A Novel Approach to Estimating Heterozygosity from Low-Coverage Genome Sequence

Katarzyna Bryc,*[,1] Nick Patterson,[†] and David Reich*
*Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, and [†]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

ABSTRACT High-throughput shotgun sequence data make it possible in principle to accurately estimate population genetic parameters without confounding by SNP ascertainment bias. One such statistic of interest is the proportion of heterozygous sites within an individual's genome, which is informative about inbreeding and effective population size. However, in many cases, the available sequence data of an individual are limited to low coverage, preventing the confident calling of genotypes necessary to directly count the proportion of heterozygous sites. Here, we present a method for estimating an individual's genome-wide rate of heterozygosity from low-coverage sequence data, without an intermediate step that calls genotypes. Our method jointly learns the shared allele distribution between the individual and a panel of other individuals, together with the sequencing error distributions and the reference bias. We show our method works well, first, by its performance on simulated sequence data and, second, on real sequence data where we obtain estimates using low-coverage data consistent with those from higher coverage. We apply our method to obtain estimates of the rate of heterozygosity for 11 humans from diverse worldwide populations and through this analysis reveal the complex dependency of local sequencing coverage on the true underlying heterozygosity, which complicates the estimation of heterozygosity from sequence data. We show how we can use filters to correct for the confounding arising from sequencing depth. We find in practice that ratios of heterozygosity are more interpretable than absolute estimates and show that we obtain excellent conformity of ratios of heterozygosity with previous estimates from higher-coverage data.

HETEROZYGOSITY, or the fraction of nucleotides within an individual that differ between the chromosomes they inherit from their parents, is a crucial number for understanding genetic variation. Estimating this simple statistic from any type of sequence data is confounded by sequencing errors, mapping errors, and imperfect power for detecting polymorphisms. Obtaining an unbiased estimate is especially difficult for ancient genomes, where the sequences have a higher error rate, or in cases of low-coverage sequence data, where there is low power to detect heterozygous sites, or for hybrid capture where there may be additional biases due to the oligonucleotides used for fishing out sequences of interest.

Several methods for estimating individual heterozygosity have been proposed (Johnson and Slatkin 2006; Hellmann *et al.* 2008; Lynch 2008; Jiang *et al.* 2009; Haubold *et al.* 2010). For an overview of these methods see Haubold *et al.* (2010). Haubold *et al.* (2010) describe mlRho, an implementation of a method that jointly infers θ, the scaled mutation rate, and ρ, the scaled recombination rate for a shotgun-sequenced genome. However, they examined performance of their method at 10X coverage and a small sequence error rate of 4×10^{-4}, which is about four times lower than encountered currently in real data (Shendure and Ji 2008). We developed a method that estimates the heterozygosity for an individual of interest by leveraging the genome-wide joint information across sequence reads from a panel of individuals. Unlike previous methods that stochastically learn population allele frequencies (and thereby inform estimates of heterozygosity) in individuals from the same population (Kim *et al.* 2011; Li 2011), our model does not require any representative individuals from the same population for inference. The advantage of leveraging the panel of individuals in our method is that it enables learning of the empirical distribution of alleles at heterozygous and homozygous positions, a distribution that encapsulates sequencing errors and the non-Bernoulli sampling of each

doi: 10.1534/genetics.113.154500
Manuscript received June 17, 2013; accepted for publication July 30, 2013
[1]Corresponding author: Department of Genetics, Harvard Medical School, 77 Ave. Louis Pasteur, Boston, MA 02115. E-mail: kbryc@genetics.med.harvard.edu

allele at a heterozygous SNP. This allows one to disentangle the rate of heterozygosity from sequencing errors and other biases and does not require explicit modeling of these platform-, batch-, and genome-specific (frequently unknown and unestimatable) error processes. As a result of including the allele or genotype information at other individuals, our method gains robustness to any unknown error sources that may also be present within the data. Furthermore, our model allows for *any* relationship between the target individual and the panel and any relationship among the panel individuals, so long as the panel individuals share genetic polymorphisms with the target individual. The estimates from our method are not affected by choice of panel or cryptic relatedness among individuals.

We use an expectation–maximization (EM) algorithm to estimate the most likely distribution of counts across the unknown underlying genotypic states, from which we obtain an estimate of the proportion of loci that are heterozygous in the target individual. An advantage of this method is that it returns an unbiased and accurate estimate of heterozygosity even when the individual has low sequence coverage. Our method learns the distribution of alleles directly from the sequence read data and does not require modeling demographic relationships among the individuals or genotype calls from the sequence reads. We validate our EM method on 1 GB of simulated sequence data of 2X, 3X, 4X, 5X, 10X, 20X and 30X coverage and find that our method performs well at estimating the true heterozygosity even when the sequence error rate is extreme and mean coverage is low. As an empirical validation of the ability of our method to perform well on low-coverage data sets, we test our method on real high-coverage (30X) sequencing data, which we subsample to lower coverage, and verify that our estimates are consistent. In particular, we show that applying our method to a lower-coverage subsampling provides the same estimates of heterozygosity as those obtained on higher-coverage data, which are concordant with estimates of heterozygosity from other methods. We also show that our estimates do not depend on the choice of reference panel composition and that our estimates are consistent even when using unrelated population panels or relatives.

We apply our method to obtain estimates of heterozygosity for 11 individuals from many worldwide human populations, from Meyer *et al.* (2012). Our finding underscores the need to compare ratios of heterozygosity across fixed genomic regions to infer the relative rates of diversity among individuals.

Materials and Methods

We apply our method to read data at sites with a target minimum coverage (for example, ≥5X coverage) for the sequenced diploid individual of interest, aligned to some reference genome of known sequence. We also use sequence read data from n other individuals likewise aligned to the reference.

Let a be the unknown diploid genotype of our target individual at some position in the genome and c be the aligned reference allele. Then the allele distribution in other individuals will depend on $g = (a, c)$. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be the vector of alleles generated by taking one randomly sampled read from each of the n individuals. Let $\mathbf{w}$ be the observed alleles from the reads for our individual. Both $\mathbf{w}$ and $\mathbf{x}$ are observed quantities for a given position in the genome for our individual, and we are interested in modeling the joint probability of $\mathbf{w}$, $\mathbf{x}$ as the product of the marginal probabilities conditional on g.

We assume conditional independence of $\mathbf{w}$, $\mathbf{x}$ on the true unobserved genotype g. This assumption holds if the allele-frequency spectrum of the panel of individuals depends only on the true underlying genotypic state of our individual and not on the allele counts we observe, and likewise the allele count distribution depends only on the true underlying genotypic state and not on the alleles observed in the other individuals. From this conditional independence property, we then derive

$$P(\mathbf{w}, \mathbf{x} \mid g) = P(\mathbf{w} \mid g)\, P(\mathbf{x} \mid g) \tag{1}$$

$$P(\mathbf{w}, \mathbf{x}) = \sum_g P(g)\, P(\mathbf{w}, \mathbf{x} \mid g) \tag{2}$$

$$= \sum_g P(\mathbf{w} \mid g)\, P(\mathbf{x} \mid g)\, P(g) \tag{3}$$

$$P(\mathbf{w}, \mathbf{x}, g) = P(\mathbf{w} \mid g)\, P(\mathbf{x} \mid g)\, P(g), \tag{4}$$

which will later provide the leverage to infer the most likely values for the above probabilities, including $P(g)$, which gives us the genomic rate of heterozygosity.

For every site that has sufficient coverage in our individual and for which we have complete information of the panel, we add this site to the corresponding bin of observed alleles $\mathbf{w}$ and panel $\mathbf{x}$. This full matrix would be inconveniently large, so to simplify the data matrix of counts, we polarize our allele counts with respect to the reference, restricting to biallelic SNPs, which constitute the majority of sites. We denote the reference allele as 0 and allow only a single other variant per site, summarizing the observed alleles from the reads by the number of nonreference alleles. Thus, we denote genotypes as $g \in \{0, 1, 2\}$, which we refer to as the homozygous ancestral, heterozygous, and homozygous derived states, respectively. If, for example, we consider only sites with a coverage of 4, then $\mathbf{w} \in \{(4, 0), (3, 1), (2, 2), (1, 3), (0, 4)\}$. We can also easily represent $\mathbf{x}$ as a vector of 0's and 1's, referring to the reference or the nonreference allele present in the randomly sampled reads; for example, $\mathbf{x} = (0, 1, 1, 0, 1)$, where the length of $\mathbf{x}$ is determined by the number of individuals we sample.

We create a count matrix N of dimension $\|\mathbf{w}\| \times \|\mathbf{x}\|$, corresponding to the number of observed sites with each particular combination of $\mathbf{w}$ and $\mathbf{x}$. The counts of the

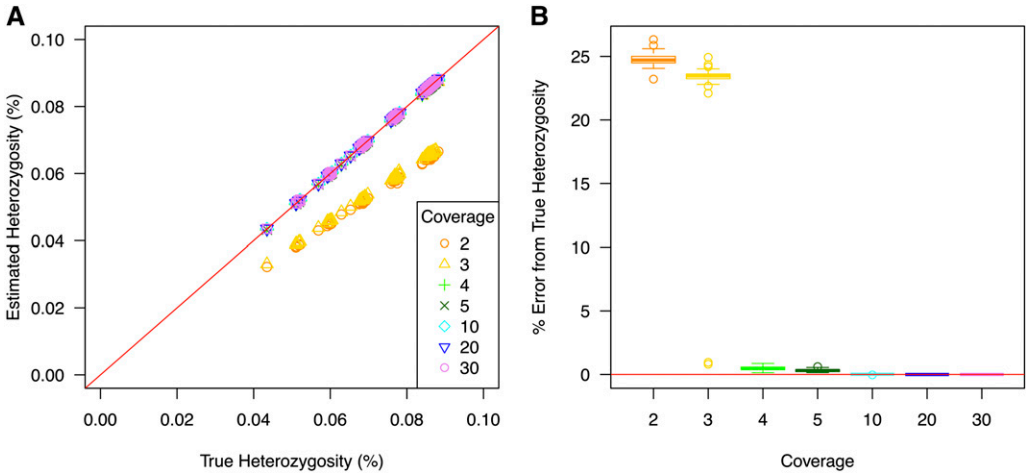

Figure 1 True *vs.* estimated rates of heterozygosity for 100 simulated read data sets. (A) Each data set has been downsampled to different coverage levels, denoted by symbol color and shape. The red line corresponds to true = estimated or perfect estimation of heterozygosity. (B) Run-by-run differences between true and estimated heterozygosity rates, stratified by downsampling coverage. The *y*-axis shows the percentage of error from the true value of heterozygosity.

numbers of loci where the individual is **w** and the panel of individuals comprises the alleles **x** are represented by the corresponding row and column entries in the matrix N.

From the matrix N we estimate the true values of $P(g)$, $P(\mathbf{w}|g)$, and $P(\mathbf{x}|g)$, using the EM algorithm. Let Y_{obs} be the observed counts of alleles in the matrix $N_{\mathbf{w},\mathbf{x}}$. Let Y_{mis} be $N_{\mathbf{w},\mathbf{x},g}$, the missing or unobserved counts of the alleles with the true parameter state g. Then the likelihood of the data is

$$\mathcal{L} = \sum_{\mathbf{w},\mathbf{x}} N_{\mathbf{w},\mathbf{x}} \log P(\mathbf{w},\mathbf{x}). \tag{5}$$

If the hidden variable, g, corresponding to the true underlying genotypic state were observed, the log-likelihood would be

$$\mathcal{L}' = \sum_{\mathbf{w},\mathbf{x},g} N_{\mathbf{w},\mathbf{x},g} \log P(\mathbf{w},\mathbf{x},g). \tag{6}$$

But this would require fitting $\|g\|$ different parameters per observed data point (*i.e.*, count entry of $N_{\mathbf{w},\mathbf{x}}$). This would require fitting three times as many parameters as there are data points. However, by relying on our conditional independence from Equation 4 above we can reduce the number of parameters to be fitted from the data.

By EM theory, the Q function $Q(P,\hat{P})$ is given by

$$\begin{aligned} Q\left(P,\hat{P}\right) &= E_{\text{post}} L'\left(\hat{P}\right) \\ &= \sum_{\mathbf{w},\mathbf{x},g} \hat{N}_{\mathbf{w},\mathbf{x},g} \log \hat{P}(\mathbf{w},\mathbf{x},g), \end{aligned}$$

where $\hat{N}_{\mathbf{w},\mathbf{x},g}$ is the expected value of $N_{\mathbf{w},\mathbf{x},g}$, which in our case derives from the multinomial distribution, under the posterior distribution calculated with the old parameters P.

The estimates for $\hat{P}$ that maximize Q, also derived from the maximum-likelihood estimates (MLEs) for the multinomial distribution, are

$$\hat{P}(\mathbf{w}|g) = \frac{\sum_{\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}{\sum_{\mathbf{w},\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}$$

$$\hat{P}(\mathbf{x}|g) = \frac{\sum_{\mathbf{w}} \hat{N}_{\mathbf{w},\mathbf{x},g}}{\sum_{\mathbf{w},\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}$$

$$\hat{P}(g) = \frac{\sum_{\mathbf{w},\mathbf{x}} \hat{N}_{\mathbf{w},\mathbf{x},g}}{N}.$$

Further, by Bayes' theorem this expands to

$$\begin{aligned} \hat{N}_{\mathbf{w},\mathbf{x},g} &= N_{\mathbf{w},\mathbf{x}} \cdot \frac{\hat{P}(\mathbf{w},\mathbf{x},g)}{\hat{P}(\mathbf{w},\mathbf{x})} \\ &= N_{\mathbf{w},\mathbf{x}} \cdot \frac{\hat{P}(\mathbf{w}\,|\,g)\hat{P}(\mathbf{x}\,|\,g)\hat{P}(g)}{\sum_z \hat{P}(\mathbf{w}\,|\,g)\hat{P}(\mathbf{x}\,|\,g)\hat{P}(g)}. \end{aligned}$$

By basic EM theory these reestimated values of $\hat{P}$ will generate a nondecreasing sequence of values for the log-likelihood $\mathcal{L}$. Finally, we obtain the parameter of interest $\hat{P}(g=1)$ after convergence.

Implementation

In practice, without constraining the parameters $\hat{P}(\mathbf{w}|g)$ we reach local but not consistently global likelihood maxima, which do not necessarily correspond to the genotypic state parameters we wish to obtain. To improve the ability of the EM to achieve maxima of genotypic states, we fit β-binomial distributions (effectively an overdispersed binomial distribution) to the probabilities of the number of nonreference alleles $P(\mathbf{w}|g)$ for each possible genotypic state g. Under this constraint, as well as the choice of reasonable starting parameters for the EM initialization, in practice, the EM consistently converges to a local maximum corresponding to the homozygous ancestral, heterozygous, and homozygous derived genotypic states.

Like those for the Beta-distribution, the MLEs for the Beta-binomial distribution do not have a closed form, although they can be found using direct numerical optimization (such as a fixed-point iteration or a Newton–Raphson iteration). However, instead, we estimate the two parameters

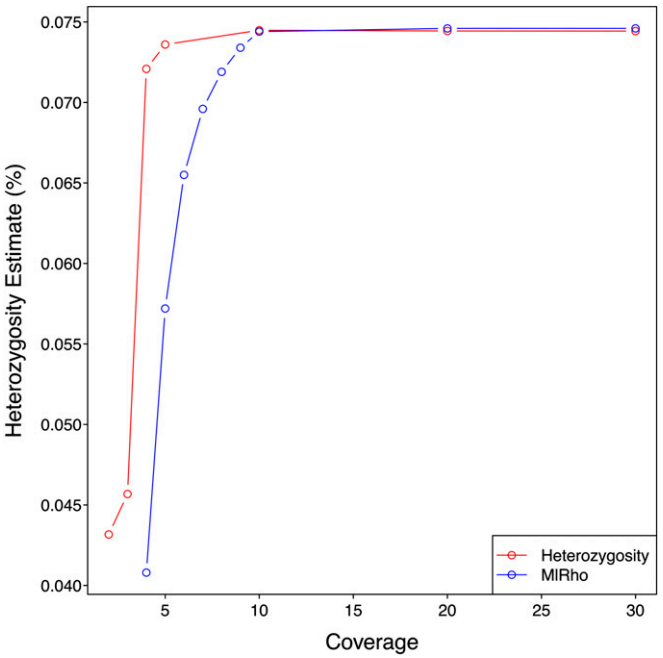

Figure 2 Our EM heterozygosity estimates (red) and MlRho estimates (blue) on the regions of a San individual genome sequenced to 30–45X and randomly downsampled. At higher coverage, both methods converge to an estimate of 7.45×10^{-4}. We note that our estimates for 4X and 5X coverage are much more accurate than those of MlRho. Results for <4X coverage were not possible to obtain from MlRho.

(α, β), using method-of-moments (MOM) estimators for the Beta-binomial, by setting

$$\hat{\alpha} = \frac{(n - \bar{x} - s^2/\bar{x})\bar{x}}{(s^2/\bar{x} + \bar{x}/n - 1)n}$$

$$\hat{\beta} = \frac{(n - \bar{x} - s^2/\bar{x})(n - \bar{x})}{(s^2/\bar{x} + \bar{x}/n - 1)n}.$$

In the case of underdispersed data, it is possible to obtain MOM estimates that are invalid. Although unlikely to occur in the read data, for this contingency, we instead fit a binomial distribution to the data.

There are several challenges in implementing the EM for our problem. The first is that, as with all likelihood calculations, our probabilities approach very small numbers. To avoid numerical error due to underflow of small likelihoods and parameter estimates, we implement the EM, storing all probabilities and likelihoods in the log form.

When any of the parameters we are interested in estimating approach 0, then the probabilities become numerically unstable and may have underflow issues (or in log space, overflow issues). To avoid this situation, we add a "prior" ε to the likelihood calculation, which adds a small count value in the step calculating the parameters to avoid probabilities reaching 0. This is a standard approach using "pseudocounts" for an EM, which also avoids an ill-defined likelihood calculation involving p^x when both $p \to 0$ and $x \to 0$. In implementing these pseudocounts, we calculate the posterior

$$L' = \hat{N}_{p,x,z} \log(P(p,x,z)) + L'$$

rather than the maximum-likelihood estimation; hence, we obtain a maximum *a posteriori* (MAP) estimate, which is a Bayesian method that incorporates a prior over the distribution to be estimated (in this case, a small uniform prior). We choose a small prior (less than in total counting one site across all possible matrices) that does not affect our estimates. In general, our estimates are robust to choice of this prior, within a range examined of 1×10^{-10} to 1×10^{-50}, and we continue to refer to our method as an EM implementation although in fact we use a non-MLE method. In effect, our equations for each step remain the same, except that in the M-step of the EM (where we estimate the parameters) we instead estimate the MAP using the prior. Specifically, we estimate

$$\text{Posterior} = \sum_{i,j} \left(N_{i,j} \cdot \log\left(\sum_z P_{i,j,z} \right) + \varepsilon \cdot \sum_z \log(P_{i,j,z}) \right).$$

In practice, we set ε to 1×10^{-20}, which does not alter estimates of the probabilities while preventing numerical instability issues.

Finally, likelihood maximization occurs on an arbitrary base, so to avoid numerical issues due to any remaining underflow of the likelihood calculation, we compute a factor F at the start of the EM. For each iteration, we compute the likelihood of the data minus this constant factor, which is a standard practice and does not affect the computation of the maximum. This is equivalent to calculating the log odds

$$\begin{aligned} L &= \left(\sum_{i,j} N_{i,j} \cdot \log(P_{i,j}) \right) - (F) \\ &= \left(\sum_{i,j} N_{i,j} \cdot \log(P_{i,j}) \right) - \left(\sum_{i,j} N_{i,j} \cdot \log(F_{i,j}) \right) \\ &= \sum_{i,j} N_{i,j} \cdot \left(\log(P_{i,j}) - \log(F_{i,j}) \right) \\ &= \sum_{i,j} N_{i,j} \cdot \log\left(\frac{P_{i,j}}{F_{i,j}} \right) \end{aligned}$$

for some constants $F_{i,j}$. In practice, we set $F_{i,j}$ to be the likelihood at initialization of the EM. We then iterate the EM until both the change in parameters and the change in the likelihood are smaller than our chosen threshold, which in practice we set as 1×10^{-50}.

It should be noted that any form for tallying read counts may be used, including the allele profile used in Haubold *et al.* (2010), other summaries of data such as number of derived reads, or genotype calls should they be available; our choice was motivated by a choice of dimensionality that is a compromise between simplicity and capturing relevant information. Our method is highly generalizable to any

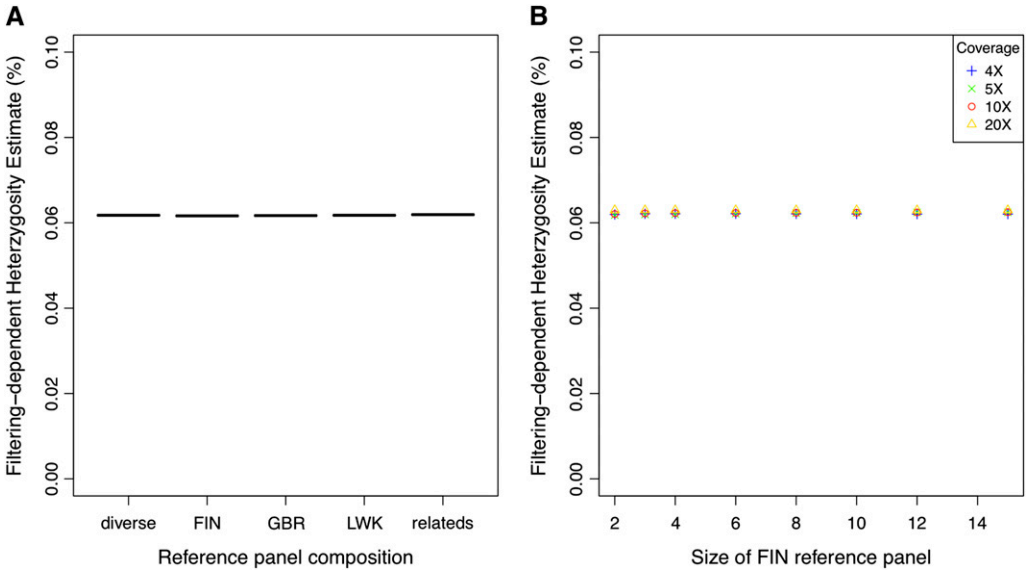

Figure 3 Estimates of heterozygosity for CEU trio individual NA12892, using a variety of reference panel compositions. (A) Heterozygosity estimates for each of the reference panels composed of five individuals from different populations as described in *Proof of principle 2: Downsampling high-coverage genomes*. (B) Heterozygosity estimates using different reference panel size (*x*-axis) and downsampled to different coverage (symbol color/shape).

choice of count data and could be implemented assuming that a reasonable starting position for the EM could be proposed, such that the iterations are likely, although not guaranteed, to converge to a local maximum corresponding to the genotypic states. Furthermore, the simple framework of our method allows for future directions, such as incorporating quality scores into the summaries of the data, which may result in better estimates and also inform the reliability of quality scores.

Proof of principle 1: Application to simulated data

We generated simulated sequence data and applied our EM method for estimating heterozygosity to assess the accuracy of our estimation procedure.

Generating coalescent simulations: We generated 100 replicate data sets of sequence data, using MaCS (Chen *et al.* 2009). Each replicate data set contains 10 independent regions of length 100 Mb for a total of 1 GB of sequence, for (each haploid) one chimpanzee chromosome, seven African chromosomes, five European chromosomes, and five East Asian chromosomes, using demographic parameters fitted by Gutenkunst *et al.* (2009). We include a chimpanzee chromosome assuming a constant ancestral population size of 50,000 individuals and a split time from humans of 6 MYA, using the same generation time as humans of 25 years per generation.

Adding simulated error: We simulate sequence data from the true genotypes by adding errors to reads. First, for all variable loci in the target individual, we randomly choose which allele is on a read and then add errors to each read (with a high error rate of 0.002) to generate the total number of derived reads for the individual at the locus from the total sequencing depth. For each other sequenced chromosome, we add errors with a lower error rate of 0.0001 (since we assume the panel is composed of higher-quality genomes) and then add a count for the final simulated locus in the appropriate hash bin. Finally, for each invariant locus, we add errors to the target individual's reads and, at a lower rate, add errors to the other sequenced chromosomes and input these counts into the hash bin. With an error rate of 0.001 we add errors to the chimpanzee chromosome, which inverts the ancestral and derived reads.

Proof of principle 2: Downsampling high-coverage genomes

To assess the efficacy of our method at lower coverages on real sequence data, we begin by obtaining estimates of heterozygosity for a San individual from the Human Genome Diversity Project (HGDP), sequenced to higher coverage, using Illumina's Genome Analyzer IIx next-generation sequencing technology, which we then downsample to varying levels of low coverage. We use this data set of sequence reads to explore the ability of our method to perform on low-coverage sequence data and the lower bound of coverage at which we are able to obtain accurate estimates of heterozygosity. We compare the performance of our method to the estimate of θ obtained from MlRho (Haubold *et al.* 2010).

We also examine the robustness of our method to a variety of reference panel compositions, examining performance for panels from different populations, panels of different size, and panels including known relatives of the target individual. We estimate the heterozygosity of one parent from a HapMap CEU (Utah residents with Northern and Western European ancestry) trio (NA12892) sequenced by the 1000 Genomes Project. We compare estimates at 10X coverage, using a reference panel of (a) FIN, 5 Finnish individuals; (b) GBR, 5 British individuals from England and Scotland; (c) LWK, 5 Luhya from Webuye, Kenya; (d) diverse, 2 FIN, 2 GBR, and 1 LWK individual; and (e) relateds, the target individual's trio and a FIN, a GBR, and an LWK individual. We also examined the effect that changing the size of the reference panel had on power and the

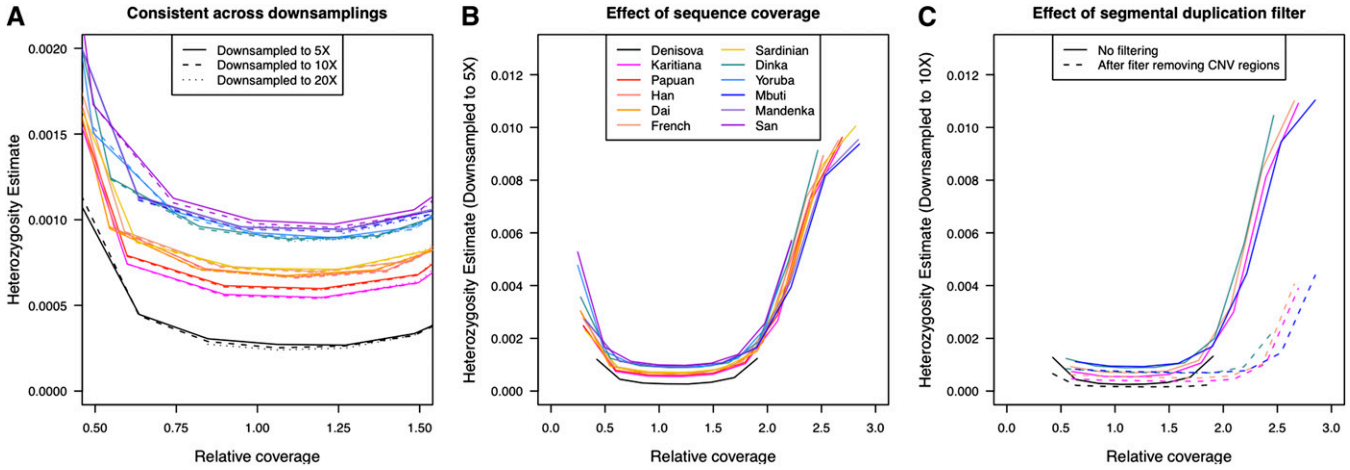

Figure 4 Estimates of heterozygosity for each of the 11 present-day human genomes and Denisova, where each individual is denoted by a unique color. Relative coverage is defined as the lower bound of the sequencing bin, divided by the mean sequencing depth for the individual. (A) Heterozygosity estimates are consistent across downsampling levels. Downsampling to 5X, 10X, and 20X levels is denoted by line type. Each individual is denoted by line color. (B) All individuals show an increase in estimated heterozygosity at higher (and lower) relative coverage. (C) Effect of removing known regions with segmental duplications. Estimates of heterozygosity are shown for a sample of five of the individuals. Without filtering, estimates for each bin are shown with solid lines. After exclusion of regions within known copy-number-variable and segmental duplications, the heterozygosity estimates display a flatter distribution (dashed lines).

lower limit of convergence by calculating heterozygosity across a range of reference panel sizes from 2 to 15 FIN individuals.

Application to 11 worldwide human genomes

We align sequence data from 11 human genomes from worldwide populations and an archaic Denisovan genome to the chimpanzee reference genome to avoid introducing human-reference population biases. For details on the populations, the samples, and the sequencing performed, see Meyer *et al.* (2012). We generate a counts matrix for each of the genomes, using a panel generated by a single read sampled from each of the other 11 genomes.

We include only sites where there is a chimpanzee reference allele and exclude sites where two or more nonreference bases are equally present or if there are more than five reads showing a third (nonvariant and nonreference) allele. We also exclude CpG sites, as well as sites where any individual from the panel has no coverage or sites that have insufficient coverage for the target individual.

To demonstrate the relationship between sequencing coverage and the true rate of heterozygosity of different regions, we generate count data for each bin of 5X coverage ranging between 5X and 50X, where for each bin data set, we include only sites where the coverage of the individual falls within the target range. We downsample coverage at each bin (where possible) to 5X, 10X, and 20X and compare results stratified by downsampling, as well as by genomic coverage.

Finally, we produce estimates for the 11 present-day genomes and the archaic Denisova genome on a fixed set of sites and compare them to previous estimates for these samples (Meyer *et al.* 2012).

Results

Simulation results

We obtain accurate estimates of heterozygosity across a variety of coverage levels (from 2X and 30X) (see Figure 1). We note a tiny bias of 0.3% (in relative terms) from the true rate for 5X coverage read data, but with higher coverage this bias goes to zero. We note that performance deteriorates sharply below 4X, as a result of failure of convergence of the EM to a maximum corresponding to genotypic states. At 2X and 3X coverage, the size of the data matrix is not significantly greater than the number of parameters being estimated, and the data do not have sufficient information to allow the EM to consistently converge to an optimum corresponding to the genotypic states. It is possible to consider multiple initializations to increase the likelihood of converging to a maximum corresponding to genotypic states, but since the EM is not consistently robust to this choice at 2X and 3X coverage, we do not present these results.

Downsampling results

Figure 2 illustrates that our EM estimation method and MlRho give consistent estimates of heterozygosity for the HGDP San individual starting at ~10X coverage and higher. However, at lower coverage (4–10X) our method significantly outperforms MlRho, giving a nearly convergent estimate, while mlRho does not.

Reference panel results

We observe consistent heterozygosity estimates independent of reference panel size and composition and relatedness to the target individual, confirming our method's independence of reference panel composition (Figure 3).

We note that there is a small correction of the bias seen when using lower-coverage data with larger reference panels (Figure 3B), but this effect is slight.

Heterozygosity estimates for 11 present-day and Denisovan genomes

We present our initial estimates of heterozygosity, downsampled to three different depths, for each sequencing coverage bin (normalized by individual mean sequence coverage) in Figure 4A. Our estimates of heterozygosity are consistent and independent of count matrix (*i.e.*, downsampling) size, as would be expected from our simulated downsampling results shown in Figure 2. However, we find a strong signal that the estimates of heterozygosity are correlated to sequencing coverage of the region. We note that this is not an artifact of the larger amount of data available at higher coverage, since each bin is calculated after being downsampled to the same depth. Instead, the U-shaped curves in Figure 4B indicate that the apparent next-generation sequencing coverage is dependent on properties of the underlying genomic sequence. In particular, we find that regions of lower coverage and higher coverage (relative to the mean sequencing depth) show higher heterozygosity.

We witnessed increased heterozygosity at regions of higher coverage, which we suspected was due at least in part to artifactual genetic diversity due to cryptic segmental duplications. To explore this hypothesis, we restricted our analyses to regions of the genome that have been identified as unlikely to contain segmental duplications, available on the Eichler Laboratory website (http://eichlerlab.gs.washington.edu/database.html). We find that this filter strongly reduces the effects (Figure 4C), confirming that unidentified segmental duplications, which result in a net higher sequencing coverage of the region, result in a high estimate of heterozygosity for such regions. Removing these regions with known segmental duplications reduces this effect at regions with higher sequencing coverage. However, the increase in heterozygosity at higher coverage still is present even after this correction (see Figure 4C), suggesting that this filter, while helpful, does not completely solve the problem.

Using only data that passed the segmental duplication filter, we obtain estimates for the sequenced genomes on the same set of regions, restricting to regions with sequencing coverage between 20X and 40X. Using the EM, we estimate the total genome-wide fraction of heterozygosity for each individual, and we also can extract estimates of the allelic distribution of heterozygous and homozygous sites (Figure 5). We present the absolute estimates we obtain in Table 1, as well as the ratio of heterozygosity in the Denisova genome relative to the other individuals. We find the highest estimates of heterozygosity for the San African individual and the next-highest estimates of heterozygosity for other African individuals from the Mandenka, Yoruba, Mbuti, and Dinka populations. The next-highest levels of heterozygosity are in individuals from European populations (French, Sardinian), followed by East Asian populations (Dai, Han). We find the lowest estimates of heterozygosity in the individuals from Melanesia (Papuan) and from a Native American population (Karitiana).

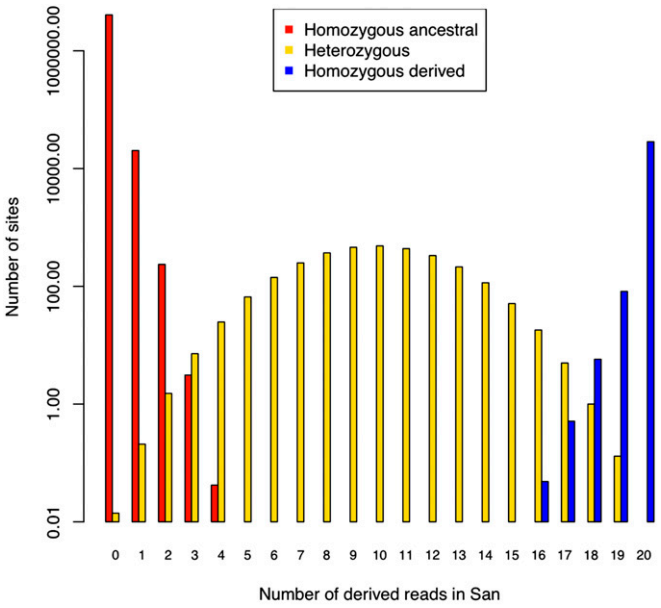

Figure 5 Inferred distribution of homozygous ancestral (red), heterozygous (yellow), and homozygous derived (blue) sites for the San HGDP individual. The *y*-axis is presented on a log scale, and counts with expected value <0.1 have been omitted from the plot.

Discussion

We have shown that our heterozygosity estimation method both performs well in low-coverage simulated sequence data and provides consistent estimates on real low-coverage data downsampled from higher coverage. In particular, our method outperforms other methods on data that have been sequenced at <10X coverage and provides reasonable estimates for as low as 4X coverage. Our method does not assume any relationship between the target individual and the reference panel individuals, making it useful even in situations where there are no other sequenced individuals from the same population or when the population is unknown.

Our estimates for 11 worldwide human genomes and the archaic Denisovan genome provide important insights into the distribution of heterozygosity across human populations. Furthermore, our results show that estimates of heterozygosity are strongly affected by genomic properties such as copy-number variability, and these properties affect sequencing coverage. Hence, we show that the heterozygosity is not independent of sequencing coverage even within one genome and is elevated in both regions with low coverage (relative to the mean sequencing depth) and regions with high coverage. This is an unexpected result if one assumes

Table 1 Heterozygosity estimates for the 11 present-day individuals from worldwide populations and Denisova

Individual	Heterozygosity estimate (%)	Ratio (%)
Denisova	0.0165	—
San	0.0721	23
Mandenka	0.0686	24
Yoruba	0.0649	25
Mbuti	0.0657	25
Dinka	0.0635	26
Sardinian	0.0490	34
French	0.0473	35
Dai	0.0465	35
Han	0.0454	36
Papuan	0.0386	43
Karitiana	0.0353	47

The ratio presented is the relative heterozygosity in the Denisova genome as a percentage of that found in the present-day individual.

a "Lander–Waterman" Poisson distribution of read depth (Lander and Waterman 1988; Weber and Myers 1997). Furthermore, even after excluding regions with known copy-number variants, an increase in heterozygosity is still present at the more extreme levels of sequence coverage, suggesting that other correlations of sequence diversity with coverage, or possibly individual-specific segmental duplications, still remain. Implications from our results suggest that using the higher tail of sequencing coverage for population genetic inference may result in a biased set of genomic regions with selectively higher heterozygosity, possibly due to population and individual segmental duplications.

Our absolute estimates of heterozygosity are lower than those reported for these genomes in other articles using other methods (Meyer *et al.* 2012); however, the *relative* estimates are consistent with those that have been previously documented. Because the regions of the genome that pass our filters are likely to be lower in complexity and substantially biased toward lower diversity due to alignment biases, the lower absolute values of heterozygosity are expected. However, our relative heterozygosity estimates are consistent with previously documented levels of genetic diversity, with African populations showing the highest levels of diversity and with decreasing levels with distance away from Africa (Jakobsson *et al.* 2008; Li *et al.* 2008). Our estimates confirm previous findings that the archaic Denisovan genome shows substantially lower levels of heterozygosity than any of the other present-day populations, with only a fraction of the rate of heterozygosity.

Other likelihood methods that study the allele-frequency spectrum from low-coverage sequence data are not well suited to our context of a single individual from a different or an unknown population (Kim *et al.* 2011; Li 2011). However, an interesting future avenue would be to explore the performance of these methods outside their intended scope, testing for robustness to population substructure among samples (Kim *et al.* 2011) or applicability to inferring heterozygosity for one individual across the whole genome rather than for the conditioned site-frequency spectrum (Li 2011).

More generally, we emphasize that absolute heterozygosity is not a well-defined quantity in the analysis of genomic data, as it strongly depends on the particular filters that are used to select the regions being analyzed and may be an implausible concept in highly repetitive regions (such as centromeres and telomeres) and copy-number-variable regions. The absolute value of heterozygosity can vary based on the regions chosen to be examined, but the relative heterozygosity estimates or ratios among individuals (using the same regions and filters) are consistent. Hence, in practice, heterozygosity estimates are most meaningful when viewed as relative ratios among individuals for the same regions of the genome and not as absolute values inherent to diploid genomes.

Acknowledgments

K.B. gratefully acknowledges that this investigation was supported by the National Institutes of Health (NIH) under Ruth L. Kirschstein National Research Service Award 5F32HG006411. This work was also supported by National Science Foundation grant 1032255 and NIH grant GM100233.

Literature Cited

Chen, G., P. Marjoram, and J. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. 19: 136–142.

Gutenkunst, R., R. Hernandez, S. Williamson, and C. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695.

Haubold, B., P. Pfaffelhuber, and M. Lynch, 2010 mlRho—a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. Mol. Ecol. 19: 277–284.

Hellmann, I., Y. Mang, Z. Gu, P. Li, M. Francisco *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Res. 18: 1020–1029.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451: 998–1003.

Jiang, R., S. Tavaré, and P. Marjoram, 2009 Population genetic inference from resequencing data. Genetics 181: 187–197.

Johnson, P., and M. Slatkin, 2006 Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res. 16: 1320–1327.

Kim, S. Y., K. 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.

Lander, E., and M. Waterman, 1988 Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231–239.

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, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.

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

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo *et al.*, 2012 A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222–226.

Shendure, J., and H. Ji, 2008 Next-generation DNA sequencing. Nat. Biotechnol. 26: 1135–1145.

Weber, J., and E. Myers, 1997 Human whole-genome shotgun sequencing. Genome Res. 7: 401–409.

Communicating editor: J. Wall