Inference of Population Mutation Rate and Detection of Segregating Sites from Next-Generation Sequence Data

Chul Joo Kang¹ and Paul Marjoram

Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089

ABSTRACT We live in an age in which our ability to collect large amounts of genome-wide genetic variation data offers the promise of providing the key to the understanding and treatment of genetic diseases. Over the next few years this effort will be spearheaded by so-called *next-generation* sequencing technologies, which provide vast amounts of short-read sequence data at relatively low cost. This technology is often used to detect unknown variation in regions that have been linked with a given disease or phenotype. However, error rates are significant, leading to some nontrivial issues when it comes to interpreting the data. In this article, we present a method with which to address questions of widespread interest: calling variants and estimating the population mutation rate. We show performance of the method using simulation studies before applying our approach to an analysis of data from the 1000 Genomes project.

WE live in an era in which technology drives scientific discovery. This is very clearly demonstrated in molecular genetics, in which the development of technologies such as gene expression arrays and so-called SNP chips, to name just two examples, has led to a vast increase in the amount of genetic data at our disposal. Arguably, at this time our ability to interpret such data lags somewhat behind the technology. One of the most recent and exciting technologies to be placed at our disposal is next-generation sequence (NGS) platforms, in which enormous quantities of sequence data can be collected at reasonably low cost. In this article we focus on two applications of this technology: estimation of mutation rate and polymorphism detection.

Our focus is partly motivated by a common use for NGS data in genome-wide association studies (GWAS). While GWAS have now identified a large number of loci at which polymorphism is associated with disease phenotypes, the overall amount of variance explained by these polymorphisms is low (e.g., Manolio *et al.* 2009; Frazer *et al.* 2009). One explanation for the remaining variance, so-called *dark*

matter, is the possible existence of *rare variants*, polymorphic loci at which the minor allele has low frequency. With that in mind, in addition to the wish to answer general population genetics questions, such as estimation of mutation rate, the investigator will often collect NGS data to identify polymorphic loci at which to collect data in a follow-up study using a custom SNP array, say. This article presents a computationally tractable method for estimating mutation rate and calling genotypes (and thereby detecting polymorphism).

NGS technology comes in a number of forms. Chief among these are the Illumina (San Diego, CA) Genome Analyzer, the 454 Genome Sequencers (Roche Applied Science, Basel, Switzerland), the SOLiD platform (Applied Biosystems, Foster City, CA), the Polonator (Dover Systems, Salem, NH), and the HeliScope Single Molecule Sequencer technology (Helicos, Cambridge, MA). While these methods differ in the details (see Shendure and Ji 2008 for a review), the data that are obtained have similar broad features. Essentially one obtains relatively short regions of sequenced DNA, so-called *reads*, either singly or in pairs drawn from opposite ends of a larger fragment of DNA. Typically these fragments are sequenced and then aligned using one of a number of possible algorithms [e.g., Mapping and Assembly with Quality (MAQ) (H. Li *et al.* 2008) or Short Oligonucleotide Alignment Program (SOAP) (R. Li *et al.* 2008)], but several key features are constant across platforms. First,

the reads are short (35–400 bp, depending upon platform); second, for that reason, alignment to a reference sequence is challenging; third, genotyping error rates vary along the read, typically increasing as we move along the read (modulo some variation from that trend that may exist at the beginning of the reads). In this article we do not focus on the issue of alignment—our method is designed to be applied to reads postalignment.

While these technologies are new, a number of approaches already exist for the analysis of the resulting data. In the context of estimating mutation rate, the first method was that of Hellmann *et al.* (2008). While their method was developed for shotgun-sequencing data, in which error rates are lower, it can nonetheless be applied to NGS data, albeit at some loss of performance. A similar approach was taken by Jiang *et al.* (2008), where robustness to issues such as genotyping errors or biased amplification was examined more explicitly. Furthermore, a wide variety of methods drawn from related topics also exist. Examples of this range from the extremely elegant and simple estimator due to Watterson (1975), to the more computationally intense methods of Griffiths and Tavaré (1994) and Kuhner *et al.* (1995). However, none of these methods were developed for NGS data, and, for example, they fail to allow for the possibility of genotyping error. Methods for estimating mutation rate in the presence of genotyping errors do exist, for example, approaches based upon considering nonsingleton variants (Knudsen and Miyamoto 2009), but these do not exploit the particular properties of NGS data.

In this article we use the *coalescent* as a model for genotype data for a sample of individuals drawn from a population. The coalescent was first formalized by Kingman (1982 a,b,c) and has become the most widely used model for population genetics data. For accessible introductions see Wakeley (2008) or Hein *et al.* (2005). Several algorithms now exist for detection of polymorphic sites for NGS data. Li and Leal (2009) developed a Bayesian method for computing individual genotype likelihood values from NGS data. There are also approaches that combine the resequenced data of the samples for better SNP calling. For example, Bansal *et al.* (2010) used a method containing a population error correction term to avoid systematic sequencing errors. Such methods were used in the 1000 Genomes Project Consortium (2010). After giving methodological details of our approach we demonstrate performance via a series of simulation studies before applying it to data from the 1000 Genomes Project and comparing to results from two popular algorithms: samtools and GATK.

Methods

Overview of the E–M algorithm

We assume we have read data that have been aligned to a reference sequence and, in the simplest form of our algorithm, that we have known (or estimated) position-

specific error rates. Our goal is to compute individual genotype likelihoods for a sample of size n , *i.e.*, the probability of the genotype for each site for each member of the sample. For computational tractability our method ignores information due to linkage disequilibrium (see *Discussion*).

Our method is an E–M algorithm that explores possible values for Θ and the genotype data while producing the maximum-likelihood estimate of the population mutation rate $4N_e\mu = \Theta$, where N_e is the effective population size and μ is the mutation rate per base per individual. We initialize using an arbitrarily chosen value of $\Theta = 0.001$. In outline the algorithm proceeds as follows:

1. Expectation step: Conditional on the current Θ -value and the observed read data, we calculate the expected number of polymorphic sites, $E(S)$.
2. Maximization step: Update Θ on the basis of $E(S)$.

In analyses in which we also wish to infer error rates, those error rates are updated in step 2, along with Θ .

We iterate the above steps 100 times and then, using the final estimated Θ , the posterior probability of being polymorphic is computed for each site. For each site that is predicted to be polymorphic, the genotype and output, for each individual in the sample, are inferred. While our method is computationally intense, for the examples given in this article, in which we simulate 20 diploid samples over a 100-kb region, the run time is short: 1 min for estimation of Θ or 30 min with coestimation of Θ and error rates. Run time is approximately linear with the length of region being analyzed.

We now give more details on each of the above steps.

Calculation of the likelihood

We begin by deriving the likelihood of Θ conditional on the read data. Let $\mathcal{R} = \{R_1, R_2, \dots, R_n\}$ be the set of read data across all sites (*i.e.*, all base positions) over all samples and let G denote the unobserved genotypes across all sites. Using Bayes' theorem, the probability of Θ given the read data for a sample is

$$\text{Prob}(\Theta|\mathcal{R}) = \frac{\text{Prob}(\mathcal{R}|\Theta)\text{Prob}(\Theta)}{\text{Prob}(\mathcal{R})}. \quad (1)$$

The likelihood of read data $\mathcal{R}$ given Θ is the summation of the joint probabilities over all possible underlying genotypes:

$$L(\Theta) = \sum_G \text{Prob}(\mathcal{R}|G)\text{Prob}(G|\Theta). \quad (2)$$

By ignoring linkage disequilibrium we can compute a composite likelihood (cf. Hudson 2001) by multiplying the likelihoods at each site:

$$L(\Theta) \approx \prod_i \sum_g \text{Prob}(R_i|g)\text{Prob}(g|\Theta), \quad (3)$$

where i indexes sites, g denotes the sample genotypes at a given site, and $\text{Prob}(R_i|g)$ is set to 1 if R_i does not overlap g .

Calculation of Equation 3 requires derivation of the terms $\text{Prob}(R_i | g)$ and $\text{Prob}(g | \Theta)$. We begin by deriving $\text{Prob}(R_i | g)$.

Let $\mathbf{r}$ be the set of read data at a given position in the genome for a given individual, r_k refer to the k th element of $\mathbf{r}$, and $e(r_k)$ be the error rate of the base called at this position for r_k . This can be estimated using quality scores, or from external estimates of read accuracy. While it is generally the case that error rate can be reasonably assumed to vary across positions within a read, but to be constant across reads at each position, we do not make this restriction here. Let a_0 be the base at this position in the reference sequence and a_1 denote a possible variant type. We now compute the individual genotype likelihoods for each possible genotype. The possible genotypes can be categorized as homologous consensus/consensus, heterozygous consensus/variant, and homologous variant/variant. To improve computational efficiency we assume at most a single mutation at each position in the ancestry of the sample, as is the case with the so-called *infinite sites* mode, for example. Thus, we write

$$\begin{aligned} \text{Prob}(\mathbf{r}|a_j, a_j) &\sim \prod_{r_k=a_j} (1 - e(r_k)) \times \prod_{r_k \neq a_j} (e(r_k)) \\ \text{Prob}(\mathbf{r}|a_0, a_1) &\sim \prod_{r_k=a_0} \left[\frac{1}{2}(1 - e(r_k)) + \frac{1}{2}e(r_k) \right] \\ &\quad \times \prod_{r_k=a_1} \left[\frac{1}{2}(1 - e(r_k)) + \frac{1}{2}e(r_k) \right] \\ &= \frac{1}{2}^{|r_k=a_0|+|r_k=a_1|} \end{aligned} \quad (4)$$

We then compute the joint likelihood of the sample genotypes at a given position by multiplying these individual likelihoods.

Next we derive $\text{Prob}(g | \Theta)$. The prior probabilities for each genotype are calculated from the expected allele frequency spectrum under the coalescent model with constant population size or expanding population size (as appropriate). The joint prior probabilities of each genotype are then approximated from the expected allele frequency spectrum (Griffiths and Tavaré 1998).

$$\text{Prob}(g(0)|\Theta) = 1 - \Theta \times \sum_j^{n-1} \frac{1}{j}$$

$$\text{Prob}(g(m)|\Theta) = (1 - \text{Prob}(g(0)|\Theta)) \frac{1}{m} \left(\sum_j^{n-1} \frac{1}{j} \right)^{-1} \binom{n}{m}^{-1} \quad (m > 0)$$

where $g(m)$ represents a genotype with m mutant alleles across the sample.

Estimation of Θ

The value of Θ maximizing the likelihood is derived using an E-M algorithm. The derivative of the likelihood function (3) is

$$L(\Theta)' = \sum_i \frac{\sum_{m=1}^{n-1} \text{Prob}(R_i|g(m)) \text{Prob}(g(m)|\Theta)}{\sum_{m=0}^{n-1} \text{Prob}(R_i|g(m)) \text{Prob}(g(m)|\Theta)} \left(n \times \sum_j^{n-1} \frac{1}{j} \right)^{-1}. \quad (5)$$

From this we see that in the expectation step we will estimate the expected number of polymorphic sites, by summing the probability of each site being polymorphic, where that calculation is performed using the existing estimate of Θ at each site:

$$S_E = \sum_i \frac{\sum_{m=1}^{n-1} \text{Prob}(R_i|g(m)) \text{Prob}(g(m)|\Theta)}{\sum_{m=0}^{n-1} \text{Prob}(R_i|g(m)) \text{Prob}(g(m)|\Theta)} \quad (6)$$

We then re-estimate Θ , per Equation 5, which can be understood as Watterson's estimator of Θ but using the expected (rather than observed) number of polymorphic sites.

Unknown error rates

In its simplest form, our method requires the error rates of the resequencing process as input. While these will generally be available (via quality scores for example) we note that, if not, our method can coestimate either a single overall error rate E^* or a vector E of position-specific error rates, while estimating Θ and calling polymorphism. In this case we model the error rates as growing exponentially along the read, which reflects the generally observed behavior. For example, the likelihood that (3) is rewritten as

$$L(\Theta, E) \approx \prod_i \sum_g \text{Prob}(R_i|g, E^*) \text{Prob}(g|\Theta)$$

and the vector of position-specific error rates, $E = \{E_1, E_2, \dots, E_L\}$ (say), can be estimated within the E-M algorithm

$$E_k = \frac{\sum_i \sum_g a_{i,k}^g \times \lambda(g, i)}{\sum_i \sum_g n_{i,k} \times \lambda(g, i)} = \frac{\sum_i \sum_g a_{i,k}^g \times \lambda(g, i)}{\sum_i n_{i,k}},$$

where $a_{i,k}^g$ is the number of reads that do not match genotype g at site i in the resequencing position k and $n_{i,k}$ is the total number of reads at site i . $\lambda(g, i)$ is the posterior probability of genotype g at site i :

$$\lambda(g, i) = \frac{\text{Prob}(R_i|g, E) \text{Prob}(g|\Theta)}{\sum_g \text{Prob}(R_i|g, E) \text{Prob}(g|\Theta)}$$

To lessen the computational burden, we restrict this calculation to sites that are currently called nonpolymorphic:

Table 1 The mean estimated Θ value and RMSEs (see text for definitions) for varying population parameters

Coverage	$\rho = 0$	$\rho = 10^{-4}$	$\rho = 10^{-3}$	$\rho = 10^{-2}$
1x				
$\hat{\Theta}$	0.00106	0.00102	0.00101	0.00099
R_t	5.016×10^{-4}	2.852×10^{-4}	1.223×10^{-4}	7.26×10^{-5}
R_w	2.253×10^{-4}	1.294×10^{-4}	6.34×10^{-5}	5.18×10^{-5}
2x				
$\hat{\Theta}$	0.00106	0.00102	0.00100	0.00100
R_t	4.344×10^{-4}	2.529×10^{-4}	1.213×10^{-4}	6.48×10^{-5}
R_w	1.415×10^{-4}	9.13×10^{-5}	4.83×10^{-5}	3.76×10^{-5}
4x				
$\hat{\Theta}$	0.00106	0.00102	0.00101	0.00100
R_t	3.870×10^{-4}	2.195×10^{-4}	1.061×10^{-4}	5.42×10^{-5}
R_w	7.88×10^{-5}	5.20×10^{-5}	2.91×10^{-5}	2.69×10^{-5}
8x				
$\hat{\Theta}$	0.00105	0.00102	0.00101	0.00100
R_t	3.458×10^{-4}	1.992×10^{-4}	1.013×10^{-4}	5.33×10^{-5}
R_w	2.91×10^{-5}	2.14×10^{-5}	1.47×10^{-5}	1.38×10^{-5}
16x				
$\hat{\Theta}$	0.00105	0.00102	0.00101	0.00100
R_t	3.258×10^{-4}	1.922×10^{-4}	9.75×10^{-5}	5.33×10^{-5}
R_w	4.5×10^{-6}	5.3×10^{-6}	4.7×10^{-6}	4.3×10^{-6}
	$\Theta = 0.0001$	$\Theta = 0.001$	$\Theta = 0.005$	$\Theta = 0.01$
1x				
$\hat{\Theta}$	0.000094	0.00102	0.00483	0.0104
R_t	3.644×10^{-5}	2.852×10^{-4}	1.234×10^{-3}	2.398×10^{-3}
R_w	2.289×10^{-5}	1.294×10^{-4}	4.15×10^{-4}	9.28×10^{-4}
2x				
$\hat{\Theta}$	0.000095	0.00102	0.00485	0.0102
R_t	3.075×10^{-5}	2.529×10^{-4}	1.112×10^{-3}	2.144×10^{-3}
R_w	1.522×10^{-5}	9.13×10^{-5}	3.23×10^{-4}	6.19×10^{-4}
4x				
$\hat{\Theta}$	0.000094	0.00102	0.00488	0.0102
R_t	2.823×10^{-5}	2.195×10^{-4}	9.98×10^{-4}	1.927×10^{-3}
R_w	9.55×10^{-6}	5.20×10^{-5}	1.77×10^{-4}	3.32×10^{-4}
8x				
$\hat{\Theta}$	0.000095	0.00102	0.00489	0.0102
R_t	2.525×10^{-5}	1.992×10^{-4}	9.175×10^{-4}	1.805×10^{-3}
R_w	4.75×10^{-6}	2.14×10^{-5}	6.25×10^{-5}	1.14×10^{-4}
16x				
$\hat{\Theta}$	0.000095	0.00102	0.00489	0.0102
R_t	2.431×10^{-5}	1.922×10^{-4}	8.925×10^{-4}	1.756×10^{-3}
R_w	1.61×10^{-6}	5.3×10^{-6}	1.1×10^{-5}	1.5×10^{-5}

Top: $\Theta = 0.001$ while ρ varies; bottom: Θ varies while ρ is fixed at 10^{-4} . In all cases, 25 diploid samples were generated on the basis of constant population size model.

$$E_k \approx \frac{\sum_i a_{i,k}^{g(0)} \times \lambda(g(0), i)}{\sum_i n_{i,k} \times \lambda(g(0), i)}.$$

After estimation of error rate at each read position, we fit an exponential curve to derive smoothed estimates of position-specific error rates.

Table 2 The mean estimated Θ value and RMSEs for varying sample sizes, n

Coverages	$n = 10$	$n = 25$	$n = 50$
1x			
$\hat{\Theta}$	0.00099	0.00102	0.00097
R_t	2.907×10^{-4}	2.852×10^{-4}	2.573×10^{-4}
R_w	1.483×10^{-4}	1.294×10^{-4}	1.043×10^{-4}
2x			
$\hat{\Theta}$	0.00099	0.00102	0.00097
R_t	2.614×10^{-4}	2.529×10^{-4}	2.304×10^{-4}
R_w	1.106×10^{-4}	9.13×10^{-5}	7.09×10^{-5}
4x			
$\hat{\Theta}$	0.00099	0.00102	0.00097
R_t	2.395×10^{-4}	2.195×10^{-4}	1.996×10^{-4}
R_w	6.28×10^{-5}	5.20×10^{-5}	3.84×10^{-5}
8x			
$\hat{\Theta}$	0.00099	0.00102	0.00098
R_t	2.206×10^{-4}	1.992×10^{-4}	1.845×10^{-4}
R_w	1.97×10^{-5}	2.14×10^{-5}	1.66×10^{-5}
16x			
$\hat{\Theta}$	0.00099	0.00102	0.00098
R_t	2.123×10^{-4}	1.922×10^{-4}	1.773×10^{-4}
R_w	5.7×10^{-6}	5.3×10^{-6}	4.1×10^{-6}

Here $\Theta = 0.001$ and $\rho = 10^{-4}$ in all cases. Twenty-five diploid samples were generated on the basis of constant population size model.

Output of polymorphic sites

At the conclusion of our algorithm, we calculate the posterior probability of polymorphism at each site by conditioning on the final estimated Θ -value and computing the posterior probability of polymorphism for each site as

$$\text{Prob}(\text{polymorphic} | R_i) = \frac{\sum_{\text{polymorphic } g} \text{Prob}(R_i | g) \text{Prob}(g | \Theta)}{\sum_g \text{Prob}(R_i | g) \text{Prob}(g | \Theta)}.$$

Since all sites will have a nonzero probability of being polymorphic, but for most sites that probability will likely be small, we adopt the following heuristic. We use a predetermined threshold T . If the posterior probability of polymorphism at a site exceeds T , that site is reported as a candidate polymorphic site. In the examples in this article we use $T = 0.9$. For sites that are predicted to be polymorphic, the genotype with the highest posterior probability is then output for each individual.

Results

Simulation of data

For all analyses in the article that are based upon simulated haplotypes, we simulated over a 100-kb region using Hudson's coalescent simulator ms (Hudson 2002) under a constant population-size model. We assume sequencing occurs on an individual level. Twenty diploid genomes were

Figure 1 The false-positive rates (FP) and false-negative rates (FN) for polymorphic site detection as a function of minor allele frequency. When ρ varies (top two plots), Θ is fixed at 0.001; when Θ varies (bottom two plots), ρ is fixed at 10^{-4} .

formed by random pairing of haplotypes. We then generated paired-end NGS data, with reads occurring as a Poisson process, each end consisting of a read of 35 bases and the distance between ends being Normally distributed with mean 500 and standard deviation of 20. Unless otherwise specified the NGS data were generated with an overall average error rate of 0.005, but with error rates increasing as the read is traversed. Specifically, the error rate at read position i was assumed to be

$$E_i = e^{\alpha + i\beta}, \quad (7)$$

with $\beta = 0.1$.

Estimation of Θ from simulated data

From the simulated data set we estimate Θ for a range of coverages per diploid and/or a range of sample sizes n . For each combination of parameters (mutation rate, Θ , recombination rate, ρ , coverage, X , and diploid sample size, n) we simulated 100 data sets. We assumed that the position-specific error rates were known. Table 1 shows results for a variety of values of ρ and X .

In the top half of Table 1 Θ is fixed at 0.001 and ρ varies, while in the bottom half Θ varies while ρ is fixed at 10^{-4} . We show the average of the estimated Θ -values, $\hat{\Theta}$, along with, R_Θ , defined as the root mean squared error (RMSE) of the ratio of our estimate of Θ compared to the true Θ value

(we normalized by dividing by Θ to ease interpretation). For benchmark purposes, we also compare our estimate of Θ , based upon the resequencing data, to that obtained by applying Watterson's estimator Watterson (1975) to the true (and, in fact, unobserved) haplotype data. For this purpose we report (R_w) , the root mean squared error of the ratio of Watterson's estimate of Θ compared to the true Θ . R_w represents a good estimate of the limit of what is attainable

Figure 2 The effect of minor allele frequency on polymorphic site detection. Twenty diploid samples were simulated with $\Theta = 10^{-3}$ and $\rho = 10^{-4}$. We show FN rate (y-axis) as a function of MAF (x-axis) for a variety of coverage levels.

Table 3 The effect of population growth on polymorphic site detection

Coverages	$r = 0$	$r = 10^{-4}$	$r = 10^{-3}$
1×	0.4700	0.5390	0.6861
2×	0.3246	0.3804	0.5194
4×	0.1804	0.2146	0.3159
8×	0.0564	0.0719	0.1092
16×	0.0063	0.0076	0.0105

Population data were simulated with different exponential growth rates r and current population size $N_0 = 10,000$ (see text for details). We report the false negative rate.

due to the stochastic noise within the coalescent process itself (since Watterson's estimate is known to be a good estimator of Θ , and we apply it here to the unobserved true genotypes).

The estimates of Θ are close to the true value of Θ for all combinations of parameters, even at 1× coverage. As expected, RMSE values are bigger when coverage is low. We also see that Θ is slightly overestimated for lower recombination rates, which is a reflection of the increased variance of our, or any other estimator of Θ as LD in the data increases.

In the lower part of Table 1 we show results for a range of Θ -values with ρ fixed. Once again the estimator appears to perform very well in all cases.

Next, in Table 2, we show results as a function of sample size for scenarios in which $\Theta = 0.001$ and $\rho = 10^{-4}$. Performance is robust to change of sample size n , with RMSEs decreasing as n increases, as is expected.

Detection of polymorphic sites

In Figure 1 we show the false-positive rate, FP, and false-negative rate, FN, as a function of coverage. For a range of Θ - and ρ -values we show the percentage of sites called as segregating that are not in fact polymorphic, FP (y-axis), against coverage X (x-axis). We also show plots for the FN rate. We see that both the FP and FN rates are reduced as coverages increases, as would be expected. Both FP and FN are robust to changes in recombination rate, but both are sensitive to changes in mutation rate. The power to detect polymorphic sites is affected by minor allele frequency (MAF). In Figure 2, we see that our method results in good detection rates for common variants (MAF > 0.05) given reasonable coverage (say 4×), but that, unsurprisingly, many rarer variants are missed, especially with lower coverages.

Given that our method is based on the constant population size model, it is important to explore robustness to departures from that assumption. We focus on situations in which, as is common, population size is growing. Two data sets were simulated under an exponentially growing population size, where the population size at time t units back into the past, $N(t)$, is modeled as $N(t) = N_0 e^{-rt}$, where N_0 is the current population size and r is the exponential growth rate. Table 3 shows the false-negative rates of polymorphic

site calling for three different growth rates. Since exponential growth leads to there being more rare variants, variation on average become harder to detect.

Effects of resequencing error rate

We also explored the effect on performance of changes to the overall base-calling error rate and to (read) position-specific base-calling error rates. Assuming exponentially distributed position-specific error rates, per Equation 7, we first explore changes to the overall rate, while setting $\beta = 0.1$ (both are assumed known). Table 4 shows that performance is robust to this change. However, the false-positive and -negative rates do grow higher as the overall error rate increases, as might be expected (Figure 3).

Second, we assumed the overall error rate is fixed, but changed β , reflecting varying degrees of degradation of calling accuracy along the reads. Once again our method is robust to this change (Table 5) but again the FP and FN rates are affected (Figure 3).

It is also of interest to see how performance is affected if incorrect error rate estimates are used during the analysis. This introduces biases to the estimated Θ -values. When the actual error rate is lower than the assumed error rate, Θ is underestimated, and when the actual error rate is higher than assumed, Θ is typically overestimated. These biases decreased as the level of coverage increases (Table 6).

It is also possible to jointly estimate (i) average error rate, E_t , and associated exponential parameter β , and (ii) Θ . Table 7 shows results for this case. For all coverage levels, the estimated values are close to the true values, and

Table 4 The mean of estimated Θ and RMSEs from resequencing data simulated with differing average error rates

Coverages	$E = 0.001$	$E = 0.005$	$E = 0.01$	$E = 0.05$
1×				
$\hat{\Theta}$	0.00103	0.00102	0.00102	0.00104
R_t	2.708×10^{-4}	2.852×10^{-4}	2.936×10^{-4}	3.518×10^{-4}
R_w	1.106×10^{-4}	1.294×10^{-4}	1.406×10^{-4}	2.121×10^{-4}
2×				
$\hat{\Theta}$	0.00102	0.00102	0.00103	0.00102
R_t	2.352×10^{-4}	2.529×10^{-4}	2.640×10^{-4}	3.203×10^{-4}
R_w	7.11×10^{-5}	9.13×10^{-5}	1.047×10^{-4}	1.752×10^{-4}
4×				
$\hat{\Theta}$	0.00102	0.00102	0.00102	0.00102
R_t	2.101×10^{-4}	2.195×10^{-4}	2.264×10^{-4}	2.820×10^{-4}
R_w	3.77×10^{-5}	5.20×10^{-5}	6.22×10^{-5}	1.277×10^{-4}
8×				
$\hat{\Theta}$	0.00102	0.00102	0.00102	0.00102
R_t	1.962×10^{-4}	1.992×10^{-4}	2.041×10^{-4}	2.415×10^{-4}
R_w	1.47×10^{-5}	2.14×10^{-5}	2.64×10^{-5}	7.92×10^{-5}
16×				
$\hat{\Theta}$	0.00102	0.00102	0.00102	0.00102
R_t	1.911×10^{-4}	1.922×10^{-4}	1.926×10^{-4}	2.071×10^{-4}
R_w	2.7×10^{-6}	5.3×10^{-6}	6.3×10^{-6}	3.3×10^{-5}

Data sets were simulated with $\Theta = 0.001$ and $\rho = 10^{-4}$. β is fixed at 0.1.

Figure 3 The effects of different resequencing error rates on polymorphic site detection.

accuracy of polymorphic site detection is much the same as for the analysis that assumes the true error rate (data not shown).

Comparison with other polymorphic site-detection methods

We next use simulated data to compare performance of our method, when focused on polymorphism detection, to that of other detection tools (samtools version 0.1.12) (Li and Leal 2009), and GATK Unified Genotyper (v. 1.0) (McKenna *et al.* 2010). To do this we created four 100-kb data sets, each containing 40 diploid samples, using four different mutation parameters ($\Theta = 0.00001, 0.0001, 0.0002, \text{ and } 0.0005$). This resulted in data with 52 (data set 1), 531 (data set 2), 915 (data set 3), and 2482 polymorphic sites (data set 4), respectively. We then simulated 100 replicate sets of read data at three different coverages ($1\times, 4\times, 8\times$). The simulated reads were then mapped with BWA (Burrows-Wheeler Alignment Tool) (Li and Durbin 2009). We then analyzed each of these data sets using our method, samtools, and GATK using default parameter values for the latter two methods. Figure 4 shows a summary of results over the 100 replicates. Results for our method are a function of the threshold, T ; for calling a polymorphism, this results in a curve of results, as a function of T , compared to the single datapoint for the other two approaches. For all data sets and coverages, our method shows good polymorphic site detection rate.

Analysis of 1000 Genomes Project data

We also applied our method to a 10-Mb region (16 Mb $\sim$ 26 Mb) on chromosome 21 from the 1000 Genomes Project data. We use only the Illumina data; for the purpose of simplicity, data from other platforms are ignored. Aligned NGS data from 57 African samples (Yoruba), 51 European samples [CEPH (Centre d'Etude du Polymorphisme Humain)], and 58 Asian samples [CHB (Chinese from Beijing) + JPN (Japanese)] were used. The analyzed data are from the Low Coverage Pilot project, in which the whole genome was sequenced at low coverage (the average coverage in the analyzed region is $3.4\times$). To improve computational efficiency and allow for local variation of population mutation rate, we divided the 10-Mb region into 100-kb subregions (with no overlap), and analyzed each separately. For each subregion, Θ -values were estimated and polymorphic sites and MAFs were predicted. We analyzed data from each population separately. For quality control, when applying our genotyping algorithm, reads with more than two mismatches to consensus sequence or lower mapping quality score (< 30) were not included in the analysis. The sequencing error rate per single read per site was inferred from the quality score, as in Cock *et al.* (2010). To improve comparability between results from our method, and those from 1000 Genomes, we use a probability threshold of $T = 0.9$, as is commonly used in similar analyses of 1000 Genomes Data (cf. Li *et al.* 2011). Results are shown in Figure 5.

Table 5 The mean of estimated Θ , and RMSEs, from resequencing data with differing error rate distributions along the read

Coverages	$\beta = 0$	$\beta = 0.1$	$\beta = 0.2$	$\beta = 1.0$
1x				
$\hat{\Theta}$	0.00103	0.00103	0.00102	0.00102
R_t	2.949×10^{-4}	2.852×10^{-4}	2.686×10^{-4}	2.292×10^{-4}
R_w	1.382×10^{-4}	1.294×10^{-4}	1.109×10^{-4}	6.59×10^{-5}
2x				
$\hat{\Theta}$	0.00102	0.00102	0.00102	0.00102
R_t	2.544×10^{-4}	2.529×10^{-4}	2.375×10^{-4}	2.138×10^{-4}
R_w	9.57×10^{-5}	9.13×10^{-5}	7.46×10^{-5}	4.31×10^{-5}
4x				
$\hat{\Theta}$	0.00103	0.00102	0.00102	0.00102
R_t	2.222×10^{-4}	2.195×10^{-4}	2.130×10^{-4}	2.000×10^{-4}
R_w	5.40×10^{-5}	5.20×10^{-5}	4.17×10^{-5}	1.92×10^{-5}
8x				
$\hat{\Theta}$	0.00102	0.00102	0.00102	0.00102
R_t	2.024×10^{-4}	1.992×10^{-4}	1.970×10^{-4}	1.911×10^{-4}
R_w	2.36×10^{-5}	2.14×10^{-5}	1.69×10^{-5}	6.2×10^{-6}
16x				
$\hat{\Theta}$	0.00102	0.00102	0.00102	0.00102
R_t	1.916×10^{-4}	1.922×10^{-4}	1.913×10^{-4}	1.906×10^{-4}
R_w	5.2×10^{-6}	5.3×10^{-6}	3.8×10^{-6}	1.4×10^{-6}

Data sets were simulated with $\Theta = 0.001$ and $\rho = 10^{-4}$. The average error rate E is fixed at 0.005.

As expected, the estimated Θ from the African data are bigger than from other populations. The average values over the region for the African, European, and Asian populations

Table 6 The mean of estimated Θ from resequencing data with differing average resequencing error rate E_t

Coverages	$E_t = 0.005$	$E_t = 0.01$	$E_t = 0.015$
1x			
$\hat{\Theta}$	0.00071	0.00102	0.00247
R_t	3.572×10^{-4}	2.852×10^{-4}	1.5456×10^{-3}
R_w	3.245×10^{-4}	1.294×10^{-4}	1.4808×10^{-3}
2x			
$\hat{\Theta}$	0.0078	0.00102	0.00213
R_t	2.988×10^{-4}	2.529×10^{-4}	1.1981×10^{-3}
R_w	2.523×10^{-4}	9.13×10^{-5}	1.132×10^{-3}
4x			
$\hat{\Theta}$	0.0087	0.00102	0.00167
R_t	2.309×10^{-4}	2.195×10^{-4}	7.329×10^{-4}
R_w	1.579×10^{-4}	5.20×10^{-4}	6.562×10^{-4}
8x			
$\hat{\Theta}$	0.00096	0.00102	0.00128
R_t	1.918×10^{-4}	1.992×10^{-4}	3.597×10^{-4}
R_w	6.08×10^{-5}	2.1×10^{-5}	2.074×10^{-4}
16x			
$\hat{\Theta}$	0.00102	0.00102	0.00108
R_t	1.898×10^{-4}	1.922×10^{-4}	2.128×10^{-4}
R_w	8.4×10^{-6}	5.3×10^{-6}	5.51×10^{-5}

Data were simulated with $\Theta = 0.001$, $\rho = 10^{-4}$, and $\rho = 0.1$.

Table 7 The mean of co-estimated Θ , average error rate E and β

Coverages	$\hat{\Theta}$	E	β
1x	0.00103	0.00499	0.1005
2x	0.00103	0.00500	0.1001
4x	0.00102	0.00500	0.1000
8x	0.00102	0.00500	0.1000
16x	0.00102	0.00500	0.1000

The data were simulated with $\Theta = 0.001$, $\rho = 10^{-4}$, $E = 0.005$, and $\beta = 0.1$.

are 0.00112, 0.000742, and 0.000650, respectively. The local variation of estimated Θ -values between populations is highly correlated. The non-African populations show highest correlation ($R^2 = 0.727$). The correlation between African and non-African populations is also high (African-European $R^2 = 0.560$ and African-Asian $R^2 = 0.489$) but some regions show differences. (17.9–18.1 Mb, 21.6–21.7 Mb, 24.1–24.4 Mb, and 25.3–25.5 Mb). These lower estimates for the non-African populations may be evidence of local selection, but might also reflect systematic differences in data production across centers, or simply noise.

Figure 6 shows a summary of our polymorphic site calls. First we show the number of polymorphic sites detected. Second, we calculate and report the expectation of the minor allele frequency, using the posterior genotype probabilities at all sites that are called polymorphic. More polymorphic sites are inferred from the African population than from the Asian or European populations. The mean MAF at polymorphic sites from the African, Asian, and European populations are about 0.145, 0.189, and 0.169, respectively. This is bigger than the expected MAF derived from even a constant population size model (≈ 0.053). This reflects an ascertainment bias due to the difficulty of detecting rare variants. The average MAF of population-private polymorphic sites is 0.048 and is much lower than polymorphic sites common to all three populations (0.22), again as would be expected (assuming that private polymorphisms are likely to have arisen more recently).

For the Yoruban population we compared our polymorphic site calls to those in dbSNP (version 129) and to calls from the 1000 Genomes Project (Figure 7). A comparison between estimated MAFs of novel and known variants shows that the polymorphic sites reported in dbSNP are relatively common variants. About 82% of variable sites from the 1000 genome study are inferred as polymorphic by our method. Our method infers 2410 more known sites than the 1000 Genomes Project. As a quality check of polymorphic site calling, we calculated the transition/transversion ratios (ts/tv) using the predicted genotypes at sites that were detected as polymorphic. The overall ts/tv ratio using calls from our algorithm is 2.28, but the ts/tv ratio of novel sites is significantly lower (1.83), suggesting a greater number of false positives at novel sites.

Our method can simultaneously estimate resequencing error rates. We therefore estimated error rates for the 1000 Genome Project data. The error rate of each read position is

Figure 4 Polymorphic site-calling performances compared across three methods. Each data set was resequenced with different coverages: 1× (left), 4× (middle) and 8× (right). The figure shows the true positive rate (y-axis) and false-positive rate (x-axis) of the analysis results from three methods: samtools (triangle), GATK unified Genotyper (circle), and our proposed method (line).

estimated from data for the Yoruban population, using a constant-sized population model. Results are shown in Figure 8. Our estimates are highly correlated with the error rate reported in 1000 Geomes; however, our estimated error rates show a systematic tendency to be lower than those derived during the 1000 Genomes analysis. This may reflect differences in the preprocessing of the read data. For example, in our analysis all reads with two or more mismatches were excluded.

Discussion

In this article we have presented a method for both (i) estimation of overall mutation rate Θ , and (ii) calling polymorphic sites, within a region for which NGS data have been collected. The posterior probability of a site being polymorphic is computed for each site, along with an estimate of Θ . If position-specific error rates are not known, or the accuracy

of estimates of these are in doubt, our method can also estimate those error rates within its likelihood framework. However, in general, estimation of error rates is unlikely to be needed. Our method appears to work well, in part at least because it exploits the coalescent model to leverage power when genotyping jointly across multiple individuals.

There are at least three sources of information that can be used when calling genotypes at a given locus, L , for a given individual, I , (say):

1. The read data for that individual at that position;
2. the data for other individuals at that position; and
3. the data at other positions.

Each of these three sources could be used to inform either genotype calls or determination of whether the locus is polymorphic. When interpreting the first source of information, the key question to address is “Do the data result from genuine heterozygosity within that individual, or are they a

Figure 5 The distribution of Θ over the 10-Mb region of chromosome 21 of the African, European, and Eastern Asian populations from 1000 genome sequence data.

result of base-calling errors from the sequencing platform?" A variety of methods exists for dealing with this issue, such as those compared in this article. Intuitively speaking, the likelihood of interpreting a set of read data at L for I as evidence for heterozygosity should increase if polymorphism at that locus is also observed in other individuals. (Once we have seen a variant allele at L we are more likely to see it at that location in subsequent individuals.) This information is exploited by our algorithm. Patterns of linkage disequilibrium (LD) are also informative regarding genotyping (and hence calling of polymorphism). Loosely speaking, we should be more likely to call a polymorphism at L if the data support a pattern of polymorphism at L that would be in LD with other nearby loci. Detailed modeling of patterns of LD is a complex issue, generally implying significant computational complexity.

It is also worth noting that LD information will not greatly help to detect a polymorphism that has not already been observed in an existing sample. For these reasons we omit it from the present method, but our results show that

Figure 6 The number of inferred polymorphic sites, mean of the expected MAF (in parentheses), and ts/tv ratio (in *italics*) in each population in the 10-Mb region of chromosome 21.

Figure 7 A comparison of inferred polymorphic sites in the African population using our analysis, calls from the 1000 Genomes Project, and known polymorphic sites from dbSNP. We show the number of inferred polymorphic sites, mean of the expected MAF, and the ts/tv ratio.

our method performs well despite making this simplification (a choice that ensures computational tractability). Nonetheless, it has been successfully exploited using a number of different underlying probabilistic models, in algorithms such as fastPHASE and Li and Stephens' PAC method) (Li and Stephens 2003; Scheet and Stephens 2006). However, applications to SNP data are of much lower dimension than to NGS data, and it is not clear how tractable those methods might remain when applied to NGS data, where we are looking at many millions of bases. Neither will it be straightforward to allow for the specific error structure of read data, in which, broadly speaking, error rates increase as we move along the read. It is reasonable to think that algorithms that eventually also include this third source of information will result in a further improvement in performance over this method, but the computational challenges will be significant. Our method exploits the first two sources of

Figure 8 The inferred position-specific error rates. The estimated error rates for each read position are reported (solid line). For comparison, the error rates from 1000 Genome Project are reported (dashed line).

information and already requires significant computational resources, even though it does not use LD information for genotyping. However, we have shown that it results in good ability to accurately infer mutation rates and to call genotypes.

In this article, our method was applied to simulated data in which we assumed that reads had been aligned without error. This is not, of course, the case with real data (e.g., our analysis of the 1000 Genomes data) or for the analysis in Figure 4, in which read alignment algorithms were applied to simulated read data. Our method does work well in these more realistic scenarios, but it is important to note that the read alignment process can be expected to induce biases in estimation of mutation rate and/or polymorphism calling, for example, in regions in which alignment is difficult (e.g., near indels). Since most alignment approaches discard reads that have a nontrivial number of mismatches, a downward bias in the estimation of Θ can be expected to occur (since some of these mismatches presumably represent true polymorphism.)

Software for implementation of our method is available upon request from the authors. We note that, as well as outputting most likely genotype calls for loci that are determined to be polymorphic, the algorithm can be used to output posterior genotype probabilities at those loci. This would allow for explicit incorporation of uncertainty in genotype calls to be incorporated in subsequent analyses, such as genome-wide association studies.

Acknowledgments

The authors gratefully acknowledge funding from National Institutes of Health award 3R01 MH084678 and the helpful comments of two reviewers.

Literature Cited

- 1000 Genomes Project Consortium, 2010 A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.
- Bansal, V., O. Harismendy, R. Tewhey, S. S. Murray, N. J. Schork *et al.*, 2010 Accurate detection and genotyping of snps utilizing population sequencing data. *Genome Res.* 20: 537–545.
- Cock, P. J. A., C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, 2010 The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res.* 38: 1767–1771.
- Frazer, K., S. Murray, N. Schork, and E. Topol, 2009 Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* 10: 241–251.
- Griffiths, R., and S. Tavaré, 1994 Ancestral inference in population genetics. *Stat. Sci.* 9: 307–319.
- Griffiths, R., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. *Stoch. Models* 14: 273–295.
- Hein, J., M. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution*. Oxford University Press, Oxford.
- Hellmann, I., Y. Mang, Z. Gu, P. Li, F. De La Vega *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequence from multiple individuals. *Genome Res.* 18: 1020–1029.
- Hudson, R., 2001 Two-locus sampling distributions and their application. *Genetics* 159: 1805–1817.
- Hudson, R., 2002 Generating samples under a Wright–Fisher neutral model. *Bioinformatics* 18: 337–338.
- Jiang, R., S. Tavaré, and P. Marjoram, 2008 Population genetic inference from resequencing data. *Genetics* 181: 187–197.
- Kingman, J., 1982a The coalescent. *Stoch. Proc. Appl.* 13: 235–248.
- Kingman, J., 1982b Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino. North-Holland Publishing, Amsterdam.
- Kingman, J., 1982c On the genealogy of large populations. *J. Appl. Probab.* 19A: 27–43.
- Knudsen, B., and M. Miyamoto, 2009 Accurate and fast methods to estimate the population mutation rate from error-prone sequences. *BMC Bioinformatics* 10: 247.
- Kuhner, M. K., J. Yamato, and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* 140: 1421–1430.
- Li, B., and S. Leal, 2009 Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet.* 5: e1000481.
- Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics* 25: 1754–1760.
- Li, H., J. Ruan, and R. Durbin, 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18: 1851–1858.
- Li, N., and M. Stephens, 2003 Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* 165: 2213–2233.
- Li, R., Y. Li, K. Kristiansen, and J. Wang, 2008 SOAP: short oligonucleotide alignment program. *Bioinformatics* 24: 713–714.
- Li, Y., C. Sidore, H. M. Kang, M. Boehnke, and G. Abecasis, 2011 Low coverage sequencing: implications for the design of complex trait association studies. *Genome Res.* 21: 940–951.
- Manolio, T., F. Collins, N. Cox, D. Goldstein, L. Hindorf *et al.*, 2009 Finding the missing heritability of complex diseases. *Nature* 461: 747–753.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.
- Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78: 629–644.
- Shendure, J., and H. Ji, 2008 Next-generation DNA sequencing. *Nat. Biotechnol.* 26: 1135–1145.
- Wakeley, J., 2008 *Coalescent Theory: An Introduction*. Roberts & Company, Greenwood Village, CO.
- Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7: 256–276.

Communicating editor: Y. S. Song