Published in final edited form as:

Genet Epidemiol. 2017 July ; 41(5): 375–387. doi:10.1002/gepi.22048.

PhredEM: A Phred-Score-Informed Genotype-Calling Approach for Next-Generation Sequencing Studies

Peizhou Liao[1], **Glen A. Satten**[2], and **Yi-Juan Hu**[1]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia

[2]Centers for Disease Control and Prevention, Atlanta, Georgia

Abstract

A fundamental challenge in analyzing next-generation sequencing data is to determine an individual's genotype accurately, as the accuracy of the inferred genotype is essential to downstream analyses. Correctly estimating the base-calling error rate is critical to accurate genotype calls. *Phred* scores that accompany each call can be used to decide which calls are reliable. Some genotype callers, such as GATK and SAMtools, directly calculate the base-calling error rates from *phred* scores or recalibrated base quality scores. Others, such as SeqEM, estimate error rates from the read data without using any quality scores. It is also a common quality control procedure to filter out reads with low *phred* scores. However, choosing an appropriate *phred* score threshold is problematic as a too-high threshold may lose data while a too-low threshold may introduce errors. We propose a new likelihood-based genotype-calling approach that exploits all reads and estimates the per-base error rates by incorporating *phred* scores through a logistic regression model. The approach, which we call PhredEM, uses the Expectation-Maximization (EM) algorithm to obtain consistent estimates of genotype frequencies and logistic regression parameters. It also includes a simple, computationally efficient screening algorithm to identify loci that are estimated to be monomorphic, so that only loci estimated to be non-monomorphic require application of the EM algorithm. Like GATK, PhredEM can be used together with a linkage-disequilibrium-based method such as Beagle, which can further improve genotype calling as a refinement step. We evaluate the performance of PhredEM using both simulated data and real sequencing data from the UK10K project and the 1000 Genomes project. The results demonstrate that PhredEM performs better than either GATK or SeqEM, and that PhredEM is an improved, robust and widely applicable genotype-calling approach for next-generation sequencing studies. The relevant software is freely available.

Keywords

common variant; EM algorithm; rare variant; read data

Address for Correspondence: Yi-Juan Hu, Ph.D., Department of Biostatistics and Bioinformatics, Emory University, 1518 Clifton Rd NE, Atlanta, Georgia 30322, Phone: (404) 712-4466, Fax: (404) 727-1370, yijuan.hu@emory.edu.

DISCLAIMER

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

INTRODUCTION

The recent advancement of next-generation sequencing (NGS) technologies and the rapid reduction of sequencing costs have led to extensive use of sequencing data in disease association studies and population genetic studies [Ng et al., 2010; The 1000 Genomes Project Consortium, 2015]. However, it is still difficult and costly to perform whole-genome sequencing (WGS) at high depth in large cohorts [Sims et al., 2014]. Instead, many studies have adopted whole-exome sequencing (WES) [The 1000 Genomes Project Consortium, 2015; Muddyman et al., 2013]. Despite the high average depth that is typically attainable in WES studies, some regions within a gene may still have much lower depth than the average due to the inefficiency of exome capture technologies [Do et al., 2012]. Other studies have kept the design of WGS but have chosen low or moderate depths. For example, the UK10K project (www.uk10k.org) sequenced the two population cohorts genome wide at depth of ~6x. Although sequencing costs are declining, we anticipate that many NGS studies will continue to employ WES or WGS with low or medium depth for some time to come.

A fundamental challenge in analyzing NGS data is to determine an individual's genotype correctly, as the accuracy of the inferred genotype is essential to downstream analyses. It is difficult to call genotypes for two reasons. First, NGS data can suffer from errors introduced in the base-calling process. The base-calling error rate ranges from a few tenths of a percent to several percent [Nielsen et al., 2011], can vary from base to base as a result of machine cycle and sequence context [Kircher et al., 2009], and also varies dramatically across different sequencing platforms. For instance, the Illumina MiSeq platform has an error rate of ~0.8% [Quail et al., 2012] whereas the Roche 454 System has ~0.1% error rate [Liu et al., 2012]. Second, the quality of called genotypes depends heavily on the read depth. Genotypes covered by many reads can typically be called reliably. However, when a locus is covered by only a few reads, genotype calling is challenging because minor allele reads are indistinguishable from sequencing errors.

All major sequencing platforms assign each called base of a raw sequence a *phred* score,[3] a widely-accepted measure of the probability that the base is called incorrectly [Ewing et al., 1998; Ewing and Green, 1998]. *Phred* scores are determined using various predictors of possible errors such as peak spacing, uncalled/called peak ratio and peak resolution. Nominally, the *phred* score is defined as

$$Q=-10\log_{10}\Pr(\text{observed allele} \neq \text{true allele}) \quad (1)$$

so that, for example, $Q = 30$ nominally corresponds to a 0.1% error rate. Despite their widespread use, *phred* scores may not accurately reflect the true error rates in base calling because they fail to account for some important factors. For instance, the specific error pattern inherent in each nucleotide base (i.e., A, C, T and G) is not considered in *phred* scores [Li et al., 2004]. Additionally, *phred* scores do not account for the position of the base within a read [DePristo et al., 2011]. Since *phred* scores might be inaccurate representations of true base-calling error rates, methods have been developed to recalibrate base quality scores, such as the base quality score recalibration (BQSR) option in GATK [DePristo et al., 2011] and the base alignment quality (BAQ) option in SAMtools [Li, 2011]. However, the effectiveness of recalibration highly depends on whether all important error predictors (e.g., machine cycle and dinucleotide context) are included in the recalibration model. In addition, the recalibration process can be computationally intensive [Yu et al., 2015].

A genotype-calling method generally uses a probabilistic framework, combining base-calling error rates and a marginal (population-level) distribution of genotype frequencies to provide an individual-level probability for each genotype [Mckenna et al., 2010; Li et al., 2009a; Martin et al., 2010]. Because the error rate is critical in probabilistic genotype-calling algorithms, it is crucial that it be correctly specified, especially when sequencing depth is low to moderate. GATK uses error rates that are calculated directly from *phred* scores or recalibrated scores by applying equation (1), neither of which is precisely correct. SAMtools obtains an error rate from the minimum of the *phred*-based error rate and the mapping error rate, so that the error rate is always adjusted downwards [Li, 2011]. In addition, bases with low *phred* scores (e.g., $Q < 20$ or 30) are typically filtered out as part of quality control (QC) procedures. However, choosing a threshold for *phred* scores always involves a tradeoff: high thresholds may result in loss of useful information by eliminating bases that are correctly called, while low thresholds leave a large number of erroneously-called bases in the data, leading to false-positive variant calls.

Instead of relying on *phred* scores, Martin et al. [2010] proposed SeqEM, a genotype-calling algorithm that estimates the error rate using the read data itself. However, the fundamental assumption of SeqEM that, at each locus, there is a uniform error rate for each read is generally not true, given the considerable variability in error rates implied by the variability in *phred* scores. Because SeqEM ignores *phred* scores entirely, the valuable information about errors encoded in *phred* scores is lost.

In this paper, we propose a new genotype-calling approach which estimates base-calling error rates from the read data while incorporating the information in *phred* scores. We model an error rate as a logistic function of a *phred* score. The logistic regression model is readily integrated into a modification of the SeqEM likelihood which allows for a base-specific error probability. Like SeqEM, our approach also uses the Expectation-Maximization (EM) algorithm [Dempster et al., 1977]. Information from all individuals is used to estimate the unknown genotype frequencies and logistic regression parameters. We compute the probability of each latent genotype for each individual based on parameter estimates and use the empirical Bayes approach to assign the most likely genotype to each individual. We show that the logistic model fits real sequencing data well, and that the unknown parameters in our likelihood are consistently estimated. Because we allow separate logistic regression parameters at each locus, error predictors that are the same for all bases at a given locus (e.g., dinucleotide context) are automatically accounted for, as in SeqEM.

To minimize the effort of calling genotypes for the large majority of loci that are estimated to have no variation, we develop a simple, computationally efficient screening algorithm to identify loci that are estimated to be monomorphic and therefore do not require parameter estimation using the EM algorithm. Furthermore, we show that our approach can be used together with a linkage-disequilibrium (LD)-based method such as Beagle to improve genotype calling. Finally, we demonstrate through simulation studies and by comparison to gene array data that our approach is more accurate than both SeqEM and GATK (even after *phred* score recalibration). We illustrate our new approach through an application to two real sequencing datasets, one from the UK10K project and the other from the 1000 Genomes project.

METHODS

We first consider one biallelic locus at a time. For the *i*-th individual, let G_i denote the underlying true genotype (coded as the number of minor alleles), T_i denote the total number of alleles that are mapped to the locus, and R_i ($R_i \leq T_i$) denote the number of mapped alleles that are called to be the minor allele. The *phred* scores are represented by $\boldsymbol{Q}_i = (Q_{i1}, \ldots, Q_{iT_i})'$, where Q_{ik} is the *phred* score associated with the *k*-th called allele and the prime ($'$) indicates the transpose of a vector. At each locus, values of T_i, R_i, and $\boldsymbol{Q}_i$ can be easily extracted from the pileup files produced by SAMtools. Let ε_{ik} be the true base-calling error rate of the *k*-th allele. We relate ε_{ik} to Q_{ik} through the logistic regression model

$$\log\left(\frac{\varepsilon_{ik}}{1-\varepsilon_{ik}}\right)=\beta_0+\beta_1 Q_{ik}, \qquad (2)$$

where β_0 and β_1 are unknown regression parameters that are locus specific. Let $\boldsymbol{\theta} = (\beta_0, \beta_1)'$ and $\varepsilon_{ik}(\boldsymbol{\theta}) = \exp(\beta_0 + \beta_1 Q_{ik})/\{1 + \exp(\beta_0 + \beta_1 Q_{ik})\}$. Equation (2) is motivated by the fact that the *phred* score is a highly informative predictor of the base-calling error, even though (1) does not hold in the exact sense. In the Results section, we demonstrated that the logistic model fits the real sequencing data well.

Without loss of generality, we order the T_i alleles so that the first R_i alleles are called to be the minor allele and the rest the major allele. Assuming that the errors of the T_i alleles are independent of each other, the probability of observing R_i copies of the minor allele out of T_i alleles can be described as a sequence of independent Bernoulli trials. Specifically, given the true genotype G_i, the total number of alleles T_i, and the *phred* scores $\boldsymbol{Q}_i$, the probability of observing R_i is written as

$$P_{\boldsymbol{\theta}}(R_i|G_i,T_i,\boldsymbol{Q}_i)=\begin{cases}\prod_{k=1}^{R_i}\varepsilon_{ik}(\boldsymbol{\theta})\prod_{k=R_i+1}^{T_i}\{1-\varepsilon_{ik}(\boldsymbol{\theta})\} & G_i=0\\(0.5)^{T_i} & G_i=1\\\prod_{k=1}^{R_i}\{1-\varepsilon_{ik}(\boldsymbol{\theta})\}\prod_{k=R_i+1}^{T_i}\varepsilon_{ik}(\boldsymbol{\theta}) & G_i=2.\end{cases} \qquad (3)$$

Suppose that the sample consists of *n* unrelated individuals. Then the likelihood function takes the form

$$L_o(\boldsymbol{\theta}, \boldsymbol{\pi})=\prod_{i=1}^{n}\sum_{g=0,1,2} P_{\boldsymbol{\theta}}(R_i|g, T_i, \boldsymbol{Q}_i)P_{\boldsymbol{\pi}}(g), \quad (4)$$

where $P_{\boldsymbol{\pi}}(g)$ is the genotype frequency characterized by $\boldsymbol{\pi}$. Under Hardy-Weinberg Equilibrium (HWE), $\boldsymbol{\pi}$ consists of a single parameter π for the minor allele frequency (MAF). Then, $P_{\boldsymbol{\pi}}(0) = (1 - \pi)^2$, $P_{\boldsymbol{\pi}}(1) = 2\pi(1 - \pi)$, and $P_{\boldsymbol{\pi}}(2) = \pi^2$. Under Hardy-Weinberg Disequilibrium (HWD), $\boldsymbol{\pi} = (\pi, f)'$ where π and f are the MAF and the fixation index F_{st}, respectively. Then, $P_{\boldsymbol{\pi}}(0) = (1 - f)(1 - \pi)^2 + f(1 - \pi)$, $P_{\boldsymbol{\pi}}(1) = 2\pi(1 - \pi)(1 - f)$, and $P_{\boldsymbol{\pi}}(2) = (1 - f)\pi^2 + f\pi$.

The proposed likelihood is closely related to several existing methods. When $\beta_1 = 0$, the error rate is independent of the *phred* score, and expression (4) reduces to the likelihood of SeqEM. When $\beta_0 = 0$, $\beta_1 = -\log(10)=10$ and $\boldsymbol{\varepsilon}$ is small, expression (2) is approximately equal to (1), and our model reduces to the Bayesian genotyper implemented in GATK. However, our likelihood fully exploits the read data and the *phred* scores, both of which could improve genotype-calling accuracy. Note that it is not necessary to filter out low-quality alleles, which still provide some information about $\boldsymbol{\theta}$. Because our model uses the read call data to adjust the relationship between *phred* scores and the error rate at each locus, it can be considered as a kind of *phred* score recalibration, except that the recalibration is done simultaneously with fitting other parameters to best fit the observed data. Like other multi-sample calling methods, our method also estimates the genotype frequencies and regression parameters by utilizing information across all individuals in the sample.

We may obtain the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ by maximizing the likelihood (4) via the EM algorithm described in the Appendix. However, if a locus has little variability (e.g., a monomorphic locus, singleton or doubleton) so that there are very few reads for the minor allele in the study sample, the MLE of β_1 based on (4) may be unreliable [Firth, 1993]. To improve stability, we propose to modify the MLE of β_1 by leveraging information from other loci. Specifically, we introduce a Gamma distribution $\Gamma(-\beta_1;\ \kappa, \phi)$ as a penalty (or prior) for $-\beta_1$, where κ and ϕ are the shape and scale hyper-parameters, respectively. We first use the method of moments to obtain estimates $\hat{\kappa}$ and $\hat{\phi}$ based on the MLEs of β_1 from a set of loci that are either all or mostly estimated to be monomorphic; for loci that are estimated to be monomorphic, all reads for the minor allele can be treated as errors, and ordinary logistic regression can be used to estimate $\boldsymbol{\theta}$ at each locus. For genome- or exome-wide data, any region can be used as most loci are estimated to be monomorphic; the full EM algorithm only needs to be run for the few loci that are estimated to be polymorphic. We then obtain the maximum penalized likelihood estimators (MPLEs) by maximizing the penalized likelihood

$$L_o^*(\boldsymbol{\theta}, \boldsymbol{\pi})=\Gamma(-\beta_1;\hat{\kappa}, \hat{\phi})L_o(\boldsymbol{\theta}, \boldsymbol{\pi}). \quad (5)$$

Note that the MPLEs are asymptotically equivalent to the MLEs, as the Gamma penalty becomes negligible when the sample size *n* grows.

Denote the MPLEs by $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\theta}}$. We can estimate the probability distribution of the true genotype G_i for the *i*-th individual from their read count data T_i and R_i and their *phred* scores $\boldsymbol{Q}_i$ using the formula

$$\Pr(G_i=g|R_i,T_i,\boldsymbol{Q}_i;\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\pi}})=\frac{P_{\hat{\boldsymbol{\theta}}}(R_i|g,T_i,\boldsymbol{Q}_i)P_{\hat{\boldsymbol{\pi}}}(g)}{\sum_{g'=0}^{2}P_{\hat{\boldsymbol{\theta}}}(R_i|g',T_i,\boldsymbol{Q}_i)P_{\hat{\boldsymbol{\pi}}}(g')}, \quad (6)$$

for $g = 0$, 1 and 2. At a single locus, genotype calls can be made by assigning each individual the genotype that their data assigns the highest estimated probability. Individuals with no read covering the locus are not assigned any genotype. Because the proposed method incorporates the *phred* scores and uses the EM algorithm, we refer to it as PhredEM.

The majority of loci in the human genome are monomorphic [The International SNP Map Working Group, 2011], and are as such of little interest in downstream analyses. To avoid running the full PhredEM algorithm at loci that are estimated to be monomorphic, we propose a simple and computationally efficient algorithm to identify and 'screen out' these loci; an earlier version of this screening algorithm that does not incorporate *phred* scores was first proposed in Hu et al. [2016]. We assume HWE holds, as loci that might be called monomorphic must have either zero or extremely low MAFs. Then $\boldsymbol{\pi}$ contains only a single parameter $\boldsymbol{\pi}$. We see that formula (6) assigns all mass to $G_i = 0$ when $\hat{\boldsymbol{\pi}} = 0$; thus loci with $\hat{\boldsymbol{\pi}} = 0$ would be called monomorphic if PhredEM was applied to obtain $\hat{\boldsymbol{\pi}}$. To determine whether $\hat{\boldsymbol{\pi}} = 0$ without fitting PhredEM, let $pl^*(\boldsymbol{\pi})$ denote the profile likelihood for $\boldsymbol{\pi}$, namely,

$$pl^*(\pi)=\max_{\boldsymbol{\theta}}\log L_o^*(\boldsymbol{\theta},\pi).$$

We show in the Appendix that $pl^*(\boldsymbol{\pi})$ is a concave function of $\boldsymbol{\pi}$, so that a negative value for the derivative of $pl^*(\boldsymbol{\pi})$ at $\boldsymbol{\pi} = 0$ implies $\hat{\boldsymbol{\pi}} = 0$; in other words, we should screen out loci at which the derivative of $pl^*(\boldsymbol{\pi})$ at $\boldsymbol{\pi} = 0$ is negative. At $\boldsymbol{\pi} = 0$, we can easily evaluate this derivative, because the part $L_o(\boldsymbol{\theta}, \boldsymbol{\pi})$ reduces to that of a logistic regression model in which we assign an outcome variable $Y_{ik} = 1$ to a minor allele read and $Y_{ik} = 0$ to a major allele read and regress Y_{ik} on Q_{ik}. Since our screening algorithm only involves fitting a standard logistic regression model plus a penalty term to solve for $\boldsymbol{\theta}$ and calculating a derivative function, it can significantly reduce the computing time that is needed to run PhredEM on whole exome or genome data.

A simple variant of the screening algorithm can also be used when estimating the parameters κ and ϕ for the gamma penalty term. If we first apply the screening algorithm using the *unpenalized* profile likelihood $pl(\boldsymbol{\pi}) = \max_{\boldsymbol{\theta}} \log L_o(\boldsymbol{\theta}, \boldsymbol{\pi})$, we can easily find all loci having $\hat{\boldsymbol{\pi}} = 0$ without running the full EM algorithm to maximize (4) at all loci. If the MLE of $\boldsymbol{\pi}$ is zero, then β_0 and β_1 can be estimated using standard logistic regression since all minor allele reads are errors. The few loci for which $\hat{\pi} > 0$ can either be excluded, or the full EM algorithm can be used to estimate β_0 and β_1.

Our approach does not use LD information. It is well known that use of LD patterns can substantially improve genotype calling for variants having moderate or high minor allele frequencies [Nielsen et al., 2011]. However, we can easily incorporate LD information into our approach by calculating the genotype likelihood at each locus using (3), evaluated at the MPLE, and then using this genotype likelihood as input to Beagle [Browning and Yu, 2009].

RESULTS

SIMULATION STUDIES

We conducted simulation studies to assess the performance of PhredEM and PhredEM followed by Beagle (PB), relative to SeqEM and SeqEM followed by Beagle (SB). We considered a sample size of 1,000 (results based on a sample size of 200 are reported in Supplemental Figure S1 and Supplemental Tables S1 and S2). In each replicate, for each individual we first generated a pair of haplotypes of European ancestry having length 100 kb using the coalescent simulator cosi [Schaffner et al., 2005]. We then generated sequencing reads with fixed length 100 bp that mimic reads from the Illumina HiSeq 2000 single-end sequencing platform [Minoche et al., 2011]. Specifically, for each read from an individual, we randomly selected one of the two haplotypes, randomly picked the starting position of the read along the haplotype, and simulated 100 *phred* scores from the empirical distribution observed in the UK10K data (Figure 2[a]). To incorporate the fact that base-calling errors occur at the end of the reads more frequently than at the beginning [Minoche et al., 2011], we rear-ranged the *phred* scores so that the last 15 bases of the read had the 15 lowest scores in a descending order; the first 85 bases thus received a random permutation of the remaining scores. Then, the base calls of the read were generated based on the underlying haplotype and error rates calculated from equation (1); we used because it is more favorable to GATK than to our method. For each individual, we drew the number of reads to be generated from a negative-binomial distribution with mean $1,000 \times c$ so as to achieve a pre-specified average read depth c. We considered three average depths: 6x, 10x, and 30x. In applying PhredEM and SeqEM, we first called genotypes with HWE and, if the estimated MAF was greater than 5%, we re-called genotypes with HWD (starting at parameter values obtained from HWE). The hyper-parameters for the Gamma prior of β_1 were estimated based on the MLEs of β_1 from the 100k loci in each replicate. All results reported here were based on 200 replicates of the entire process.

We first assessed the performance of PhredEM, SeqEM, PB, and SB in truly monomorphic loci. A monomorphic locus is mis-called if there is at least one call of the minor allele in the study sample. Figure 1(a) shows that, with or without LD refinement, PhredEM made fewer mistakes among monomorphic loci than SeqEM at all depths. In addition, LD-refinement has negligible improvement upon PhredEM at monomorphic loci.

We then compared the four methods in calling genotypes for rare variants. We grouped variants into four categories based on the true minor allele counts (MACs): 1, [2, 10], [11, 20], and [21, 100], where MAC = 1 corresponds to singletons. As shown in Table I, the overall number of mis-called genotypes obtained by PhredEM was less than that by SeqEM in all scenarios; for most cases, PhredEM reduced by almost one half the number of mis-called genotypes compared with SeqEM. For instance, when the MAC was between 11 and 20 and depth was 6x, SeqEM mis-called an average of 2.96 genotypes among 997 individuals whereas PhredEM mis-called 1.58. As expected, both methods became more accurate as the average read depth increased. Nevertheless, the performance of PhredEM was noticeably better than SeqEM at depth as high as 30x. We further examined the mis-called genotypes stratified by the underlying genotype. In both the strata of homozygotes ($G = 0$) and heterozygotes ($G = 1$), PhredEM mis-called fewer genotypes than SeqEM. Applying Beagle after PhredEM substantially improved the performance of PhredEM alone, except for singletons at which the two methods have comparable mis-call rates. The superiority of PhredEM over SeqEM remained after applying Beagle to both methods.

For common variants, we stratified the results based on five MAF intervals. As shown in Table II, PhredEM outperformed SeqEM in both the overall and stratified number mis-called. Overall, PhredEM correctly called 3–4 more genotypes than SeqEM at depth ≤10x. The number mis-called by PhredEM increases as the MAF increases because the information in the *phred* scores is not used when $G = 1$, which can be seen from (3). Furthermore, minor allele homozygotes are more likely to be mis-called than major allele homozygotes due to the smaller prior probability of the former. As expected, applying Beagle after PhredEM substantially improved genotype calling by PhredEM alone for common variants, and the improvement was most profound for heterozygotes ($G = 1$). This marked improvement was also shown in Supplemental Table S3 where the error rates are reported given the called variants instead of the true variants as in Table II.

We further examined the *phred* scores at loci having genotypes that are called differently by PhredEM and SeqEM. In Table III, we displayed the average *phred* score associated with major and minor alleles at such loci, stratified by the underlying genotype (G) and genotypes called by PhredEM (G_P) and SeqEM (G_S). At loci with (G_P, G_S) = (0, 1), regardless of the value of G, the major alleles tend to have high *phred* scores whereas the minor alleles tend to have low scores, explaining why PhredEM called these loci major allele homozygotes; the average *phred* scores for minor alleles are consistently lower under $G = 0$ than that under $G = 1$, because in the former case the minor alleles are all errors and in the latter case the minor alleles are a mixture of errors and true alleles. Similarly, for loci with (G_P, G_S) = (2, 1), the major alleles tend to have low scores, which are even lower under $G = 2$ than those under $G = 1$. In other cases when PhredEM called heterozygous genotypes, we observe high average *phred* scores for both major and minor alleles. These patterns of *phred* scores confirm that PhredEM worked as expected. While the results in Table III pertain to common variants, those for rare variants are similar and are shown in Supplemental Table S4.

UK10K SCOOP DATA

To confirm that the results from our simulations hold when analyzing real sequencing data, we analyzed data from the Severe Childhood Onset Obesity Project (SCOOP) cohort sequenced as part of the UK10K project. The sequenced SCOOP cohort consists of 784 UK

Caucasian patients with severe early onset obesity, who were whole-exome sequenced using the Illumina HiSeq 2000 platform with an average depth of ~60x. We first used SAMtools to generate pileup files from BAM files, filtering out reads that are PCR duplicates, with mapping score ≤ 30, or with improperly mapped mates. From the pileup files, we extracted read count data and *phred* scores. The distribution of the *phred* scores is shown in Figure 2(a).

Using the SCOOP sequencing data, we checked the fit of the logistic regression model in (2). First, we applied our screening algorithm to identify loci that were estimated to be monomorphic (i.e., $\hat{\pi} = 0$). At such loci, we could reliably treat all minor allele reads as errors. Assigning $Y = 1$ and 0 for minor allele reads and major allele reads, respectively, we can determine the relationship between $\Pr(Y = 1)$ and the corresponding *phred* scores Q. To create a subset of such data that is computationally manageable, we randomly selected 1,000 monomorphic loci from each of the 22 chromosomes and randomly picked one individual from each locus, forming a dataset of 22,000 (Y, Q) pairs. Then, we fit the logistic regression model in [2] and, as a gold standard, fit a smooth spline function of *phred* scores using the generalized additive model (GAM) [Wood, 2006]. Figure 2(b) shows the fitted curves and pointwise 95% confidence intervals from the two models. The logistic regression fit always fell within the 95% confidence region of the GAM. Thus, we conclude that over the range of *phred* scores found in real data, the logistic model adequately describes the relationship between *phred* scores and base-calling error rates well.

To facilitate the evaluation of PhredEM and especially the comparison with SeqEM, we first selected a set of genotypes that can serve as the gold standard. Specifically, we downloaded from the UK10K website the VCF files for the SCOOP cohort, which contained genotypes called by SAMtools and filtered by GATK. In addition, we excluded a variant if its average depth across samples is less than 20. We excluded a genotype whose genotype likelihood (on the *phred* scale) was ≤ 20 (i.e., nominal genotyping error rate ≥ 0.01) and excluded a variant completely if it has more than 20% of genotypes with likelihood ≤ 20. These exclusion criteria ensured that all selected genotypes were called with particularly high quality. We thus refer to these genotypes as 'true' genotypes. After applying the exclusion criteria, there remain 416,402 loci in the entire exome. Since the loci with true genotypes were selected towards having high read depth, both PhredEM and SeqEM would perform well if applied to the original data. To create sequencing data with low or median depth, we then subsampled the observed reads with equal probability.

We based the estimation of hyper-parameters κ and ϕ on 100k random loci that were reliably estimated to be monomorphic (i.e., with coverage $>$ 60x and the MLE of the MAF π is zero); these 100k loci mimic real sequencing data in which the vast majority of loci are monomorphic whereas the 416,402 loci extracted from the VCF files are mostly polymorphic. We then applied PhredEM and SeqEM to call genotypes assuming HWE at first and, if the estimated MAF was over 5%, we re-called genotypes assuming HWD. The computation time of PhredEM to call the subsampled UK10K data depends on the average depth. For example, it took ~5 h on an Intel Xeon E5-2660 machine with 2.60 GHz and 6.4 GB memory to call genotypes at the 416,402 loci in the 6x dataset.

The numbers of mis-called genotypes, averaged over all variants on chromosomes 1–22 and stratified by MAF ranges, are displayed in Table IV. For rare variants (MAF ≤ 0.05), the pattern in the number of mis-called genotypes by PhredEM and SeqEM agreed well with the results in the simulation section, with PhredEM generally producing more accurate genotype calls. The biggest difference occurred when the variants were relatively rare, i.e., MAF $\in (0.001, 0.01]$; when the average read depth was ~6x, PhredEM generated an average of 1.9 more correct genotypes out of 757 individuals than SeqEM for loci with MAFs in this range. For common variants (MAF > 0.05), the differences between the two methods were smaller, possibly because *phred* scores at heterozygous loci are not informative; this also explains the increase in genotype-calling error rates with increasing MAF found in Table IV. As seen in the simulation results, applying Beagle after PhredEM improved the performance of PhredEM alone for all variants except for the very rare ones (e.g., MAF $\in (0, 0.001]$). The *phred* scores at loci with differently called genotypes by PhredEM and SeqEM are summarized in Supplemental Table S5. These results exhibited the same patterns seen in the simulated data. The mis-call rate at monomorphic loci (Figure 1 [b]) also show the same pattern seen in the simulated data (Figure 1 [a]).

To gain more insights into the mechanisms of PhredEM and SeqEM, we listed in Table V the raw data at eight loci (from the subsampled dataset at 6x) that were called differently by PhredEM and SeqEM. Generally, base calls with low *phred* score are error-prone, and PhredEM treats these unreliable calls as likely errors when calling the genotype. By contrast, SeqEM depends heavily on the proportion of minor allele reads among the total reads and ignores the quality measure of each allele. For example, at Locus 1, the six major alleles were of high quality while the two minor alleles were likely to be errors. In this case, PhredEM distinguishes between alleles of different qualities and produced the correct genotype but SeqEM, which cannot account for low quality alleles, calls the incorrect genotype.

1000 GENOMES CEU DATA

To compare PhredEM to GATK, we considered data from the CEU samples in the 1000 Genomes project. It is hard to make this comparison using simulated data, since it is difficult to construct BAM files for the simulated data, and because the 100KB region we simulated is to short to train the BQSR model used in GATK. It is also hard to make this comparison using the UK10K SCOOP data, as BAM files for the subsampled data are not easily available. In the CEU cohort, 99 unrelated individuals were whole-genome sequenced with an average depth of ~7.3x. We adopted the same filters for the reads as in the analysis of UK10K SCOOP data. As the 99 CEU samples have also been genotyped on the Illumina Omni 2.5 array, we treated these array genotypes as the gold standard. We excluded array SNPs at which $\ge 5\%$ of the samples have missing array genotypes or are not covered by any reads. We also removed 11,119 array SNPs where the genotypes called using sequencing data for all three methods (SeqEM, PhredEM and GATK) indicated a MAF that differed by more than 0.2 from the MAF based on the array genotypes. After these exclusions, there were 1,842,422 array SNPs available for comparison.

We estimated the hyper-parameters for PhredEM based on a random subset of 100k array SNPs that are called as monomorphic using the genotype array in the 99 CEU cohort. In addition to PhredEM and SeqEM, we also applied GATK, using the base quality score recalibration step implemented in BQSR (GATK version 3.6) and a genotype calling step by UnifiedGenotyper with default options. It took 3.4 days for BQSR and 1.3 days for UnifiedGenotyper to run; in contrast, it took a total of 1.7 days for PhredEM to call the same set of genotypes.

PhredEM performed better than SeqEM and GATK in general. Figure 1(c) shows that, at monomorphic loci (i.e., no polymorphism in the array genotypes of the 99 samples), PhredEM has the smallest mis-call rate with or without LD refinement whereas GATK has the highest mis-call rate. Table VI displays the numbers of mis-called genotypes at polymorphic loci, stratified by the ‘true’ MAFs (i.e., based on array genotypes). In most strata, the numbers for PhredEM are smaller than that for GATK, with or without Beagle. The results stratified on the estimated MAF by each method are presented in Table S6, which shows similar patterns. All results consistently indicate that GATK tends to call too many heterozygotes at rare variants and monomorphic loci. Table S7 compares the sensitivity (i.e., the probability of calling a minor allele given a minor allele is truly present) and specificity (probability of calling the major allele given the major allele is truly present) for the methods we consider in Table S7. We find that PhredEM with the LD refinement has the highest specificity (although the differences are tiny, they are significant and when amplified to the genome-wide scale can represent a meaningful difference). PhredEM with LD refinement has the best sensitivity at very low MAF by a considerable amount (0.828, compared to 0.748 for GATK with LD refinement); for higher MAFs, GATK with LD refinement outperforms PhredEM with LD refinement by smaller amounts (e.g., 0.954 for PhredEM with LD vs. 0.958 for GATK with LD). When evaluating the importance of the differences reported in Table S7, it is worth noting that the number of truly polymorphic alleles with low MAF is much smaller than the number of monomorphic alleles, so that a small difference in specificity results in more mis-calls than a larger difference in sensitivity. This explains how GATK with LD can have a higher sensitivity but a lower accuracy as reported in Table VI.

DISCUSSION

We have developed a *phred*-score-informed genotype-calling approach for NGS studies, called PhredEM. We also proposed a simple and computationally efficient screening algorithm to identify loci that would be called as monomorphic. PhredEM improves the accuracy of genotype-calling by estimating base-calling errors from both read data and *phred* scores, and by using all sequencing reads available without setting a *phred*-score-based quality threshold. PhredEM is closely related to the SeqEM approach, which can be viewed as a special case of PhredEM. We showed that the logistic model relating *phred* score to base-calling error rate used in PhredEM fits real sequencing data well. The software program implementing PhredEM, also called PhredEM, is freely available at http://web1.sph.emory.edu/users/yhu30/software.html. The webpage also contains a link to utility programs that process raw BAM files for use as inputs to PhredEM.

In our logistic regression model (2), the *phred* score is the only predictor for the base-calling error. Other important predictors for base-calling quality could also be included. One interesting factor is the position in the read [Brockman et al., 2008], although it is unclear whether this has an independent effect once the *phred* score is accounted for. We did not consider the mapping score as a possible covariate because there is little variability in mapping scores [Li et al., 2008] (see Supplemental Figure S2). However, we recommend that PhredEM should be applied after excluding alignments with mapping scores less than 30.

Our approach is similar in spirit to GATK with BQSR because we allow the relationship between error and *phred* score to be determined by fit to the data, but our approach is more accurate and computationally more tractable. Because we allow a separate set of error parameters at each locus, we automatically account for any covariates that are locus-dependent such as the actual alleles at each locus. We could also consider adding other predictors of error that are included in BQSR that vary across reads.

We recommend using PhredEM with the HWE assumption first, because most loci have low MAFs and HWE has a minimal effect for them. If the estimated MAF is greater than 5%, a second pass of PhredEM could easily be made using the model assuming HWD, which is more robust. Our numerical studies (not shown) suggest that at medium or high read depth (≥10x), the estimated genotype frequencies based on the calls from PhredEM converged rapidly to their true values with increasing sample size even when assuming HWD.

PhredEM is based on several simplifying assumptions. First, the sample should consist of independent, unrelated individuals; this is essential to the likelihood in expression (4). A version of PhredEM could be constructed for trio data by modeling the joint genotypes of parents and offspring, for example, using the conditional-on-parental genotypes (CPG) approach of Schaid and Sommer [1993]. We also assume that errors are symmetric, i.e. that the probability of a read for the major allele being mis-called as the minor allele is the same as the probability of the minor allele being mis-called as the major allele. Further, PhredEM assumes that all variants are biallelic. The biallelic assumption is reasonable because only a small fraction of SNPs have been verified to carry three or more alleles [Hodgkinson and Eyre-Walker, 2010]. In analyzing the UK10K and 1000 Genomes data, we deleted in advance all calls for bases that differed from the two most frequent bases at every locus.

LD information is helpful in identifying monomorphic loci and calling genotypes for both rare and common variants. Therefore, we recommend always using Beagle in conjunction with PhredEM when calling genotypes for NGS data.

In summary, we developed PhredEM, an improved genotype caller which reduces the genotype-calling errors for NGS data. We also proposed a simple and computationally inexpensive algorithm for screening out loci that are estimated to be monomorphic. We showed in simulations that the proposed approach generates fewer incorrect calls than SeqEM regardless of the average read depth and sample size. Using the UK10K and 1000 Genomes sequencing data, we demonstrated the capability of PhredEM to improve the genotype-calling accuracy over SeqEM and GATK in real sequencing data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study makes use of data generated by the UK10K Consortium, derived from samples from the Severe Childhood Onset Obesity Project (SCOOP). A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. This research was supported in part by the University Research Committee (URC) Award at Emory and the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM116065-01A1.

References

Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. 2008; 18:763–770. [PubMed: 18212088]

Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet. 2009; 85:847–861. [PubMed: 19931040]

Boyd, S., Vandenberghe, L. Convex Optimization. Cambridge University Press; New York: 2004.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc Series B Methodol. 1977; 39:1–38.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]

Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet. 2012; 21:R1–R9. [PubMed: 22983955]

Ewing B, Green P. Base-calling of automated sequencer traces using *phred*. II. Error probabilities. Genome Res. 1998; 8:186–194. [PubMed: 9521922]

Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. Genome Res. 1998; 8:175–185. [PubMed: 9521921]

Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993; 80:27–38.

Hodgkinson A, Eyre-Walker A. Human triallelic sites: evidence for a new mutation mechanism. Genetics. 2010; 184:233–241. [PubMed: 19884308]

Hu YJ, Liao P, Johnston HR, Allen AS, Satten GA. Testing rare-variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. PLoS Genet. 2016; 12(5):e1006040.doi: 10.1371/journal.pgen.1006040 [PubMed: 27152526]

Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina genome analyzer using machine learning strategies. Genome Biol. 2009; 10:R83. [PubMed: 19682367]

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27:2987–2993. [PubMed: 21903627]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis GR, Durbin R. 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. Bioinformatics. 2009a; 25:2078–2079. [PubMed: 19505943]

Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851–1858. [PubMed: 18714091]

Li M, Nordborg M, Li LM. Adjust quality scores from alignment and improve sequencing accuracy. Nucleic Acids Res. 2004; 32:5183–5191. [PubMed: 15459287]

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012; doi: 10.1155/2012/251364

Martin ER, Kinnamon DD, Schmidt MA, Powell EH, Zuchner S, Morris RW. SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics. 2010; 26:2803–2810. [PubMed: 20861027]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]

Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 2011; 12:R112. [PubMed: 22067484]

Muddyman D, Smee C, Griffin H, Kaye J. Implementing a successful data-management framework: the UK10K managed access model. Genome Med. 2013; 5:1–9. [PubMed: 23311897]

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42:30–35. [PubMed: 19915526]

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011; 12:443–451. [PubMed: 21587300]

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13:341–353. [PubMed: 22827831]

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005; 11:1576–1583.

Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet. 1993; 53:1114–1126. [PubMed: 8213835]

Sims D, Sudbery l, IIott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014; 15:121–132. [PubMed: 24434847]

The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]

The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409:928–933. [PubMed: 11237013]

Wood, SN. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC; 2006.

Yu YW, Yorukoglu D, Peng J, Berger B. Quality score compression improves genotyping accuracy. Nat Biotechnol. 2015; 33:240–243. [PubMed: 25748910]

APPENDIX

EM ALGORITHM

In the EM algorithm, G_i ($i = 1, \ldots, n$) is treated as missing. The complete-data log-likelihood has the form

$$l_c(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{g=0}^{2} I(G_i = g) \ \{\log P_{\boldsymbol{\theta}}(R_i | g, T_i, \boldsymbol{Q}_i) + \log P_{\pi}(g)\} .$$

Let $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\pi}^{(k)}$ be the parameter values after the kth iteration. In the E-step of the $(k+1)$th iteration, we evaluate $E\{I(G_i = g) | R_i, T_i, \boldsymbol{Q}_i\}$ for $g = 0, 1, 2$, which can be shown to be

$$\omega_{ig}^{(k)} \equiv \frac{P_{\boldsymbol{\theta}^{(k)}}(R_i|g,T_i,\boldsymbol{Q}_i)P_{\boldsymbol{\pi}^{(k)}}(g)}{\sum_{g'=0}^{2}P_{\boldsymbol{\theta}^{(k)}}(R_i|g',T_i,\boldsymbol{Q}_i)P_{\boldsymbol{\pi}^{(k)}}(g')}.$$

In the M-step, we maximize $l_c(\boldsymbol{\theta},\boldsymbol{\pi})$ with $I(G_i = g)$ replaced by $\omega_{ig}^{(k)}$. Specifically, under HWE we update $\boldsymbol{\pi}$ by a closed form $\pi^{(k+1)}=(2n)^{-1}\sum_{i=1}^{n}(2\omega_{i2}^{(k)}+\omega_{i1}^{(k)})$, or under HWD we update $\boldsymbol{\pi}$ by the same $\boldsymbol{\pi}^{(k+1)}$ and update f by $f^{(k+1)}=1-\sum_{i=1}^{n}\omega_{i1}^{(k)}/\left\{2n\pi^{(k+1)}(1-\pi^{(k+1)})\right\}$. We use a one-step Newton-Raphson iteration to update $\boldsymbol{\theta}$. We iterate between the E-step and M-step until the changes in the parameter estimates are negligible.

PROOF OF CONCAVITY OF *pl**(*π*)

First, we prove that, for fixed $\boldsymbol{\theta}$, the function $h(\boldsymbol{\pi}) = \log\{\Sigma_{g=0,1,2}\, P_{\boldsymbol{\theta}}(R|g, T,\boldsymbol{Q})P_{\boldsymbol{\pi}}(g)\}$ is concave. Under HWE, we write $h(\boldsymbol{\pi}) = \log\{a\boldsymbol{\pi}^2 + b(1 - \boldsymbol{\pi})^2 + 2c\boldsymbol{\pi}(1 - \boldsymbol{\pi})\}$, where $a = P_{\boldsymbol{\theta}}(R|G = 2, T,\boldsymbol{Q})$, $b = P_{\boldsymbol{\theta}}(R|G = 0, T,\boldsymbol{Q})$, and $c = (0.5)^T$. The second derivative of $h(\boldsymbol{\pi})$ is

$$h''(\pi)=-\frac{2\{(a+b-2c)\pi+(c-b)\}^2+2(c^2-ab)}{\left\{a\pi^2+b(1-\pi)^2+2c\pi(1-\pi)\right\}^2}.$$

Because $ab=\prod_{k=1}^{T}\varepsilon_k(\boldsymbol{\theta})\,\{1-\varepsilon_k(\boldsymbol{\theta})\} \le (0.25)^T=c^2$, we obtain $h''(\boldsymbol{\pi}) \le 0$ and thus $h(\boldsymbol{\pi})$ is a concave function of $\boldsymbol{\pi}$.

Because the sum of concave functions is still concave, $\log L_o(\boldsymbol{\theta}, \boldsymbol{\pi})$ is concave in $\boldsymbol{\pi}$ for fixed $\boldsymbol{\theta}$. It follows that $L_o^*(\boldsymbol{\theta},\pi)=\log\Gamma(-\beta_1;\hat{\kappa},\hat{\phi})+\log L_o(\boldsymbol{\theta},\pi)$ is also concave in $\boldsymbol{\pi}$ for fixed $\boldsymbol{\theta}$. Because the pointwise supremum over $\boldsymbol{\theta}$ preserves the concavity [Boyd and Vandenberghe, 2004], $pl^*(\boldsymbol{\pi})$ is concave.

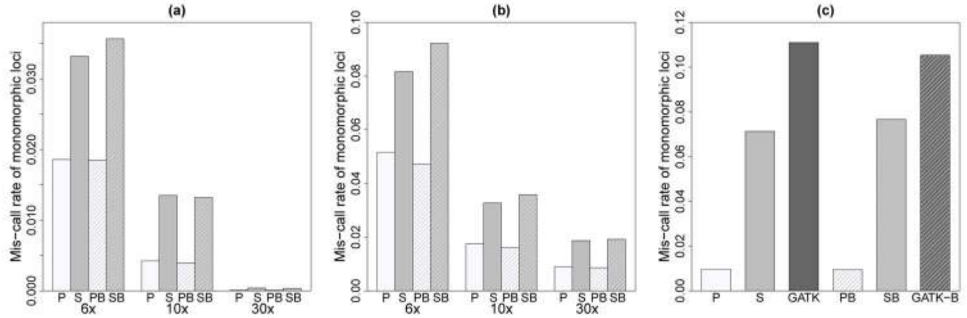

Figure 1.

Mis-call rates at monomorphic loci in the analysis of (a) the simulated data, (b) the UK10K SCOOP data, and (c) the 1000 Genomes CEU data. P and S represent PhredEM and SeqEM. PB, SB, and GATK-B represent PhredEM, SeqEM, and GATK, each followed by Beagle.

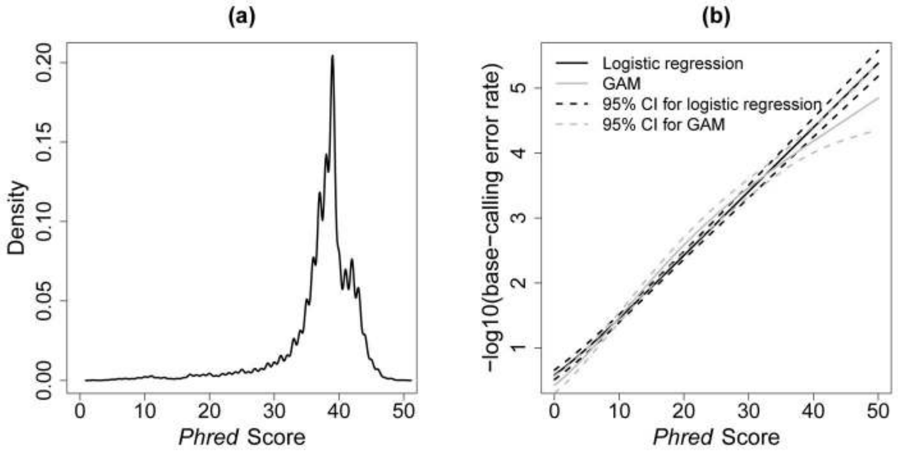

Figure 2.

UK10K SCOOP data. (a) Distribution of *phred* scores. (b) Logistic regression model and generalized additive model (GAM) fit to the sequencing data at loci that were identified as monomorphic.

Table I

Average number of mis-called genotypes per variant for rare variants in the simulation studies.

		Overall					Stratified									
							$G = 0$					$G = 1$				
MAC	**Depth**	N	**P**	**S**	**PB**	**SB**	N_0	**P**	**S**	**PB**	**SB**	N_1	**P**	**S**	**PB**	**SB**
1	6×	997.1	0.241	0.311	0.274	0.338	996.1	0.065	0.072	0.096	0.096	1	0.176	0.239	0.178	0.242
	10×	999.7	0.074	0.135	0.088	0.159	998.7	0.016	0.033	0.039	0.051	1	0.058	0.102	0.049	0.108
	30×	1000	0.001	0.004	0.002	0.006	999.0	0	0.001	0.001	0.003	1	0.001	0.003	0.001	0.003
[2, 10]	6×	997.1	0.525	0.845	0.439	0.691	993.5	0.106	0.112	0.162	0.193	3.6	0.417	0.730	0.275	0.496
	10×	999.7	0.191	0.315	0.142	0.243	996.2	0.049	0.060	0.063	0.082	3.5	0.140	0.253	0.079	0.161
	30×	1000	0.004	0.009	0.003	0.007	996.4	0.001	0.002	0.002	0.004	3.6	0.003	0.007	0.001	0.003
[11, 20]	6×	997.0	1.579	2.959	0.779	1.306	982.2	0.387	0.514	0.243	0.429	14.7	1.156	2.409	0.529	0.868
	10×	999.7	0.551	1.011	0.212	0.381	984.9	0.156	0.176	0.090	0.138	14.7	0.380	0.819	0.121	0.241
	30×	1000	0.011	0.026	0.005	0.010	985.1	0.004	0.007	0.003	0.005	14.8	0.007	0.019	0.002	0.005
[21, 100]	6×	997.0	4.197	7.633	1.416	2.217	947.8	0.667	2.108	0.347	0.696	48.5	3.136	5.131	1.051	1.489
	10×	999.7	1.457	2.722	0.361	0.603	949.9	0.347	0.606	0.126	0.210	49.1	1.002	1.998	0.230	0.381
	30×	1000	0.032	0.068	0.009	0.016	949.6	0.008	0.015	0.004	0.007	49.6	0.024	0.051	0.005	0.009

P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively. N, N_0 and N_1 are the average numbers of individuals covered by at least one read. G is the true genotype; the case $G = 2$ is omitted as it is barely seen for rare variants. MACs of 1, 10, 20, and 100 correspond to MAFs of 0.0005, 0.005, 0.01, and 0.05, respectively, given the sample size of 1,000.

Table II

Average number of mis-called genotypes per variant for common variants in the simulation studies.

		Overall					Stratified														
							G = 0					*G* = 1					*G* = 2				
MAF	**Depth**	***N***	**P**	**S**	**PB**	**SB**	N_0	**P**	**S**	**PB**	**SB**	N_1	**P**	**S**	**PB**	**SB**	N_2	**P**	**S**	**PB**	**SB**
(0.05, 0.1]	6×	997.0	11.47	16.94	1.97	2.85	857.9	0.95	4.61	0.44	0.77	133.5	8.80	10.58	1.49	2.01	5.6	1.72	1.75	0.04	0.07
	10×	999.7	3.62	6.28	0.49	0.71	858.9	0.59	1.61	0.14	0.22	135.4	2.59	4.18	0.33	0.47	5.4	0.44	0.49	0.02	0.02
	30×	1000	0.07	0.15	0.01	0.02	858.9	0.01	0.03	0	0.01	135.5	0.05	0.11	0.01	0.01	5.6	0.01	0.01	0	0
(0.1, 0.2]	6×	997.1	22.66	28.31	2.15	2.96	725.3	1.14	5.67	0.45	0.78	249.2	17.81	18.78	1.60	2.05	22.6	3.71	3.86	0.10	0.13
	10×	999.7	6.38	9.93	0.64	0.80	731.7	0.72	2.45	0.17	0.21	246.2	4.94	6.55	0.44	0.54	21.8	0.72	0.93	0.03	0.05
	30×	1000	0.11	0.24	0.01	0.02	727.1	0.02	0.05	0	0.01	250.2	0.08	0.17	0.01	0.01	22.7	0.01	0.02	0	0
(0.2, 0.3]	6×	997.1	35.53	40.59	2.44	3.33	568.4	1.17	5.51	0.48	0.87	367.3	29.49	29.72	1.79	2.19	61.4	4.87	5.36	0.17	0.27
	10×	999.7	9.64	13.71	0.80	0.95	562.6	0.75	2.86	0.17	0.22	373.3	8.03	9.43	0.57	0.65	63.8	0.86	1.42	0.06	0.08
	30×	1000	0.15	0.34	0.01	0.03	564.7	0.03	0.08	0	0.01	372.2	0.11	0.22	0.01	0.02	63.1	0.01	0.04	0	0
(0.3, 0.4]	6×	997.0	45.56	49.99	2.66	3.74	423.2	1.09	4.57	0.46	0.96	450.1	40.19	40.11	1.94	2.32	123.7	4.28	5.31	0.26	0.46
	10×	999.7	11.68	15.91	0.89	1.01	426.6	0.69	2.80	0.16	0.21	451.8	10.11	11.28	0.64	0.69	121.3	0.88	1.83	0.09	0.11
	30×	1000	0.18	0.38	0.02	0.03	425.0	0.03	0.07	0.01	0.01	452.7	0.13	0.25	0.01	0.02	122.3	0.02	0.06	0	0
(0.4, 0.5]	6×	997.1	50.63	54.88	2.72	4.02	305.0	1.22	3.83	0.39	0.90	491.6	45.53	45.46	2.01	2.41	200.5	3.88	5.59	0.32	0.71
	10×	999.7	12.90	17.17	0.98	1.11	302.6	0.63	2.46	0.14	0.18	493.2	11.38	12.38	0.71	0.76	203.9	0.89	2.33	0.13	0.17
	30×	1000	0.19	0.41	0.02	0.03	302.7	0.03	0.08	0.01	0.01	494.3	0.14	0.26	0.01	0.02	203.0	0.02	0.07	0	0

P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively. *N*, N_0, N_1, and N_2 are the average numbers of individuals covered by at least one read. *G* is the true genotype.

Table III

Average *phred* scores associated with major (M) and minor (m) alleles at loci that are called differently by PhredEM and SeqEM in the simulation studies for common variants.

		$G = 0$				$G = 1$								$G = 2$			
		(0, 1)		(1, 0)		(0, 1)		(1, 0)		(1, 2)		(2, 1)		(1, 2)		(2, 1)	
MAF	**Depth**	**M**	**m**	**M**	**m**	**M**	**m**	**M**	**m**	**M**	**m**	**M**	**m**	**M**	**m**	**M**	**m**
(0.05, 0.1]	6×	37.2	9.4	37.1	36.3	37.2	14.4	37.1	38.6	39.0	34.9	9.2	37.4	37.5	34.8	7.5	37.4
	10×	37.2	9.7	37.1	36.8	37.0	15.2	37.1	38.6	38.4	36.7	10.3	37.4	36.4	35.7	8.3	37.5
	30×	37.1	10.5	36.9	37.0	36.7	17.2	37.2	38.3	39.3	37.5	18.1	38.2	37.1	37.2	7.7	37.5
(0.1, 0.2]	6×	37.2	9.2	37.1	36.2	37.1	13.4	37.1	38.6	39.0	34.5	9.2	37.5	36.2	34.2	7.7	37.3
	10×	37.3	9.4	37.1	37.0	37.1	14.3	37.1	38.6	38.5	36.9	10.8	37.2	36.2	36.1	8.6	37.6
	30×	37.2	10.4	37.2	37.1	37.1	16.7	37.1	38.3	38.5	37.3	13.2	37.6	36.2	37.3	9.2	37.2
(0.2, 0.3]	6×	37.2	9.1	37.1	36.2	37.2	12.5	37.1	38.6	38.6	35.3	9.6	37.4	35.9	33.3	8.0	37.2
	10×	37.2	9.3	37.2	36.9	37.2	13.5	37.1	38.6	38.6	36.9	11.1	37.1	36.9	36.7	8.8	36.8
	30×	37.3	10.0	37.1	37.7	36.9	14.8	37.2	38.5	38.5	37.0	14.4	37.4	37.2	36.9	9.6	37.1
(0.3, 0.4]	6×	37.4	8.8	37.3	36.6	37.2	12.0	37.1	38.6	38.8	36.1	10.3	37.2	35.2	33.3	8.3	37.0
	10×	37.0	9.2	37.1	36.7	37.1	13.2	37.1	38.6	38.5	37.4	11.7	37.2	36.6	37.0	8.9	37.2
	30×	37.2	10.2	36.8	37.5	37.4	14.8	37.1	38.4	38.4	37.0	14.9	36.6	37.3	37.3	10.2	36.8
(0.4, 0.5]	6×	37.1	8.5	36.3	36.4	37.2	11.4	37.0	38.6	38.7	36.9	10.9	37.2	35.8	35.8	8.5	37.2
	10×	37.3	9.1	37.1	36.7	37.1	12.7	37.1	38.6	38.6	37.1	12.3	37.1	36.3	37.1	9.0	37.1
	30×	37.2	10.3	37.2	37.1	36.9	14.6	37.0	38.5	38.3	37.3	15.1	37.7	37.5	37.3	10.3	37.2

G is the true genotype. (G_P, G_S) = (0, 1), (1, 0), et al. represent loci that are called to be G_P and G_S by PhredEM and SeqEM, respectively.

Table IV

Average number of mis-called genotypes per variant in analysis of the UK10K SCOOP data (subsampled to achieve different depths).

		Overall					Stratified														
							G = 0					*G* = 1					*G* = 2				
MAF	**Depth**	***N***	**P**	**S**	**PB**	**SB**	N_0	**P**	**S**	**PB**	**SB**	N_1	**P**	**S**	**PB**	**SB**	N_2	**P**	**S**	**PB**	**SB**
(0, 0.001]	6×	762.4	0.30	0.98	0.33	1.23	761.4	0.11	0.74	0.13	0.99	1.0	0.19	0.24	0.20	0.24	0	0	0	0	0
	10×	778.7	0.26	0.94	0.30	1.02	777.7	0.08	0.73	0.12	0.82	1.0	0.18	0.21	0.18	0.20	0	0	0	0	0
	30×	783.6	0.16	0.77	0.21	0.91	782.6	0.05	0.63	0.11	0.75	1.0	0.11	0.14	0.10	0.16	0	0	0	0	0
(0.001, 0.01]	6×	757.0	1.77	3.68	1.65	3.43	752.4	0.32	2.21	0.53	2.10	4.5	1.36	1.38	1.05	1.25	0.1	0.09	0.09	0.07	0.08
	10×	776.1	1.64	3.32	1.43	2.95	771.4	0.30	1.92	0.43	1.91	4.6	1.26	1.32	0.93	0.97	0.1	0.08	0.08	0.07	0.07
	30×	782.2	1.02	2.25	0.81	1.90	777.5	0.27	1.06	0.35	1.31	4.6	0.69	1.13	0.42	0.54	0.1	0.06	0.06	0.04	0.05
(0.01, 0.05]	6×	751.1	10.45	11.52	7.30	8.84	713.1	1.09	3.21	0.85	1.68	37.3	8.87	7.80	6.22	6.91	0.7	0.49	0.51	0.23	0.25
	10×	772.3	8.22	9.17	6.35	7.67	733.4	0.95	2.84	0.66	1.39	38.2	6.88	5.91	5.50	6.08	0.7	0.39	0.42	0.19	0.20
	30×	779.1	1.41	2.33	0.89	1.37	739.9	0.44	1.03	0.28	0.63	38.5	0.78	1.10	0.48	0.60	0.7	0.19	0.20	0.13	0.14
(0.05, 0.1]	6×	749.5	19.52	20.28	11.27	12.25	646.8	1.35	2.14	1.58	1.94	98.4	15.87	15.78	8.89	9.48	4.3	2.30	2.36	0.80	0.83
	10×	772.4	11.99	12.76	7.54	8.13	666.7	1.20	1.73	1.11	1.40	101.3	9.33	9.48	5.89	6.12	4.4	1.46	1.55	0.54	0.61
	30×	779.8	2.34	2.52	1.15	1.47	673.3	0.68	0.72	0.44	0.63	102.0	1.29	1.40	0.43	0.51	4.5	0.37	0.40	0.28	0.33
(0.1, 0.2]	6×	748.5	38.06	38.54	13.47	13.90	546.4	2.12	2.47	1.74	1.95	184.9	28.69	28.81	9.82	9.94	17.2	7.25	7.26	1.91	2.01
	10×	772.3	21.36	21.65	7.95	8.30	563.9	1.91	2.28	1.33	1.52	190.6	15.56	15.44	5.41	5.56	17.8	3.89	3.93	1.21	1.22
	30×	779.4	3.43	3.55	1.10	1.13	569.4	0.84	0.86	0.46	0.50	191.9	1.77	1.87	0.28	0.26	18.1	0.82	0.82	0.36	0.37
(0.2, 0.3]	6×	747.3	62.54	62.93	14.70	15.21	423.7	2.76	3.15	1.75	1.79	276.6	46.20	46.15	10.69	11.05	47.0	13.58	13.63	2.26	2.37
	10×	771.8	33.96	34.41	8.51	8.72	437.7	2.58	2.85	1.37	1.49	285.5	24.84	24.94	5.78	5.81	48.6	6.54	6.62	1.36	1.42
	30×	779.6	4.70	4.86	1.22	1.31	442.3	1.17	1.22	0.45	0.50	288.0	2.48	2.57	0.43	0.45	49.3	1.05	1.07	0.34	0.36
(0.3, 0.4]	6×	748.3	81.03	81.28	15.37	15.91	317.9	2.99	3.30	1.94	2.03	338.2	62.50	62.37	11.02	11.39	92.2	15.54	15.61	2.41	2.49
	10×	772.1	42.04	42.40	9.07	9.21	328.1	2.74	3.02	1.40	1.48	349.0	32.34	32.33	6.15	6.20	95.0	6.96	7.05	1.52	1.53
	30×	780.5	5.43	5.51	1.55	1.63	331.9	1.23	1.27	0.48	0.51	352.1	3.03	3.04	0.65	0.70	96.5	1.17	1.20	0.42	0.42
(0.4, 0.5]	6×	747.3	95.93	96.22	15.82	16.40	221.1	4.41	4.58	2.16	2.21	378.7	75.66	75.69	11.19	11.61	147.5	15.86	15.95	2.47	2.58
	10×	771.4	49.78	50.13	9.47	9.68	228.3	3.39	3.56	1.53	1.58	390.5	39.19	39.35	6.32	6.39	152.6	7.20	7.22	1.62	1.71
	30×	778.7	6.88	7.15	1.82	1.93	230.8	1.34	1.39	0.51	0.60	393.4	4.32	4.46	0.88	0.87	154.5	1.22	1.30	0.43	0.46

G is the true genotype. *N*, N_0, N_1, and N_2 are the average numbers of individuals covered by at least one read. P, S, PB and SB represent PhredEM, SeqEM, PhredEM followed by Beagle, and SeqEM followed by Beagle, respectively.

Table V

Eight example loci in the UK10K SCOOP data (subsampled to 6×).

	Reads		***Phred* scores**		**Genotype**		
Locus	**M**	**m**	**M**	**m**	**True**	**P**	**S**
1	6	2	21 36 37 38 39 42	9 16	0	0	1
2	6	1	18 18 27 36 39 40	33	0	1	0
3	4	1	20 34 34 36	15	1	0	1
4	5	1	25 32 32 34 39	37	1	1	0
5	1	5	35	20 25 38 40 40	1	1	2
6	1	5	14	33 37 38 38 40	1	2	1
7	1	4	32	30 34 37 39	2	1	2
8	2	5	11 17	30 34 35 36 39	2	2	1

M and m represent major and minor alleles, respectively. True is the true genotype. P and S represent the called genotypes by PhredEM and SeqEM, respectively.

Table VI

Average number of mis-called genotypes per variant in the analysis of the 1000 Genomes CEU data.

MAF	*N*	P	S	GATK	PB	SB	GATK-B
(0, 0.01]	98.07	0.185	0.203	0.808	0.197	0.220	0.701
(0.01, 0.05]	98.08	0.546	0.562	0.716	0.285	0.306	0.326
(0.05, 0.1]	98.04	1.330	1.334	1.541	0.451	0.482	0.445
(0.1, 0.2]	98.02	2.553	2.519	2.781	0.724	0.749	0.685
(0.2, 0.3]	98.03	3.716	3.889	3.919	0.727	0.794	0.742
(0.3, 0.4]	98.02	4.648	4.827	4.733	0.835	0.923	0.865
(0.4, 0.5]	98.01	5.189	5.380	5.118	0.886	0.979	0.915

MAF is the minor allele frequency observed in the array genotype data. P and S represent PhredEM and SeqEM, respectively. PB, SB, and GATK-B represent PhredEM, SeqEM, and GATK followed by Beagle. *N* is the average number of individuals covered by at least one read.