.
- Genovese, G., R. E. Handsaker, H. Li, N. Altomose, A. M. Lindgren *et al.*, 2013 Using population admixture to help complete maps of the human genome. *Nat. Genet.* 45: 406–414e2.
- Glenn, T. C., 2011 Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11: 759–769.
- Han, E., J. S. Sinsheimer, and J. Novembre, 2014 Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol. Biol. Evol.* 31: 723–735.
- Hellmann, I., Y. Mang, Z. Gu, P. Li, F. M. de la Vega *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.* 18: 1020–1029.
- Hodgkinson, A., and A. Eyre-Walker, 2010 The genomic distribution and local context of coincident SNPs in human and chimpanzee. *Genome Biol. Evol.* 2: 547–557.
- Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson *et al.*, 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* 6: e1000862.
- Jiang, R., S. Tavare, and P. Marjoram, 2009 Population genetic inference from resequencing data. *Genetics* 181: 187–197.
- Johnson, P. L., and M. Slatkin, 2008 Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* 25: 199–206.
- Keightley, P. D., and D. L. Halligan, 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.
- Keinan, A., and A. G. Clark, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336: 740–743.
- Kendall, M., and A. Stuart, 1979 *The Advanced Theory of Statistics*, Vol. 2, Ed. 4. Charles Griffin & Co. London, UK.
- Kim, S. Y., K. E. Lohmueller, A. Albrechtsen, Y. Li, T. Korneliussen *et al.*, 2011 Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* 12: 231.
- Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15: 356.
- Le, S. Q., and R. Durbin, 2011 SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* 21: 952–960.
- 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.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009a The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Li, R., Y. Li, X. Fang, H. Yang, J. Wang *et al.*, 2009b SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 19: 1124–1132.
- Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet, 2003 The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4: 981–994.
- Lynch, M., 2009 Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* 182: 295–301.
- Lynch, M., D. Bost, S. Wilson, T. Maruki, and S. Harrison, 2014 Population-genetic inference from pooled-sequencing data. *Genome Biol. Evol.* 6: 1210–1218.
- Martin, E. R., D. D. Kinnam, M. A. Schmidt, E. H. Powell, S. Zuchner *et al.*, 2010 SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics* 26: 2803–2810.
- Maruki, T., and M. Lynch, 2014 Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. *Genetics* 197: 1303–1313.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.
- Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa *et al.*, 2011 Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 39: e90.
- Nei, M., and R. K. Chesser, 1983 Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.* 47: 253–259.
- Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337: 100–104.

- Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang, 2012 SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One* 7: e37558.
- 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.
- Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. *Genome Res.* 20: 291–300.
- Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris *et al.*, 2012 A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341.
- Solberg, O. D., S. J. Mack, A. K. Lancaster, R. M. Single, Y. Tsai *et al.*, 2008 Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum. Immunol.* 69: 443–464.
- Vieira, F. G., M. Fumagalli, A. Albrechtsen, and R. Nielsen, 2013 Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. *Genome Res.* 23: 1852–1861.
- Wall, J. D., L. F. Tang, B. Zerbe, M. N. Kvale, P. Y. Kwok *et al.*, 2014 Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res.* 24: 1734–1739.
- Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, Massachusetts.
- Weir, B. S., W. G. Hill, L. R. Cardon, and S. N. P. Consortium, 2004 Allelic association patterns for a dense SNP map. *Genet. Epidemiol.* 27: 442–450.

Communicating editor: N. Risch

Appendix

Preliminary ML estimates of the major-allele frequency and error rate can be analytically found in the following way. Let ϕ_M , ϕ_m , and ϕ_e denote the probability of the observed major-allele read count, minor-allele read count, and read count of the other nucleotides, respectively. Then, each of these is calculated by summing the probability of the observed read over the three genotypes:

$$\begin{aligned}\phi_M &= (p^2 + D_A)(1 - \epsilon) + 2\{p(1 - p) - D_A\}\left(\frac{1}{2} - \frac{\epsilon}{3}\right) \\ &\quad + \{(1 - p)^2 + D_A\}\frac{\epsilon}{3} \\ &= \left\{1 - \frac{4}{3}\epsilon\right\}p + \frac{\epsilon}{3} \\ \phi_m &= (p^2 + D_A)\frac{\epsilon}{3} + 2\{p(1 - p) - D_A\}\left(\frac{1}{2} - \frac{\epsilon}{3}\right) \\ &\quad + \{(1 - p)^2 + D_A\}(1 - \epsilon) \\ &= 1 - \left\{1 - \frac{4}{3}\epsilon\right\}p - \epsilon \\ \phi_e &= \frac{\epsilon}{3}.\end{aligned}$$

The log-likelihood of the observed set of site-specific sequence reads for individual i , $\ln L_i$, is

$$\ln L_i = n_{iM} \ln \phi_M + n_{im} \ln \phi_m + n_{ie} \ln \phi_e,$$

where $n_{ie} = n_{ie_1} + n_{ie_2}$. The preliminary ML estimates of the major-allele frequency $\hat{p}$ and error rate $\hat{\epsilon}$ are obtained by taking the derivative of the log-likelihood with respect to p and ϵ , summing each of the derivatives over the individuals, and equating them to zero,

$$\begin{aligned}\hat{p} &= \frac{2n_M - n_e}{2(n_M + n_m - n_e)}, \\ \hat{\epsilon} &= \frac{3}{2} \cdot \frac{n_e}{n_T},\end{aligned}$$

where n_M , n_m , n_e , and n_T are the read counts of the major allele, the minor allele, other nucleotides, and their sum, respectively, in the population sample. These estimators are identical to those that we derived for pooled-sample high-throughput sequencing data (Lynch *et al.* 2014).

GENETICS

Supporting Information

www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179077/-/DC1

Genotype-Frequency Estimation from High-Throughput Sequencing Data

Takahiro Maruki and Michael Lynch

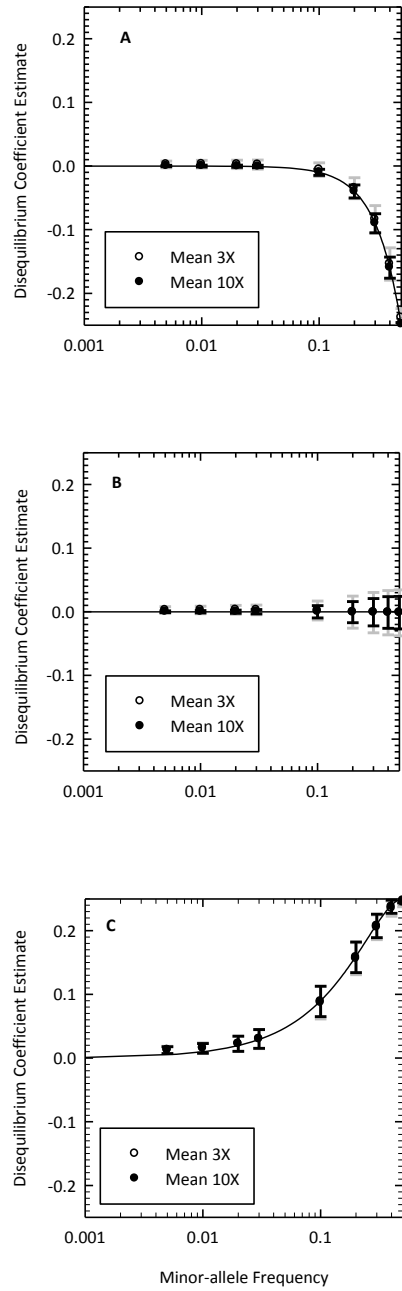


Figure S1 ML estimate of the disequilibrium coefficient. The ML estimate of the disequilibrium coefficient D_A as a function of the minor-allele frequency is shown when D_A is A) minimized, B) equal to zero (Hardy-Weinberg equilibrium), or C) maximized. The results are conditioned on significant polymorphism at the 5% level. The mean and standard deviation of the estimated D_A are shown by the point and bar (gray for mean 3 $\times$ and black for mean 10 $\times$), respectively. A, C) The curve represents the ideal situation where the ML estimate is equal to the true value. B) The true value of D_A is zero (shown by the line). Number of sampled individuals $N = 100$, error rate $\epsilon = 0.01$. A total of 10,000 simulation replications were run for each set of parameter values.

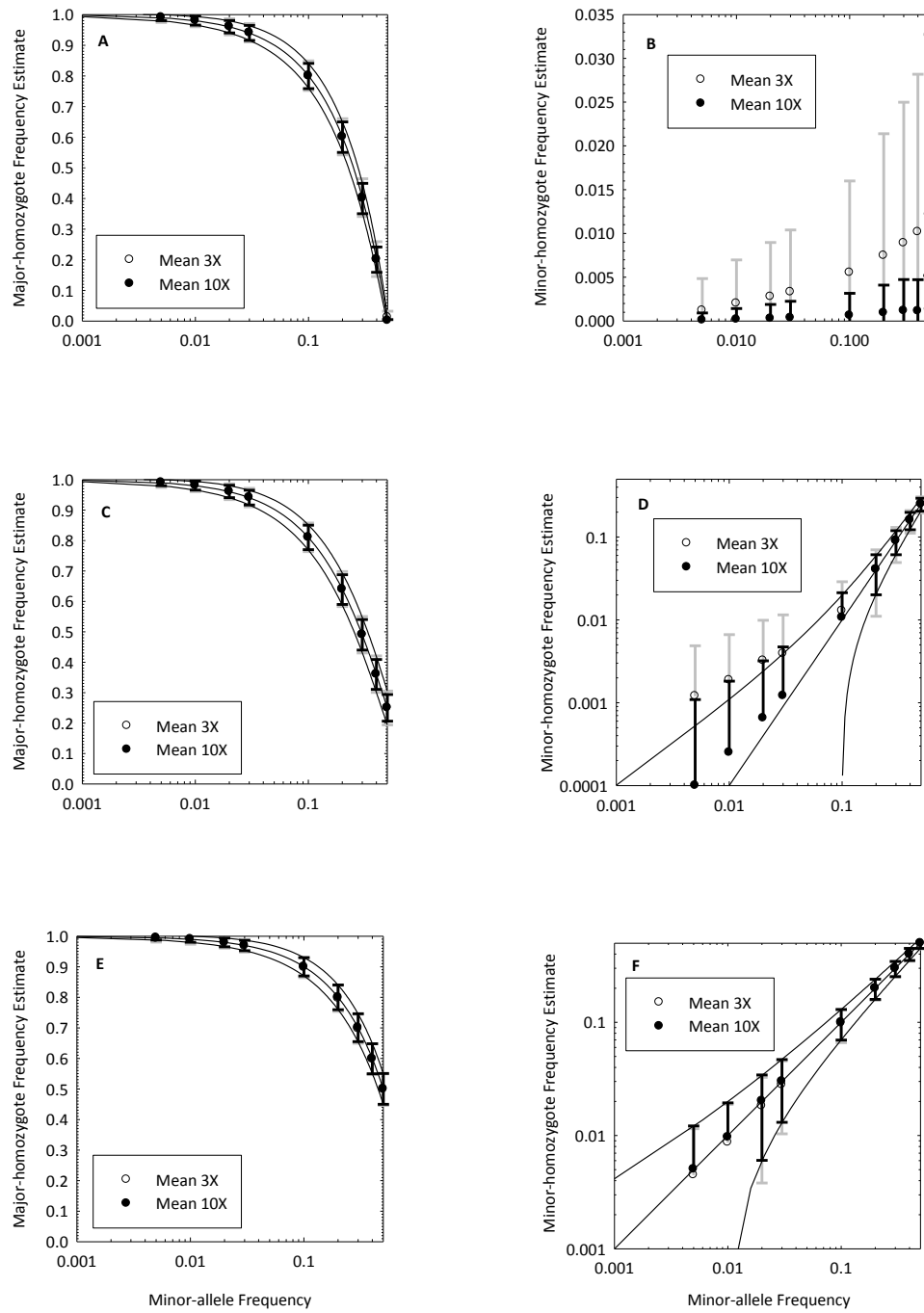


Figure S2 ML estimates of the genotype frequencies. The ML estimates of the A, C, E) major-homozygote frequency and B, D, F) minor-homozygote frequency as functions of the minor-allele frequency are shown when the inbreeding coefficient is A, B) minimized, C, D) equal to zero, or E, F) maximized. The mean and standard deviation of the estimates are shown by the points and bars (gray for mean 3 $\times$ and black for mean 10 $\times$), respectively. The middle curve represents the ideal situation where the estimate is equal to the true value. The upper and lower curves represent the theoretical asymptotic sampling standard deviation from the mean (calculated as the square root of Equation 17 or 18). Number of sampled individuals $N = 100$, error rate $\epsilon = 0.01$. A total of 10,000 simulation replications were run for each set of parameter values.

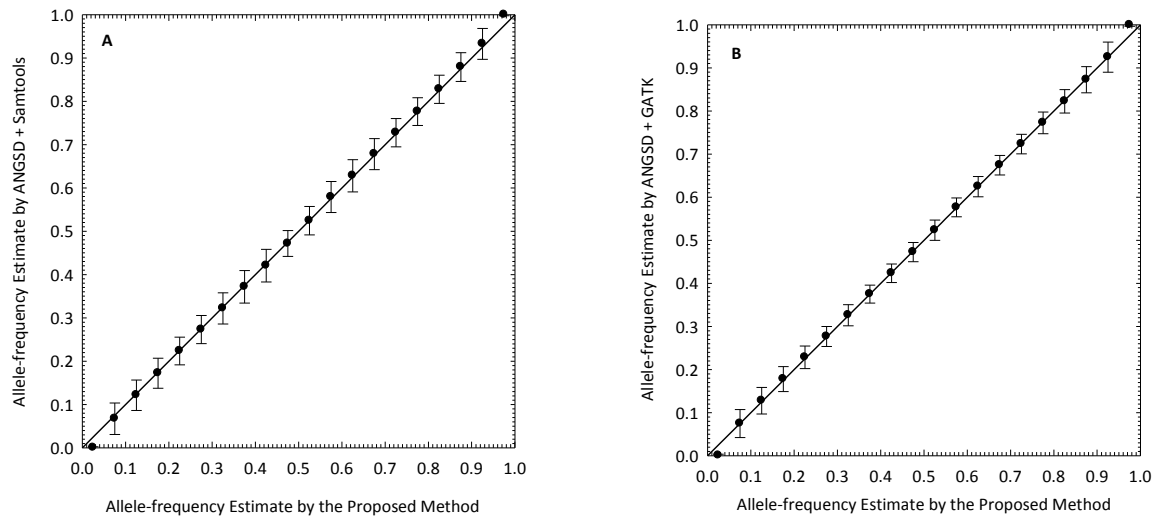


Figure S3 Comparison of the allele-frequency estimates on human chromosome 6 by different methods. The allele frequencies estimated using the method by Kim *et al.* (2011) (ANGSD) and the A) Samtools or B) GATK genotype likelihood model is compared to those estimated using the proposed method. The points and bars represent the mean and standard deviation of the ANGSD allele-frequency estimates in each bin of the allele-frequency estimates of size 0.05 by the proposed method.

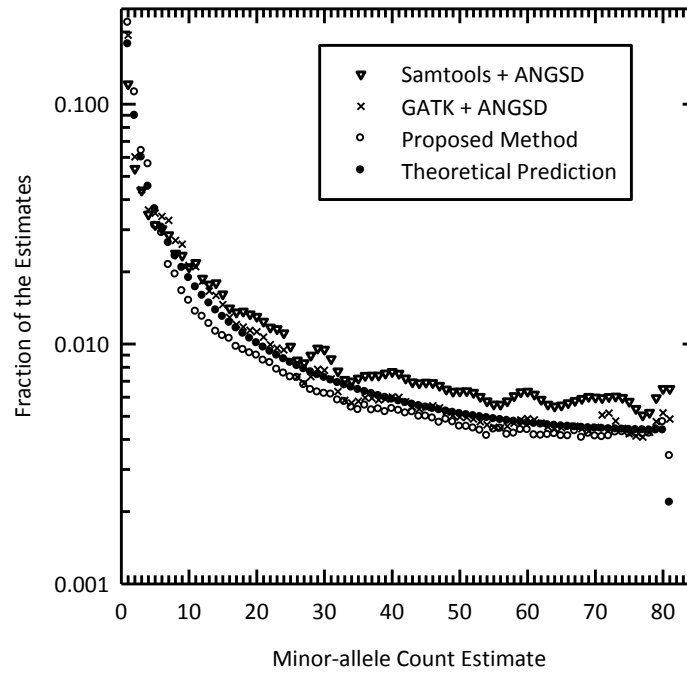


Figure S4 Comparison of site-frequency spectra on human chromosome 6 estimated by different methods. The site-frequency spectrum estimated using the method by Nielsen *et al.* (2012) (ANGSD) and the Samtools or GATK genotype likelihood model is compared to that estimated using the proposed method. The theoretical prediction is in a population with constant size, and is given by $[1/j + 1/(162-j)]/[\sum_{k=1}^{161} 1/k]$ for $1 \leq j < 81$ and $(1/j)/[\sum_{k=1}^{161} 1/k]$ for $j = 81$, where j is the minor-allele count (Ewens 2004).

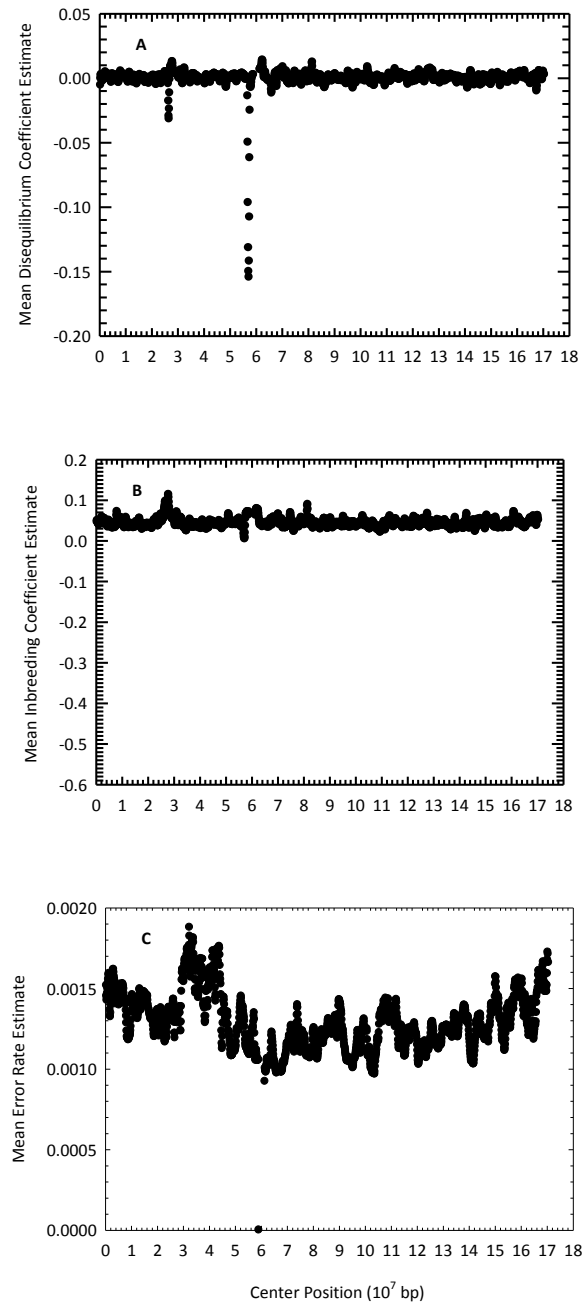


Figure S5 Sliding-window analyses of parameter estimates on human chromosome 6. A) Sliding-window analysis of the disequilibrium coefficient estimates by the proposed method. B) Sliding-window analysis of the per-site inbreeding coefficients estimated by the method by Vieira *et al.* (2013). C) Sliding-window analysis of the error rate estimates by the proposed method. A, B) Results are conditioned on significant polymorphism at the 5% level. C) Results are based on all sites with sequence-read data. $s = 100,000$ (bp) and $w = 150,000$ (bp) used for defining the step size and width of the windows, respectively.

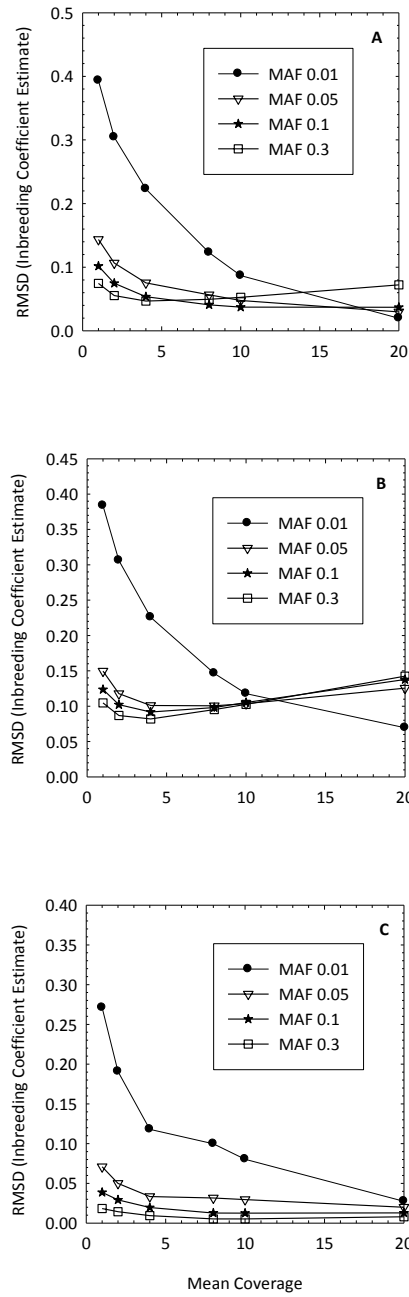


Figure S6 Root mean square deviation of the inbreeding coefficient estimates, as a function of the mean depth of coverage, estimated from a fixed number of total sequence reads. The product of the mean depth of coverage μ and number of sampled individuals N is fixed at 1,000 such that a two-fold increase in μ resulted in a two-fold decrease in N . The root mean square deviation (RMSD) when the inbreeding coefficient is A) minimized, B) equal to zero (Hardy-Weinberg equilibrium), or C) maximized is shown. Results are conditioned on significant polymorphism at the 5% level. The error rate $\epsilon = 0.01$. A total of 10,000 simulation replications were run for each set of parameter values.

File S1

Computer program for simulated data analysis.

File S1 is available as a text file at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179077/-/DC1.

File S2

README of the computer program for simulated data analysis.

File S2 is available as a PDF file at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179077/-/DC1.

File S3

Names of the analyzed BAM files of the Illumina-sequencing read data of 81 individuals from the CEU population.

File S3 is available as a PDF file at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179077/-/DC1.