BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 20 2012, pages 2569–2575
doi:10.1093/bioinformatics/bts501

Genome analysis

Advance Access publication August 22, 2012

An empirical Bayes mixture model for SNP detection in pooled sequencing data

Baiyu Zhou*

Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Detecting single-nucleotide polymorphism (SNP) in pooled sequencing data is more challenging than in individual sequencing because of sampling variations across pools. To effectively differentiate SNP signal from sequencing error, appropriate estimation of the sequencing error is necessary. In this article, we propose an empirical Bayes mixture (EBM) model for SNP detection and allele frequency estimation in pooled sequencing data.

Results: The proposed model reliably learns the error distribution by pooling information across pools and genomic positions. In addition, the proposed EBM model builds in characteristics unique to the pooled sequencing data, boosting the sensitivity of SNP detection. For large-scale inference in SNP detection, the EBM model provides a flexible and robust way for estimation and control of local false discovery rate. We demonstrate the performance of the proposed method through simulation studies and real data application.

Availability: Implementation of this method is available at https://sites.google.com/site/zhouby98

Contact: baiyu.zhou@einstein.yu.edu

Received on April 22, 2012; revised on July 6, 2012; accepted on August 8, 2012

1 INTRODUCTION

Recent advances in next-generation sequencing (NGS) technologies have significantly reduced the cost and time involved in sequencing the human genome in hundreds or thousands of individuals. The resulting comprehensive genomic analyses will provide powerful tools for discovering the genetic variations underlying complex phenotypic traits. NGS holds the promise of revolutionizing genetic studies due to its ability to detect rare variants that are not observed in array-based genome-wide association studies (GWAS). The importance of rare variants underlying complex diseases has been discussed in a number of recent articles (Bodmer and Bonilla, 2008; Cohen *et al.*, 2006; Ji *et al.*, 2008). It is believed that rare variants account for a significant proportion of genetic heritability missed by previous GWAS (Manolio *et al.*, 2009). Although the cost of NGS has dropped dramatically over the past few years, it is still prohibitively expensive to sequence individual genome at a sufficient read depth. An alternative approach to reduce the cost of sample preparation and sequencing is to pool genomic DNA from multiple individuals and sequence the pooled DNA samples (Kim *et al.*, 2010; Wang *et al.*, 2010). The feasibility of pooled sequencing has been demonstrated in several studies (Bansal *et al.*, 2011; Nejentsev *et al.*, 2009). Nejentsev *et al.* (2009) used pooled sequencing to identify causal rare variants in candidate genes of Type I diabetes.

When analyzing NGS data, the first step is to identify polymorphic sites in the genomic regions. A number of computational methods have been developed to identify single-nucleotide polymorphisms (SNPs) in individual sequenced samples (Bansal *et al.*, 2010; Koboldt *et al.*, 2009; Li *et al.*, 2008; Martin *et al.*, 2010; Muralidharan *et al.*, 2011; Zhou and Whittemore, 2012). These methods use sequencing and mapping quality scores to distinguish true SNPs from sequencing and mapping errors. Some methods utilize cross-sample information to estimate sequencing error at each genomic position (Bansal *et al.*, 2010; Martin *et al.*, 2010; Muralidharan *et al.*, 2011; Zhou and Whittemore, 2012). In addition, incorporation of pedigree information and linkage disequilibrium (LD) has been demonstrated to significantly improve SNP calling accuracy in individual sequencing (Zhou and Whittemore, 2012).

Methods for individual sequencing do not directly apply to SNP calling in pooled sequencing. Characteristics unique to pooled sequencing make SNP calling more challenging than in individual sequencing. First, because of sampling variation, the frequency of a variant allele will vary across the pools. Allele frequency in a pool is a function of the population allele frequency and pool size (number of individuals in a pool; Bansal, 2010). In cases in which there exists a rare variant with small pool size, the presence of the variant allele will be sporadic across pools. In contrast, in individual sequencing, the variant allele frequency for a diploid sample is 0, 0.5 or 1. Variation of allele frequency across pools makes it difficult to differentiate the true SNP signal from sequencing error. Second, external information, such as pedigree and LD information, is not available in pooled sequencing. Third, because of the small number of pools, it is almost impossible to get a good estimate of the sequencing error using only cross-pool information. Therefore, sophisticated methods are needed to call SNPs in pooled sequencing.

Recently, several SNP calling methods designed for pooled sequencing have been reported (Bansal, 2010; Druley *et al.*, 2009; Wei *et al.*, 2011). SNPSeeker (Druley *et al.*, 2009), derived from large deviation theory, requires negative control data to estimate error model. CRISP (Comprehensive Read analysis for Identification of Single Nucleotide Polymorphisms from Pooled sequencing, Bansal, 2010) models allele counts across multiple pools using contingency tables. It uses Fisher's exact test to detect uneven distributions of variant reads across pools and then uses Chernoff bound to compute an upper bound on the *P*-value of the overabundance of alternate alleles within each

*To whom correspondence should be addressed.

pool against the sequencing error. This method builds on the assumption that the presence of a rare variant in a pool is likely to cause excess variant reads compared with the other pools. Such an assumption does not hold for less rare or common variant and is questionable when the pool size is large. Consequently, CRISP is not effective for detection of less rare and common SNPs. As an alternative approach, SNVer (Wei *et al.*, 2011) models variant reads in a single pool using a binomial–binomial model and combines *P*-values from multiple pools to identify SNPs. Because it is difficult to simultaneously estimate the population allele frequency and sequencing error using the binomial–binomial model, SNVer arbitrarily specifies the sequencing error to be 0.01 for all positions. As we show in the simulation studies, inaccurate specification of sequencing error will lead to high false-positive rates and low sensitivity.

In this article, we introduce an empirical Bayes mixture (EBM) model for detecting SNPs in pooled sequencing data. In contrast to the existing methods that detect SNPs one at a time, the EBM model learns error distribution by pooling information across multiple pools and genomic positions. It is well known that the sequencing error at each position is highly reproducible across samples or pools. Cross-sample information in individual sequencing has been gathered to estimate sequencing error. However, in pooled sequencing, direct estimation sequencing error has not been explored due to the small number of pools available and the variations of allele frequency across pools. Additional information is needed to obtain a good estimate of the sequencing error in pooled sequencing. Using the EBM model, data across genomic positions can be pooled to learn the error distribution. Since the density of a SNP is estimated to be approximately 1 of 1000 bases (Wei *et al.*, 2011), leveraging the information from a large number of null positions will allow fairly accurate estimation of the error distribution.

The fundamental idea of pooling information to estimate null distribution in empirical Bayes methods has been applied to genomic data since the advent of microarray analysis (Efron *et al.*, 2001). Recently, Muralidharan *et al.* (2011) extended Efron's empirical Bayes method to detect SNPs in individual sequencing. In this article, different from previous methods, we propose an EBM model that builds in characteristics unique to pooled sequencing for SNP detection. The EBM model blurs the line between hypothesis testing and parameter estimation. In addition to SNP detection, the proposed method produces empirical Bayes shrinkage estimation of allele frequency. Allele frequency estimation is important for downstream analyses, such as association studies or population genetic analysis. Yet, it is left unexplored by the current SNP detection methods. In the EBM model, SNP detection and allele frequency estimation are nicely integrated. In the following sections, we will present mathematical formulation of the EBM model and demonstrate its performance through simulation studies and a real data application.

2 METHODS

Our objective is to build an EBM model to pool information across genomic positions and utilize sequence reads from multiple pools to identify SNPs. Consider a sequencing study in which a targeted region of I nucleotides is sequenced in K pools. Each pool contains n diploid individuals (i.e. $2n$ haplotypes, n is not necessarily equal across pools). We assume that the reads from each pool have been aligned to a reference genome. Suppose there are N_{ij} reads covering position i in pool j, among which V_{ij} are mismatches (V_{ij} is the count of the most frequent non-reference base and we ignore reads of other bases). In the absence of a variant, mismatched reads are merely the result of sequencing error (base calling error or/and alignment error) and are expected to be similar across pools since most sequencing errors are local context dependent (Bansal, 2010). On the other hand, the presence of a variant will result in excessive mismatches in one or more pools, and the variant allele reads distribution is likely to be uneven across pools for rare variants. Existing methods, such as CRISP, use the uneven distribution across pools as a signal to identify SNPs. Our proposed EBM model also incorporates such cross-pool information to enhance the sensitivity of SNP detection. In addition, the EBM model pools data across genomic positions to learn the error distribution. Due to small number of pools, sequencing error is difficult to estimate from the data for each individual position. To overcome this difficulty, empirical Bayes approach specifies a prior distribution on position-specific parameters and learns the prior distribution by aggregating the data across different positions. Simultaneously incorporating information across pools and positions increases the sensitivity for detection of true SNP sites. Next, we present the detailed formulation of the EBM model.

2.1 EBM model

For a particular position i in pool j, we model the mismatch counts V_{ij} as drawn from a binomial distribution $V_{ij} \sim Binomial(N_{ij}, p_{ij})$. Instead of specifying an EBM model for V_{ij}, we first apply a normalizing and variance stabilizing transformation to the data (V_{ij}, N_{ij}):

$$Z_{ij} = \arcsin\sqrt{\frac{V_{ij} + \frac{1}{4}}{N_{ij} + \frac{1}{2}}} \tag{1}$$

Brown (2008) has shown that transformed data are approximately normal for moderate or large size of N_{ij} (for example, $N_{ij} \geq 12$): $Z_{ij} \sim N(\mu_{ij}, \frac{1}{4N_{ij}})$ and $\mu_{ij} = \arcsin\sqrt{p_{ij}}$. By using equation (1), the count data are transformed to more familiar Gaussian data. Importantly, the transformation equation (1) ensures the variance of Z_{ij} is known and independent of μ_{ij}.

Data across pools are collected into vectors. Let $\mathbf{Z}_i = (Z_{i1,} \ldots, Z_{iK})'$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iK})'$. Bold letters are used to denote vectors and matrices. We assume $(\mathbf{Z}_i, \boldsymbol{\mu}_i)$ are generated from the following hierarchical model:

$$\boldsymbol{\mu}_i \sim h(\boldsymbol{\mu}_i), \mathbf{Z}_i|\boldsymbol{\mu}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{2}$$

where $\boldsymbol{\Sigma}_i = \text{diag}(\frac{1}{4N_{i1}}, \ldots, \frac{1}{4N_{iK}})$ is the diagonal matrix whose elements are $\frac{1}{4N_{ij}}$. In this hierarchical model, Z_{ij} is independently normal with a variance of $\frac{1}{4N_{ij}}$, conditioned on the mean vector $\boldsymbol{\mu}_i$. The empirical Bayes approach does not specify the prior distribution h in advance. Instead, it models h using a mixture of multivariate Gaussians and estimates the mixture distribution from the data. The mixture Gaussians can approximate any smooth density and gives a flexible and nearly non-parametric way to model the prior distribution (Liao et al., 2004; Muralidharan, 2010; Pan et al., 2003). Specifically, let

$$h(\boldsymbol{\mu}_i) = \sum_{g=1}^{G_0+G_1} \pi_g h^{(g)}(\boldsymbol{\mu}_i), \tag{3}$$

where each component $h^{(g)}(\boldsymbol{\mu}_i) = N(\mathbf{v}_g, \Lambda_{gi})$ is a multivariate normal with mean $\mathbf{v}_g$ and covariance Λ_{gi}. π_g are unknown mixture proportions. In the mixture distribution, $g = 1, 2 \ldots, G_0$ are null components used to capture the sequencing error distribution. The remaining non-null components capture true SNP signals that are beyond the error distribution.

For a null position i, we expect its p_{ij}s (i.e. sequencing error) is approximately the same across pools, so $\boldsymbol{\mu}_i$ is a constant vector with equal components: $\mu_{i1} = \mu_{i2} = \cdots = \mu_{iK}$. To impose this constraint on the null components of equation (3), let $\mathbf{P} = K^{-1}\mathbf{ee}'$ be the $K \times K$ projection matrix onto the rank 1 space of constant vectors, where $\mathbf{e}' = (1, \ldots, 1)$ is a $K \times 1$ vector of 1 s. Let $\mathbf{Q} = \mathbf{I}_K - \mathbf{P}$ be the projection matrix onto the orthogonal complement of $R(\mathbf{P})$. We can write any vector $\boldsymbol{\mu}_i = \mathbf{P}\boldsymbol{\mu}_i + \mathbf{Q}\boldsymbol{\mu}_i$. For the null position i, the second term $\mathbf{Q}\boldsymbol{\mu}_i$ vanishes. Given $\boldsymbol{\Sigma}_i$, we model $\boldsymbol{\mu}_i$ drawn from the g th null component as

$$\boldsymbol{\mu}_i|\boldsymbol{\Sigma}_i \sim N(\nu_g\mathbf{e}, \tau_g\mathbf{P}\boldsymbol{\Sigma}_i\mathbf{P}), \tag{4}$$

where $\nu_g$and τ_g are scalars. The covariance matrix $\tau_g\mathbf{P}\boldsymbol{\Sigma}_i\mathbf{P}$ guarantees that $\boldsymbol{\mu}_i$ is a constant vector. For a SNP position i, whose $\boldsymbol{\mu}_i$ is drawn from a non-null component g, we model the distribution of $\boldsymbol{\mu}_i$ as

$$\boldsymbol{\mu}_i|\boldsymbol{\Sigma}_i \sim N(\boldsymbol{\nu}_g, \tau_g\mathbf{P}\boldsymbol{\Sigma}_i\mathbf{P} + \kappa_g\mathbf{Q}\boldsymbol{\Sigma}_i\mathbf{Q}). \tag{5}$$

The extra component $\kappa_g\mathbf{Q}\boldsymbol{\Sigma}_i\mathbf{Q}$ adds further variance to $\boldsymbol{\mu}_i$ so that the $\boldsymbol{\mu}_i$ becomes a non-constant vector. The hyper-parameters $(\boldsymbol{\nu}_g, \tau_g, \kappa_g)$ will be estimated from the data by pooling information across positions.

Under the hierarchical model (2), the marginal distribution of $\mathbf{Z}_i$is the mixture of multivariate Gaussians:

$$f(\mathbf{Z}_i) = \sum_{g=1}^{G_0+G_1} \pi_g f^{(g)}(\mathbf{Z}_i), \tag{6}$$

where $f^{(g)}(\mathbf{Z}_i) = N(\nu_g\mathbf{e}, \tau_g\mathbf{P}\boldsymbol{\Sigma}_i\mathbf{P} + \boldsymbol{\Sigma}_i)$ for $g = 1, \ldots, G_0$, $f^{(g)}(\mathbf{Z}_i) = N(\boldsymbol{\nu}_g, \tau_g\mathbf{P}\boldsymbol{\Sigma}_i\mathbf{P} + \kappa_g\mathbf{Q}\boldsymbol{\Sigma}_i\mathbf{Q} + \boldsymbol{\Sigma}_i)$ for $g = G_0 + 1, \ldots, G_0 + G_1$. The posterior distribution of $\boldsymbol{\mu}_i|\mathbf{Z}_i$ is also the mixture of Gaussians. It can be easily shown that

$$\boldsymbol{\mu}_i|\mathbf{Z}_i \sim \sum_{g=1}^{G_0+G_1} p^{(g)}(\mathbf{Z}_i)h^{(g)}(\boldsymbol{\mu}_i|\mathbf{Z}_i), \tag{7}$$

where $p^{(g)}(\mathbf{Z}_i) = (\pi_g f^{(g)}(\mathbf{Z}_i))/(f(\mathbf{Z}_i))$ is the posterior probability that $\mathbf{Z}_i$ is drawn from component g, and $h^{(g)}(\boldsymbol{\mu}_i|\mathbf{Z}_i)$ is the posterior distribution if the prior were $h^{(g)}$. The structure of the model is illustrated in Figure 1.

With this model, we proceed in three steps to call SNPs. We first estimate the parameters $\boldsymbol{\theta} = (\pi_g, \boldsymbol{\nu}_g, \tau_g, \kappa_g)$ by maximizing the marginal likelihood using a modified expectation–maximization (EM) algorithm. Details are given in the next section. We then use the estimated parameters to find the posterior probability that $\mathbf{Z}_i$ is drawn from the null components of equation (3). Finally, we use the estimated posterior probability to call SNPs. The posterior probability that $\mathbf{Z}_i$ is drawn from the null components relates to the local false discovery rate (fdr). In statistical testing framework, let H_i be the null hypothesis that the position $i$is non-SNP. The fdr is defined as the posterior probability $fdr(z) = P(H_i \text{ is true } |\mathbf{Z}_i = z)$. It is related to the commonly used FDR and possesses features that are more attractive than FDR in certain scenarios. See Efron (2007) and Liao *et al.* (2004) for comprehensive discussions about FDR. The EBM model makes it straightforward to calculate FDR. Specifically,

$$fdr(\mathbf{Z}_i) = \sum_{g=1}^{G_0} p^{(g)}(\mathbf{Z}_i) = \sum_{g=1}^{G_0} \frac{\pi_g f^{(g)}(\mathbf{Z}_i)}{f(\mathbf{Z}_i)}. \tag{8}$$

Positions with FDR less than a pre-specified threshold are identified as SNP sites.

For SNP sites, we estimate the mean vector $\boldsymbol{\mu}_i$ by its posterior mean:

$$\hat{\boldsymbol{\mu}}_i = E(\boldsymbol{\mu}_i|\mathbf{Z}_i) = \sum_{g=1}^{G_0+G_1} p^{(g)}(\mathbf{Z}_i)E^{(g)}(\boldsymbol{\mu}_i|\mathbf{Z}_i), \tag{9}$$

where $E^{(g)}(\boldsymbol{\mu}_i|\mathbf{Z}_i)$ is the posterior mean under $h^{(g)}(\boldsymbol{\mu}_i|\mathbf{Z}_i)$. The empirical Bayes estimator shrinks the positional estimate toward the common mean shared across positions. The allele frequency estimate is obtained by $\hat{p}_i = \frac{1}{K}\sum_{j=1}^{K} \sin(\hat{\mu}_{ij})^2$.

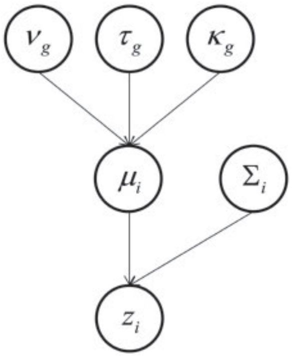

Fig. 1. Hierarchical structure of the EBM model. $\mathbf{Z}_i$ is observed. ν_g, τ_g and κ_g are hyper-parameters to be estimated from data

2.2 Model fitting and parameter estimation

We fit the EBM model using the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993). The ECM algorithm replaces the M-step of the usual EM algorithm with a series of partial maximizations. It is computationally simpler than using the EM algorithm, yet possesses many of the same convergence properties.

We first define indicator $\delta_{ig} = I(\boldsymbol{\mu}_i \text{ is from component } g)$ for each position i and treat it as missing data. The log-likelihood for the complete data $(\mathbf{Z}_i, \delta_{ig})$ is $l(\mathbf{Z}, \delta) = \sum_{i=1}^{I}\sum_{g=1}^{G_0+G_1} \delta_{ig}(\log\pi_g + f^{(g)}(\mathbf{Z}_i))$. In the E-step, we compute $E(\delta_{ig}|\mathbf{Z})$ at the current vales of the parameters. It is then straightforward to show $E(\delta_{ig}|\mathbf{Z}) = \pi_g f^{(g)}(\mathbf{Z}_i)/\sum_{g=1}^{G_0+G_1} \pi_g f^{(g)}(\mathbf{Z}_i)$. In the CM step, we sequentially optimize parameters using the expected values of the indicators (for simplicity, we will write $\delta_{ig}$for $E(\delta_{ig}|\mathbf{Z})$). π_g is updated by $\hat{\pi}_g = \frac{1}{I}\sum_{i=1}^{I} \delta_{ig}$. $\boldsymbol{\nu}_g$ is estimated by $\hat{\boldsymbol{\nu}}_g = \sum_i \delta_{ig}\mathbf{Z}_i/\sum_i \delta_{ig}$ for non-null components and $\hat{\nu}_g = \sum_i \delta_{ig}(\mathbf{PZ}_i)/\sum_i \delta_{ig}$ for null components, where $\mathbf{PZ}_i$ is the projection of $\mathbf{Z}_i$ on $R(\mathbf{P})$. Using the fitted $\hat{\pi}_g$ and $\hat{\boldsymbol{\nu}}_g$, we estimate τ_g and κ_g by maximizing the marginal likelihood of $\mathbf{Z}$.

The EBM model has two tuning parameters: the number of null components, G_0, and the number of non-null components, G_1. The tuning parameters can be determined using various model selection criteria, such as Akaike Information Criterion or the Bayesian Information Criterion (Liao *et al.*, 2004; Muralidharan, 2010; Pan *et al.*, 2003). The tuning parameters are not critical for the purpose of fitting distribution function (Pan *et al.*, 2003; Muralidharan, 2010). In the simulations and real data analysis, we choose $G_0 = 2$ and $G_1 = 4$. We use two null components to capture the sequencing error distribution and two non-null components to capture SNP signals beyond the sequencing error. We put other two non-null components at the same location of the null components but add $\kappa_g\mathbf{Q}\boldsymbol{\Sigma}_i\mathbf{Q}$ to the covariance matrix to allow cross-pool variations. These two non-null components are used to detect rare variants, whose signals are at the same level or lower than the sequencing error but exhibit cross-pool variations.

For starting values, we use $\pi_g = 0.498$ and $\pi_g = 0.001$ for the null and non-null components. We initialize the center $\boldsymbol{\nu}_g$ of the two null components (and two non-null components) at 40% and 80% quantile of the data. The centers of the two remaining non-null components are set to be at the 92.5% and 97.5% quantile of the data. We initialize τ_g and κ_g to be 1. After initialization, the ECM iterations will fine-tune the parameters and reach the final estimate.

3 RESULTS

3.1 Simulations studies

We simulate synthetic data to evaluate the performance of the EBM model and compare it with other methods. Data are

Table 1. AUC and false-positive rate of the EBM model

Read depth		10			50		
Pool size		10	25	50	10	25	50
Error	MAF	AUC					
0.01	0.01	0.809	0.955	0.994	0.817	0.956	0.994
	0.05	0.998	1	1	0.998	1	1
	0.1	0.999	1	1	1	1	1
0.05	0.01	0.762	0.802	0.842	0.822	0.954	0.986
	0.05	0.994	0.997	0.998	0.995	0.999	0.999
	0.1	0.999	0.999	1	0.999	0.999	1
Error	fdr	False-positive rate ($\times10^{-4}$)					
0.01	0.01	0.2	1.0	0.4	1.5	1.6	1.1
	0.05	1.2	2.2	1.2	2.4	2.7	2.5
	0.1	2.0	2.9	1.5	3.7	3.0	3.6
0.05	0.01	1.1	1.2	1.1	1.2	1.5	0.7
	0.05	2.7	2.2	2.2	2.6	3.0	2.1
	0.1	3.6	2.7	3.2	3.4	3.7	3.4

MAF: minor allele frequency; error: sequencing error generated from uniform (0, 0.01) or uniform (0, 0.05).

generated for a region of 1000 positions. Assuming the first 30 positions are SNPs, we simulate sequencing studies with five pools. Simulations are conducted under each combination of the following settings: (i) minor allele frequencies (MAF) of SNPs are 0.01, 0.05 and 0.1, respectively; (ii) the sequencing coverage is 10× or 50× per haplotype; (iii) the sequencing error of each position is generated from a random uniform distribution (0, e), where $e = 0.01$ or 0.05 and (iv) the number of individuals in each pool (i.e. pool size) is 10, 25 or 50.

We generate 100 datasets under each setting and rank the position based on the posterior probability of being a SNP. Positions with high posterior probability of being drawn from the non-null components in equation (3) are ranked at the top of the list. Then, we assess the quality of ranking by computing the area under the receiver operating characteristic (ROC) curve averaged over the 100 datasets. The ROC curve is the plot of 1-specificity versus sensitivity, which is equivalent to the Type I error rate versus 1-Type II error rate. The area under the curve (AUC) value close to one indicates high sensitivity and high specificity in identifying true SNPs.

Table 1 shows the AUC of the EBM model for a variety of settings. Several observations are noteworthy. First, increasing read depth significantly improves the sensitivity of SNP detection for rare variants with a high sequencing error. The largest improvement of AUC (~19%) by increasing read depth comes from MAF of 0.01 and sequencing error of 0.05. Second, large sample size helps improve sensitivity, particularly in detecting rare variants. This is because with larger sample size, rare variants are more likely to be included in pools. For example, the probability of presence in at least one pool (assuming five pools in total) for a variant with $\text{MAF} = 0.01$ is 99.3% if the pool size includes 50 samples. This probability dramatically decreases to 63.3% if the pool size is 10 (i.e. with probability 36.7%, the SNP can never be identified no matter how sensitive the method is). Third, the EBM model is sensitive enough to identify SNPs as long as the read depth is sufficient even in situations where the signal is weaker than the sequencing error. For example, with MAF 0.01 and pool size of 50, the SNP signal (1%) is lower than the sequencing error of 0.05, but the EBM model still achieves AUC of 0.986 under the read depth of 50. The EBM model is sensitive enough to detect rare variants because of two reasons. First, it borrows information from other genomic positions to help estimate the error distribution. Second, it detects the uneven distribution of variant reads across pools when a rare variant is present.

Next, we evaluate the false-positive rate of the EBM model, which is controlled by specifying the fdr to be less than a certain threshold. For the data generated above, we call SNPs by setting the threshold $\text{fdr} \leq 0.001$ (0.05 or 0.1). The FDR is estimated by equation (8). For each threshold, the FDR and false-positive rate are well controlled (data not shown). We then generate sequencing data as previously described, except all 1000 positions are non-SNP sites. We use the threshold $\text{fdr} \leq 0.001$ (0.05 or 0.1) to call SNPs. Table 1 shows the false-positive rate under each threshold, demonstrating that it is effectively controlled.

We then compare the efficiency of ranking candidate SNPs by different methods. We focus on comparisons with CRISP and SNVer, since they have been shown to outperform other methods for SNP calling in pooled sequencing (Bansal, 2010; Wei *et al.*, 2011). CRISP uses Fisher's exact test followed by additional filtering steps. The positions are ranked based on the *P*-values from Fisher's exact test. In contrast, SNVer first computes a *P*-value based on a binomial–binomial model for a single pool and then uses Simes' method to calculate a pooled *P*-value for the multiple pools. Positions are ranked by the pooled *P*-value. Table 2 shows the AUC comparisons of the three methods. We can see that the EBM model outperforms other methods in every scenario. CRISP has a good sensitivity

Table 2. Comparisons of AUC of different methods for calling SNPs

MAF	0.01		0.05		0.1	
Error	0.01	0.05	0.01	0.05	0.01	0.05
EBM	0.994	0.842	1	0.998	1	1
CRISP	0.967	0.826	0.972	0.946	0.969	0.954
SNVer	0.985	0.760	1	0.886	1	0.886

Shown are AUCs under different settings of MAF (minor allele frequency) and error (sequencing error) for pooled sequencing with read depth of 10 per haplotype and pool size of 50.

for SNP detection when the variant is rare but does less well when the allele frequency becomes less rare or common. This is because the working hypothesis of CRISP (i.e. uneven distribution of variant reads across pools) does not hold well for less rare and common variants. SNVer performs reasonably well when the assumed error rate matches the true sequencing error, but is the least efficient in ranking candidate SNPs when the true sequencing error deviates from the specified error rate. In practice, the sequencing error is unknown and varies across position. Specifying an arbitrary error rate for all positions would be problematic for application of SNVer to real data.

Different from the previous methods, the EBM model integrates hypothesis testing and allele frequency estimation in a single procedure: By fitting the hierarchical model (2), we obtain the posterior probability that $\mathbf{Z}_i$ is drawn from the null components of the prior distribution (3), based on which the position $i$is declared as SNP or null. Meanwhile, we use the empirical Bayes estimator $\hat{\boldsymbol{\mu}}_i$ to estimate the minor allele frequency of a SNP i, i.e. $\hat{p}_i = \frac{1}{K}\sum_{j=1}^{K} \sin(\hat{\mu}_{ij})^2$. As an alternative, the naive estimator for MAF of SNP i is $\hat{q}_i = \frac{1}{K}\sum_{j=1}^{K} V_{ij}/N_{ij}$. We compare the accuracy of these two estimators by simulating 1000 positions, among which the first 30 positions are SNPs with allele frequencies ranging from 0.01 to 0.3, each incrementing by 0.01. Estimation accuracy is measured by the total sum of squared errors (SSE) defined as: $SSE = \sum_{i=1}^{30}(\hat{p}_i - p_i)^2$ or $SSE = \sum_{i=1}^{30}(\hat{q}_i - p_i)^2$ for the two estimators respectively, where p_i is the true allele frequency. The SSE averaged over 100 datasets is plotted in Figure 2 for a variety of settings. The EB estimator outperforms the naive estimator in every setting. The improvement is more noticeable when the sample size is small. By pooling information across positions, the EB estimator overcomes the small sample size and improves the allele frequency estimation.

3.2 Real data application

We applied the EBM model to analyze pooled sequencing data downloaded from Bansal *et al.* (2011). One hundred HapMap samples including 20 Utah residents with ancestry from northern and western Europe (CEU), 20 Han Chinese (CHB), 20 Tuscan in Italy (TSI) and 40 Yoruba (YRI) were sequenced in five pools. The samples were pooled by population with each pool consisting of 20 individuals (YRI were divided into two equal pools: YRI_1 and YRI_2). Approximately 600 kb of coding regions were sequenced by Illumina GAIIx. The sequenced reads were mapped to reference human genome (HG18) by Burrows-Wheeler Alignment (BWA) (Li and Durbin, 2009), resulting in averaged coverage of 34× per individuals. Figure 3 shows the distribution of coverage per haplotype.

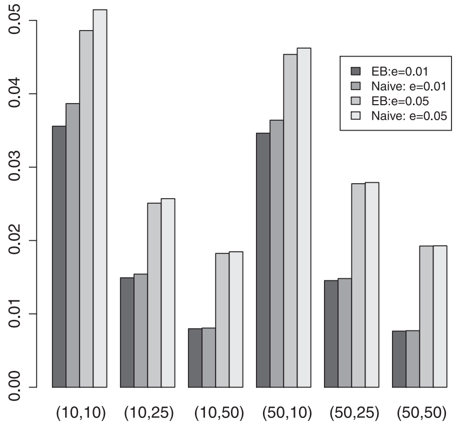

Fig. 2. Comparisons of total SSE for EB estimator and naïve estimator. The y-axis is the SSE. The pairs of numbers on x-axis are read depth and pool size. For example, (10, 25) represents read depth of 10 and pool size of 25. $e = 0.01$ (or $e = 0.05$) represents the sequencing error generated from uniform (0, 0.01) [or uniform (0, 0.05)]

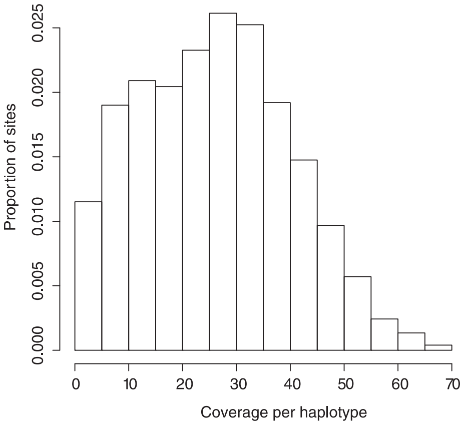

Fig. 3. Empirical distribution of sequence coverage per haplotype (chromosome 16)

We use the EBM model to identify SNPs in the ~20 kb sequenced coding region on chromosome 16. Before applying the EBM model, the data were preprocessed by filtering out reads with base calling quality score <10 and mapping quality score <20. Positions with coverage <20 in one or more pools were not included in the analysis. After applying the EBM model, 93 SNPs were identified at the threshold of fdr ≤ 0.001. As a comparison, we applied CRISP with the default parameters and identified 75 SNPs. Sixty SNPs were identified by both methods. Fifteen SNPs were identified by CRISP alone, and 33 SNPs were identified by the EBM model alone (Fig. 4a). Figure 4b plots the distribution of allele frequencies (averaged over five pools) of the 60 SNPs identified by both methods. Among these SNPs, 66.7% had MAF <0.05, 10% had MAF

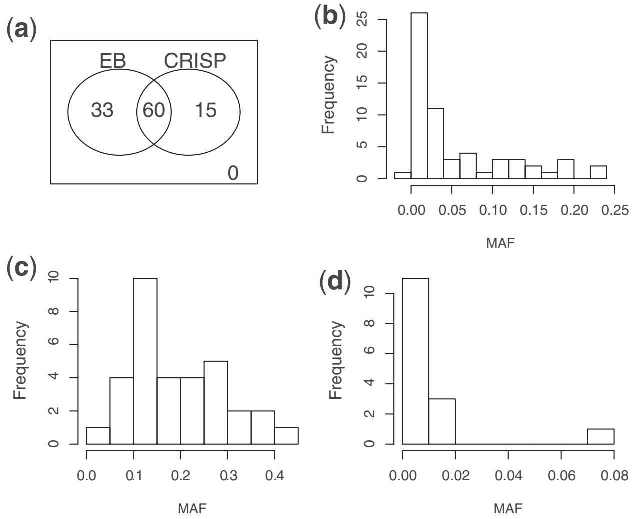

Fig. 4. (**a**) Venn diagram of the SNPs detected by the EBM model and CRISP; (**b**) distribution of allele frequencies of the 60 SNPs detected by both methods; (**c**) distribution of allele frequencies detected by the EBM model alone and (**d**) distribution of allele frequencies detected by CRISP alone

between 0.05 and 0.1 and 23.3% had MAF > 0.1. We compare with a list of variants on the same set of 100 samples from the HapMap and the 1000 Genomes Projects (HM + 1KG dataset). Due to the missing data in the 1000 genomes dataset for three of the pools, the HM + 1KG dataset is only a partial list of the variants. Forty-four (73%) of the 60 SNPs are on the list of HM + 1KG dataset.

For the 16 SNPs identified by both the EBM model and CRISP but not on the list of HM + 1KG, we inspect the aligned sequence reads from the pooled data and found that 10 of the 16 sites are strongly supported by the pooled sequence data as SNP sites. Among the 33 SNPs identified by the EBM model but not by CRISP, 97% (32 SNPs) have allele frequencies >0.05 (Fig. 4c). The excess variant allele reads strongly suggest these sites are true SNPs. Of note, CRISP failed to detect these SNP sites, which is consistent with the simulations results showing that CRISP is not sufficiently sensitive for detection of less rare and common SNPs. We further inspected the 15 SNPs identified by CRISP but not the EBM model. Fourteen of the SNPs were rare (MAF < 0.05; Fig. 4d), and eight of them had the variant allele present in only one pool. Three of the 15 sites are strongly supported by the alignment data to be true SNPs. The remaining sites were inconclusive based on the alignment data.

We then compared the allele frequency estimated by the empirical Bayes shrinkage estimator and the naive estimator. We calculated the total SSE between the estimated allele frequencies with those obtained from 1000 genome data for the overlapping 44 SNPs. The SSEs were 0.102 and 0.185 when using empirical Bayes estimate and the naive estimate, respectively. Correlations of the estimated allele frequencies and the 1000 genome allele frequencies were 0.973 for the empirical Bayes estimate and 0.951 for the naive estimate. The empirical Bayes estimator performs slightly better than the naive estimator.

4 DISCUSSIONS

We have developed an EBM model for SNP calling in the analysis of pooled sequencing data. The EBM model is able to learn the error distribution reliably by borrowing information across pools and positions. Because of the small number of pools, it is necessary to borrow information across positions to infer the error distribution in pooled sequencing. Muralidharan *et al.* (2011) developed an empirical Bayes model to detect SNPs in individual sequencing, but their model is not applicable to pooled sequencing. The proposed EBM model effectively incorporates characteristics of the pooled sequencing data to increase the sensitivity of SNP detection. Simulation studies and real data application have demonstrated the superior performance of the EBM model over other methods.

In contrast to previously available methods, such as CRISP, which use *P*-value as threshold to call SNPs, the EBM model uses fdr as the threshold for SNP calling thereby controlling the false-positive rate. It has been established that the FDR and fdr are more appropriate than *P*-values for large-scale inference of genomic data (Benjamini and Hochberg, 1995; Storey, 2002). The concept of fdr was developed by Efron *et al.* (2001) for the analysis of microarray data. In the context of SNP calling, fdr quantifies the site-specific evidence indicating that each position is a SNP site. The usual FDR, on the other hand, averages over other sites with stronger evidence. Several authors have argued that FDR may result in misleading inferences and advocated the use of fdr in genomic studies (Finner and Roters, 2002; Liao *et al.*, 2004). However, fdr is often more difficult to estimate than FDR. The FDR can be formulated in terms of the cumulative distribution functions, for which the empirical distribution is a consistent and stable estimator. But to estimate fdr, it is necessary to estimate the density of distributions, which requires parametric models or non-parametric smoothing methods (Liao *et al.*, 2004). The EBM approach provides a simple and flexible way to estimate fdr (Liao *et al.*, 2004; Muralidharan, 2010; Pan *et al.*, 2003) and effectively controls the false-positive rate. As a comparison, the *P*-value approach of CRISP has been shown to result in inflated type I errors (Wei *et al.*, 2011).

Another advantage of the EBM model is that it integrates the allele frequency estimation in a single procedure with SNP detection. The EBM model estimates allele frequency using an empirical Bayes shrinkage estimator, which shrinks the SNP-specific MAF toward a genome-wide consensus. Statistical properties of the empirical Bayes estimator have been extensively discussed in the literature (Brown, 2008; Efron and Morris, 1975; Robbins, 1954). We have shown here that it outperforms the naive estimator for MAF estimation especially when the sample size is small. In NGS, sampling bias is a non-trivial issue. Reads coverage is almost certainly non-uniform across positions. By considering one position at a time, previous methods completely ignore the non-uniform distribution of reads coverage. The EBM model is able to incorporate coverage information in the covariance matrix of equation (2) to reflect the noise level of different positions.

Before applying the EBM model, appropriate filtering of sequence reads is necessary. The EBM model uses transformation (1), which requires the coverage to be at least 12 or so (Brown, 2008). Positions with read coverage <12 are not suitable for transformation as the mean and variance formula no longer hold well. It is recommended to discard or manually inspect these poorly covered positions. Reads with low base calling or mapping qualities should also be excluded. It is possible to apply more sophisticated filtering strategies to eliminating systematic error or strand bias, see Rivas *et al.* (2011) for detailed discussions.

The EBM model can be extended in several ways for future work. First, detection of small indels is not supported by our current model. The indels impose a great challenge for NGS. DNA amplification and reads mapping techniques to handle indels are under development. In future work, we may incorporate indels calling in the EBM model. Second, we may extend the EBM framework to include downstream analysis, such as association analysis of case–control data. Currently, there is a lack of methods for association analysis for pooled sequencing data. In summary, we have developed an efficient EBM model for SNP calling in pooled sequencing data that is applicable to NGS data generated by various sequencing platforms.

ACKNOWLEDGEMENTS

I would like to thank Dr Mimi Kim, Dr Kenny Ye, Dr Harry Ostrer and members of Ostrer group for their constructive comments.

Funding: This work was supported by Einstein initiative grant.

Conflict of Interest: none declared.

REFERENCES

Bansal,V. (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.

Bansal,V. *et al.* (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.*, **20**, 537–545.

Bansal,V. *et al.* (2011) Efficient and cost effective population resequencing by pooling and in-solution hybridization. *PLoS One*, **6**, e18353.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B.*, **57**, 289–300.

Bodmer,W. and Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.*, **40**, 695–701.

Brown,L.D. (2008) In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Statist.*, **2**, 113–152.

Cohen,J.C. *et al.* (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl. Acad. Sci. USA.*, **103**, 1810–1815.

Druley,T.E. *et al.* (2009) Quantification of rare allelic variants from pooled genomic DNA. *Nat. Methods*, **6**, 263–265.

Efron,B. (2007) Size, power and false discovery rates. *Ann. Statist.*, **35**, 1351–1377.

Efron,B. *et al.* (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Efron,B. and Morris,C. (1975) Data analysis using Stein's estimator and its generalizations. *J. Amer. Stat. Assoc.*, **70**, 311–319.

Finner,H. and Roters,M. (2002) Multiple hypotheses testing and expected number of type I errors. *Ann. Stat.*, **30**, 220–238.

Ji,W. *et al.* (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.

Kim,S.Y. *et al.* (2010) Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet. Epidemiol.*, **34**, 479–491.

Koboldt,D.C. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Liao,J.G. *et al.* (2004) A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, **20**, 2694–2701.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Martin,E.R. *et al.* (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, **26**, 2803–2810.

Meng,X. and Rubin,D.B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Muralidharan,O. (2010) An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.*, **4**, 422–438.

Muralidharan,O. *et al.* (2011) A cross-sample statistical model for SNP detection in short-read sequencing data. *Nucleic Acids Res.*, doi: 10.1093/nar/gkr851.

Nejentsev,S. *et al.* (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.

Pan,W. *et al.* (2003) A mixture model approach to detecting differentially expressed genes with microarray data. *Funct. Integr. Genomics*, **3**, 117–124.

Rivas *et al.* (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **43**, 1066–1073.

Robbins,H. (1954) An empirical Bayes approach to statistics. In Neyman,J. (ed.) *Proc. Thrid Berkeley Sympos. Math. Statist. Probab. 1*. Univ. California Press, Berkeley, CA, pp. 157–163.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

Wang,T *et al.* (2010) Resequencing of pooled DNA for detecting disease associations with rare variants. *Genet. Epidemiol.*, **34**, 492–501.

Wei,Z *et al.* (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.*, **39**, e132, doi: 10.1093/nar/gkr599.

Zhou,B. and Whittemore,A.S. (2012) Improving sequence-based genotype calls with linkage disequilibrium and pedigree information. *Ann. Appl. Stat.*, **6**, 457–475.