Bioinformatics Advance Access published July 24, 2016

Bioinformatics, 2016, 1–6
doi: 10.1093/bioinformatics/btw409
Advance Access Publication Date: 4 July 2016
Original Paper

Sequence analysis

An empirical Bayes method for genotyping and SNP detection using multi-sample next-generation sequencing data

Gongyi Huang[1], Shaoli Wang[2,3], Xueqin Wang[1,4,5] and Na You[1,4,*]

[1]School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China, [2]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China, [3]Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai 200433, China, [4]South China Center for Statistical Science, Sun Yat-sen University, Guangzhou 510275, China and [5]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 4, 2016; revised on June 18, 2016; accepted on June 20, 2016

Abstract

Motivation: The development of next generation sequencing technology provides an efficient and powerful approach to rare variant detection. To identify genetic variations, the essential question is how to quantity the sequencing error rate in the data. Because of the advantage of easy implementation and the ability to integrate data from different sources, the empirical Bayes method is popularly employed to estimate the sequencing error rate for SNP detection.

Results: We propose a novel statistical model to fit the observed non-reference allele frequency data, and utilize the empirical Bayes method for both genotyping and SNP detection, where an ECM algorithm is implemented to estimate the model parameters. The performance of our proposed method is investigated via simulations and real data analysis. It is shown that our method makes less genotype-call errors, and with the parameter estimates from the ECM algorithm, it attains high detection power with FDR being well controlled.

Availability and implementation: The proposed algorithm is wrapped in the R package ebGenotyping, which can be downloaded from http://cran.r-project.org/web/packages/ebGenotyping/.

Contact: youn@mail.sysu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next-generation sequencing (NGS) technology produces vast amount of sequencing data at an unprecedented low cost, which totally changes the landscape of identification of genetic variations (Schuster, 2008; Snyder *et al.*, 2010). Thanks to its high-throughput capability, NGS provides a direct and powerful approach to the detection of rare variants (Li and Leal, 2009; Zhao *et al.*, 2013). It makes it possible to explore the DNA sequence at nucleotide level and identify mutations through single-sample analysis. However, it is demonstrated that multi-sample sequencing can increase the detection power and has better false positive control than single-sample sequencing (Le and Durbin, 2011; Murillo *et al.*, 2016), therefore, multi-sample sequencing is more widely used in the genomic research, and the analysis tools for multi-sample sequencing data are always in high demand.

As the most common type of sequence variations, the single nucleotide polymorphism (SNP) plays an important role in genetic evolution. The existence of SNPs changes the genotypes of sequenced sample different from the reference genome, resulting in the non-reference alleles being observed in the alignment data. On the contrary, the genomic positions with frequent non-reference alleles indicate potential SNPs. Nevertheless, the NGS data analysis pipeline is a complex process. Many factors during this process may contribute to the non-reference alleles, including not only the sample preparation, and PCR amplification during the sequencing experiment, but also the algorithms used for base calling and alignment. All the alleles different from the genotype allele due to these factors are called as sequencing errors. The existence of sequencing errors makes SNP detection much more challenging than expected.

The main task of SNP detection is to evaluate the sequencing error rate in the data, in order to distinguish SNPs from the sequencing errors. The base calling quality and mapping quality scores are widely used to quantity the sequencing error rate in the NGS data for SNP detection, such as GATK (DePristo *et al.*, 2011), SAMtools (Li *et al.*, 2009a), SOAPsnp (Li *et al.*, 2009b), FreeBayes (https://github.com/ekg/freebayes), Atlas-SNP2 (Shen *et al.*, 2010), etc. With an arbitrary prior, they employ the classical Bayes model to call the genotypes according to their posterior probabilities, and then call SNPs by comparing to the reference genome. Although the quality scores reflect the magnitude of the sequencing error rate to a great extent, they still have some limitations. First, the quality scores may not be able to fully account for the sequencing error rate due to the complexity of NGS data analysis pipeline. You *et al.* (2012) propose to incorporate the sample preparation error in single-sample analysis to evaluate the sequencing error rate. Their method, as well as the corresponding multi-sample method MultiGeMS (Murillo *et al.*, 2016), outperforms other SNP callers. Second, processing the quality metrics is computationally intensive and those algorithms are usually not applicable across different experimental platforms. Third, using the quality scores to determine the sequencing error rate ignores the common information from repeated alleles, such as the sequencing error rate related to the local DNA content, which is shared by the alleles aligned to that DNA region from different samples (Muralidharan *et al.*, 2012b).

Recently, more and more SNP callers based on the empirical Bayes method were proposed, where the sequencing error rate is set to be an unknown parameter and empirically estimated from the data. Martin *et al.* (2010) assume that different samples have the same sequencing error rate at each genomic locus, and use the allele frequency data from all of the samples to estimate it. Muralidharan *et al.* (2012b) employ a Dirichlet mixture model to fit the frequency vectors of four nucleotides at all the genomic loci of different samples, and combine all the data to estimate the model parameters. Muralidharan *et al.* (2012a) decompose the sequencing error rate into three types of variations, i.e. sample effect, positional effect and finite depth, and integrate the data across samples and genomic locations to estimate them. Besides, Zhou (2012) and Zhao *et al.* (2013) utilize the empirical Bayes method for pooled sequencing data analysis. Comparing to the classical Bayes models, the empirical Bayes method can combine the information from different samples at different genomic sites to improve the estimation for model parameters (Efron, 2010; Zhao *et al.*, 2013).

In this paper, we focus on the multi-sample sequencing data from the diploid organism. At each genomic locus, only three possible genotypes are considered, i.e. homozygous reference (RR), which is non-SNP, heterozygous SNP (RV) and homozygous SNP (VV). Although this may cause the incorrect genotype calls at triallelic sites, it need not significantly affect the results but highly reduces the computational burden (Murillo, *et al.*, 2016, Supplementary Material). With this assumption, the Binomial model is a natural choice to fit the allele frequency, as did Martin *et al.* (2010) and Muralidharan *et al.*(2012a). Martin *et al.* (2010) assume the independence between different genomic sites and estimate the sequencing error rate site-by-site, while Muralidharan *et al.* (2012a) introduce the sample effect and integrate the data not only across samples but also genomic sites to estimate the sequencing error rate. Comparing to Martin *et al.* (2010), Muralidharan *et al.* (2012a) improve the statistical model to admit heterogeneity among samples. However, since their model is built to fit the sequencing error frequency which is not observable, they have to incorporate a pre-genotyping step. Besides, due to the complexity of the model, they estimate the parameter using the median method and approximate the logit-normal by Beta distribution to calculate the posterior distribution.

We modify the decomposition in Muralidharan *et al.* (2012a) and propose a novel statistical model for both of genotyping and SNP detection. The model is established to fit the observed allele frequency data, so we do not need any pre-processing analysis. Moreover, benefitting from the modification, instead of the approximation methods, we implement an ECM algorithm (Meng and Rubin, 1993) to estimate the model parameters, and utilize the empirical Bayes method to genotype samples and call SNPs with false discovery rate (FDR) control. The rest of the paper is structured as follows. In Section 2, the statistical model and ECM algorithm are introduced. Their performance is illustrated via simulations in Section 3 and two real datasets in Section 4. Finally, we give a short discussion in Section 5.

2 Methods

Consider there are n sequenced samples and m genome sites. Let N_{ij} be the total number of aligned reads of sample j covering the genomic locus i, and X_{ij} be the number of non-reference alleles among N_{ij}. We assume that

$$X_{ij} \sim \text{Bin}(N_{ij}, p_{ij}),$$

where p_{ij} is the probability that a non-reference allele is being observed at this position, and $\text{Bin}(N, p)$ indicates the Binomial distribution with size N and success probability p. As mentioned previously, both the sequencing errors and SNPs can cause the non-reference alleles. Let Z_{ij} be the genotype of sample j at the genomic site i. In the simple case that there are no sequencing errors, $p_{ij} = 0$ for $Z_{ij} = \text{RR}$, $p_{ij} = 1/2$ for $Z_{ij} = \text{RV}$, and $p_{ij} = 1$ for $Z_{ij} = \text{VV}$. However, the existence of sequencing errors makes the modelling of p_{ij} difficult.

Denoting by ϕ_{ij} the sequencing error rate of sample j at genomic locus i, Muralidharan *et al.* (2012a) decompose $\text{logit}(\phi_{ij})$ into three components, i.e. $\text{logit}\phi_{ij} = \mu_i + \delta_j + \epsilon_{ij}$, where μ_i is the positional effect, δ_j is the sample-specific effect, and ϵ_{ij} follows a normal distribution $N(0, \sigma_j^2)$ measuring the variation within a sample caused by finite depth. The μ_i, δ_j and σ_j^2 are all unknown parameters and need to be estimated from the data. Due to the fact that the sequencing errors are unobservable, they have to involve a pre-genotyping step to remove the genotype effect on the observed allele frequency data to estimate those parameters.

The pre-processing analysis not only involves additional computations, but also may introduce errors. In order to model the genotype effect on the non-reference allele frequency, we define a latent variable Z_{ij} to indicate the unobserved genotype. As stated previously, only three possible genotypes are considered, therefore Z_{ij} is coded as -1, 0 and 1 corresponding to the genotype RR, RV and VV, respectively. When $Z_{ij} = -1$ (RR), the non-reference alleles are sequencing errors, so $p_{ij} = \phi_{ij}$. When $Z_{ij} = 1$(VV), the reference alleles are errors, so that $p_{ij} = 1 - \phi_{ij}$. When $Z_{ij} = 0$(RV), the error probability of R being called as V equals to that of V being called as

R, resulting in $p_{ij} = 1/2$. In summary, the model is presented as follows:

$$X_{ij} \sim \mathrm{Bin}(N_{ij}, p_{ij}),$$

$$p_{ij} = \begin{cases} \phi_{ij}, & Z_{ij} = -1, \\ 1/2, & Z_{ij} = 0, \ \mathrm{logit}\,\phi_{ij} = \mu_i + \delta_j. \\ 1 - \phi_{ij}, & Z_{ij} = 1, \end{cases}$$

Note that we exclude the finite sampling effect ϵ_{ij} from the decomposition of logitϕ_{ij}, since the Binomial distribution accounts for the sampling variation due to finite depth. We assume Z_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, are independently and identically distributed with $\Pr(Z_{ij} = -1) = p_{-1}$, $\Pr(Z_{ij} = 0) = p_0$, $\Pr(Z_{ij} = 1) = p_1$, and $p_{-1} + p_0 + p_1 = 1$.

The unknown parameters μ_i, $i = 1, 2, \ldots, m$, δ_j, $j = 1, 2, \ldots, n$, p_{-1} and p_0 are estimated and used to calculate the posterior probabilities of $Z_{ij} = \Delta$, $\Delta \in \{-1, 0, 1\}$,

$$\widehat{z}_{ij,\Delta} = \Pr(Z_{ij} = \Delta | \widehat{\mu}_i, \widehat{\delta}_j, \widehat{p}_{-1}, \widehat{p}_0, X_{ij}, N_{ij}),$$

where $\widehat{\mu}_i$, $\widehat{\delta}_j$, $\widehat{p}_{-1}$ and $\widehat{p}_0$ are the parameters' estimates obtained from the following ECM iteration. It is worth noting that, the exclusion of ϵ_{ij} from the decomposition of logitϕ_{ij} makes the calculation of $\widehat{z}_{ij,\Delta}$ much easier, whereas Muralidharan *et al.* (2012a) approximate the logit-normal distribution with a Beta distribution to calculate the marginal and posterior distributions. The genotype of sample j at genome locus i is assigned by

$$\widehat{\Delta}_{ij} = \underset{\Delta \in \{-1,0,1\}}{\mathrm{argmax}}\ \widehat{z}_{ij,\Delta}.$$

For SNP detection, we propose using the posterior probability $\widehat{z}_{ij,-1}$ as the estimate for local FDR, according to the definition in Efron (2010). Given a pre-specified cutoff α, the positions with $\widehat{z}_{ij,-1}$ smaller than α are called as SNPs, and non-SNPs otherwise.

In order to estimate the unknown parameters in our model, instead of the median method for approximate estimation, we implement an ECM algorithm (Meng and Rubin, 1993). It is clear that the complete log-likelihood function is

$$\begin{aligned} l_c = \sum_{i=1}^{m}\sum_{j=1}^{n} \Big\{ & I(Z_{ij}{=}-1)[X_{ij}(\mu_i + \delta_j) - N_{ij}\log(1 + \exp(\mu_i + \delta_j)) \\ & + \log p_{-1}] + I(Z_{ij} = 0)(-N_{ij}\log 2 + \log p_0) \\ & + I(Z_{ij}{=}\ 1)[(N_{ij} - X_{ij})(\mu_i + \delta_j) - N_{ij}\log(1 + \exp(\mu_i + \delta_j)) \\ & + \log(1 - p_{-1} - p_0)] \Big\}. \end{aligned}$$

Given the parameter estimates from the kth iteration, in the $(k+1)$th E-step, we have

$$z_{ij,\Delta}^{(k+1)} = P(Z_{ij} = \Delta | \mu_i^{(k)}, \delta_j^{(k)}, p_{-1}^{(k)}, p_0^{(k)}, X_{ij}, N_{ij}).$$

In the $(k+1)$th M-step, p_{-1} and p_0 admit closed-form maximizers, i.e.

$$p_{-1}^{(k+1)} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij,-1}^{(k+1)}}{mn}$$

and

$$p_0^{(k+1)} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} z_{ij,0}^{(k+1)}}{mn}.$$

The other parameters do not have closed-form maximizers. According to the ECM algorithm (Meng and Rubin, 1993), in order to simplify the optimization, we partition the parameter space $(\mu_1, \mu_2, \ldots, \mu_m, \delta_1, \delta_2, \ldots, \delta_n)^\top$ into two subspaces $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_\mathrm{m})^\top$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_\mathrm{n})^\top$, and find their maximizers conditionally. Given δ, all off-diagonal elements of the Hessian matrix $H_{c,\mu}$ are zero, and all diagonal elements are negative, where

$$H_{c,\mu} = \frac{\partial Q_c^2}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^\top},$$

$$Q_c = E(l_c | \boldsymbol{\mu}^{(k)}, \boldsymbol{\delta}^{(k)}, p_{-1}^{(k)}, p_0^{(k)}, X_{ij}, N_{ij}).$$

It indicates that optimizing Q_c over μ reduces to optimizing Q_{ci} over μ_i, where Q_{ci} is the ith element of Q_c in the summation, i.e. $Q_c = \sum_{i=1}^{m} Q_{ci}$. According to this property, we utilize the bisection method to find the unique maximizers $\mu_i^{(k+1)}$ of Q_{ci}, given that $\boldsymbol{\delta} = \boldsymbol{\delta}^{(k)}$, $\mathrm{p}_{-1} = \mathrm{p}_{-1}^{(k+1)}$ and $p_0 = p_0^{(k+1)}$, $i = 1, 2, \ldots, m$. The parameter $\boldsymbol{\delta}$ shows the same property as $\boldsymbol{\mu}$, and then is updated similarly, given $\boldsymbol{\mu} = \boldsymbol{\mu}^{(k+1)}$, $\mathrm{p}_{-1} = \mathrm{p}_{-1}^{(k+1)}$ and $p_0 = p_0^{(k+1)}$. The E-step and M-step iterate until the parameters converge.

The model has the nonidentifiability problem if we do not set any constraints on the parameter space, as discussed in Martin *et al.* (2010). It is necessary to restrict $\phi_{ij} < 0.5$, i.e. $\mu_i + \delta_j < 0$. In the above bisection method, we search the interval $(-M, \min(-\delta_1, -\delta_2, \ldots, -\delta_n))$ for the maximizer $\mu_i^{(k+1)}$ and the interval $(-M, \min(-\mu_1, -\mu_2, \ldots, -\mu_m))$ for $\delta_j^{(k+1)}$, where M is a large positive constant, being set to be 10 as default in our algorithm.

3 Simulations

The performance of our proposed method is investigated via simulations. The numbers of non-reference alleles X_{ij} and coverage N_{ij} of $n = 300$ samples at $m = 10\,000$ genomic sites are randomly generated. Without loss of generalization, we set the first $m_0 = 9500$ loci not bearing any SNPs and all the SNPs are distributed in the last $m_1 = 500$ genomic sites. At the SNP locus, let Q be the probability that a sample carries SNP, and R be the probability that the SNP is homozygous. We set $Q = 0.05$ or 0.8 to reflect the rare or common variant, respectively, and let $R = 0, 0.8$ or 1 indicate different types of mutations. Given Q and R, the genotype indicators $(Z_{ij})_{m\times n}$ are generated from the binomial distributions $\mathrm{Bin}(1, Q)$ and $\mathrm{Bin}(1, R)$. The positional effects μ_1, μ_2, $\ldots$, μ_m are independently generated from a distribution F and sample effects δ_1, δ_2, $\ldots$, δ_n are from a distribution G. We set F and G to be $-3 + \xi$ or the degenerate distribution at -3, corresponding to whether or not there exists the positional effect or sample effect, where ξ is a random variable following the Gamma distribution with shape = 2 and scale = 1/2, $\Gamma(2, 2)$. Note that the distribution of F or G does not affect the performance of our proposed method, since we do not make any assumption on the distribution of μ_i, $i = 1, 2, \ldots, m$, or δ_j, $j = 1, 2, \ldots, n$ in the estimation, but treat all of them as unknown parameters and estimate them from the data. Let $\phi_{ij} = \exp(\mu_i + \delta_j)/(1 + \exp(\mu_i + \delta_j))$. Given the coverage N_{ij} from $\mathrm{Bin}(N, 1/2)$, X_{ij} is generated from the Binomial mixture distribution $\mathrm{Bin}(N_{ij}, p_{ij})$, where $p_{ij} = \phi_{ij}$ when Z_{ij} is RR, 1/2 when Z_{ij} is RV, and $1 - \phi_{ij}$ when Z_{ij} is VV. We vary F, G, Q, R and N and get a series of simulation experiments with different parameter combinations, as listed in Table 1.

Table 1. The $-\log_{10}$ (genotype-call error rate) of seqEM (Martin et al., 2010) and our proposed method in simulation experiments, where ξ is a random variable following the Gamma distribution $\Gamma(2,2)$, and -3 indicates the degenerate distribution at -3

Exp	Q	R	F	G	$N=50$		$N=100$	
ID					Proposed	seqEM	Proposed	seqEM
1	0.8	0.8	-3	-3	5.25	5.12	Inf	Inf
2	0.8	0.8	-3	$\xi-3$	4.33	2.71	5.63	2.77
3	0.8	0.8	$\xi-3$	-3	3.96	3.88	4.52	4.19
4	0.8	0.8	$\xi-3$	$\xi-3$	2.89	2.16	2.70	2.11
5	0.05	0.8	-3	-3	5.88	5.78	Inf	Inf
6	0.05	0.8	-3	$\xi-3$	4.91	2.13	5.48	2.07
7	0.05	0.8	$\xi-3$	-3	5.36	4.85	5.88	5.57
8	0.05	0.8	$\xi-3$	$\xi-3$	2.75	1.94	2.47	1.85
9	0.8	0	-3	-3	4.89	1.79	Inf	Inf
10	0.8	0	-3	$\xi-3$	3.73	1.72	4.97	2.65
11	0.8	0	$\xi-3$	-3	3.26	1.75	5.27	3.76
12	0.8	0	$\xi-3$	$\xi-3$	2.27	1.44	2.38	1.78
13	0.05	0	-3	-3	6.48	5.63	Inf	Inf
14	0.05	0	-3	$\xi-3$	4.92	2.70	6.18	2.65
15	0.05	0	$\xi-3$	-3	4.97	4.80	6.00	6.18
16	0.05	0	$\xi-3$	$\xi-3$	2.77	1.88	2.56	1.79
17	0.8	1	-3	-3	Inf	6.00	Inf	Inf
18	0.8	1	-3	$\xi-3$	6.00	2.70	Inf	2.66
19	0.8	1	$\xi-3$	-3	5.88	4.84	5.78	6.18
20	0.8	1	$\xi-3$	$\xi-3$	2.80	1.87	2.57	1.79
21	0.05	1	-3	-3	Inf	6.00	Inf	Inf
22	0.05	1	-3	$\xi-3$	6.48	2.71	Inf	2.66
23	0.05	1	$\xi-3$	-3	Inf	4.96	Inf	6.18
24	0.05	1	$\xi-3$	$\xi-3$	2.92	1.89	2.62	1.79

3.1 Genotyping

We calculate the genotype-call error rates of the proposed method in all simulation experiments, and compare them to that of Martin *et al.* (2010). Note that at most positions, where the non-reference allele frequencies are close to 0, 1/2 or 1 with sufficient coverage, no matter which methods are used, their genotype-calls are similar. The differences only happen at the less ambiguous positions. Therefore, even though the absolute differences in the genotype-call error rates may not be so big, it indicates the significant improvement in genotyping, especially at the positions where one often makes mistakes. In order to clearly show the difference in genotype-call error rates of both methods, we present their $-\log_{10}$ (genotype-call error rate) in Table 1, where a larger entry indicates a smaller genotype-call error rate. It is shown that, both of our method and Martin *et al.* (2010) can focus on the genotype estimation for each individual sample, so that decreasing Q from 0.8 to 0.05 does not increase their genotype-call error rates, demonstrating their efficacy for rare variation detection. The experiment with $Q=0.05$ even presents less genotype-call errors than that with $Q=0.8$, since smaller Q indicates less samples mutate into RV, that do not make contributions to the parameter estimation.

As the coverage determination parameter N increasing from 50 to 100, our genotype-call error rate decreases as expected. In particular, when the sample effect exists, the advantage of our method becomes more significant, as shown in experiments 2, 6, 10, 14, 18, 22. Martin *et al.* (2010) call genotypes independently for each genomic locus, and at each locus, a homogenous sequencing error rate is assumed across all of the samples. In contrast, our model admits the heterogeneity among samples, and integrates the data across both of samples and genome loci to estimate the sequencing error rate, resulting in less genotype-call errors.

Table 2. The detection powers in simulation experiments with $\alpha=0.01$, where P, P_E and P_M respectively indicate the power of our proposed method, ECM-M and Muralidharan *et al.* (2012a)

$N=50$

Exp	P	P_E	P_M	Exp	P	P_E	P_M	Exp	P	P_E	P_M
1	1	1	1	9	1	1	–	17	1	1	1
2	1	1	1	10	.99	.99	.49	18	1	1	1
3	1	1	.99	11	.98	.98	.48	19	1	1	–
4	.99	.98	.82	12	.88	.87	0	20	.99	.99	.89
5	1	1	1	13	1	1	1	21	1	1	1
6	.99	.99	.99	14	.99	.98	.98	22	1	1	1
7	1	.99	–	15	.99	.98	.98	23	1	1	1
8	.98	.98	.71	16	.93	.93	0	24	.99	.99	.88

$N=100$

Exp	P	P_E	P_M	Exp	P	P_E	P_M	Exp	P	P_E	P_M
1	1	1	1	9	1	1	–	17	1	1	–
2	1	1	.99	10	1	1	.36	18	1	1	1
3	1	1	.98	11	1	1	0	19	1	1	1
4	1	1	.96	12	.99	.98	.85	20	1	1	.97
5	1	1	1	13	1	1	1	21	1	1	–
6	1	1	.97	14	1	1	.98	22	1	1	1
7	1	1	.99	15	1	1	.95	23	1	1	1
8	.99	.99	.92	16	.98	.98	.91	24	1	1	.97

– indicates failure to report the result.

3.2 SNP detection

Given different cutoff values α, we identify SNPs according to $\widehat{z}_{ij,-1}$ in each simulation experiment. For comparison, we implement the method of Muralidharan *et al.* (2012a) by a naive mixture model for pre-genotyping and the median method for the parameter estimation. After getting the P-values from the Binomial-Beta distribution, they are randomized and analyzed by the FDR method using R package locfdr (Efron 2004,2007a,b) for SNP detection. Moreover, in order to illustrate the advantage of ECM algorithm, after getting the parameter estimates from our method, we also calculate the P-values, then randomize them and use locfdr to call SNPs. We denote this procedure by ECM-M. The detection powers and FDRs of our proposed method, Muralidharan *et al.* (2012a) and ECM-M are presented in Tables 2 and 3, respectively. We only show the results with $\alpha=0.01$ for illustration. For more results, please refer to the Supplementary Materials.

As shown in Tables 2 and 3, our method performs very similarly to ECM-M across all of the simulation experiments, demonstrating the feasibility of posterior probability $\widehat{z}_{ij-1}$ for multiple testing adjustment. Both of these two methods achieve high detection powers with FDRs controlled at the nominal level, even with the low coverage when $N=50$. Except in experiments 12 and 16, their powers are all greater than 0.95. As N increasing from 50 to 100, the detection powers in experiments 12 and 16 rise up to 0.98, too. The FDRs are well controlled across all of the experiments, except in experiments 8, 16 and 24, where both of the positional effect and sample effect exist, and large FDRs are expected. The existence of both the positional effect and sample effect greatly increases the probability to get a too large sequencing error rate, causing RR is very likely to be wrongly called as RV. Although it does not influence the detection power, the number of false positives, therefore FDR, increases.

In contrast, the performance of Muralidharan *et al.* (2012a) is not so satisfied. In experiments 7, 9 and 19 when $N=50$ and experiments 9, 17 and 21 when $N=100$, it fails in fitting the empirical

Table 3. FDRs in simulation experiments with $\alpha = 0.01$, where r, r_E and r_M respectively indicate the FDR of our proposed method, ECM-M and Muralidharan *et al.* (2012a)

$N = 50$											
Exp	r	r_E	r_M	Exp	r	r_E	r_M	Exp	r	r_E	r_M
1	0	0	0	9	0	0	–	17	0	0	0
2	0	0	.05	10	0	0	.40	18	0	0	0
3	0	0	.20	11	0	0	.49	19	0	0	–
4	0	0	.14	12	.03	.02	1	20	.01	.01	.14
5	0	0	.03	13	0	0	.02	21	0	0	.04
6	0	0	.76	14	0	0	.40	22	0	0	.01
7	0	0	–	15	0	0	.91	23	0	0	.83
8	.17	.18	.91	16	.18	.18	1	24	.14	.16	.87
$N = 100$											
Exp	r	r_E	r_M	Exp	r	r_E	r_M	Exp	r	r_E	r_M
1	0	0	0	9	0	0	–	17	0	0	–
2	0	0	.07	10	0	0	.47	18	0	0	0
3	0	0	.48	11	0	0	1	19	0	0	.47
4	.03	.03	0	12	.05	.05	.33	20	.03	.04	0
5	0	0	.05	13	0	0	.01	21	0	0	–
6	0	0	0	14	0	0	.78	22	0	0	0
7	0	0	.94	15	0	0	.97	23	0	0	.94
8	.40	.51	0	16	.36	.43	.91	24	.34	.44	0

- indicates failure to report the result.

distribution and reports errors while running locfdr, whereas ECM-M does not. It seems that the parameter estimation is crucial to the results of SNP detection and ECM algorithm is necessary to be implemented to estimate the unknown parameters. In the successful experiments, the detection power and FDR of Muralidharan *et al.* (2012a) are not comparable to that of ECM-M and our method, especially in some experiments as shown in Tables 2 and 3, the detection power of Muralidharan *et al.* (2012a) goes remarkably low and its FDR runs out of control.

4 Real data analysis

4.1 Ovarian cancer transcriptomes data

To evaluate the SNP detection method developed for single-sample analysis, Goya *et al.* (2010) provide the sequencing data of 16 ovarian cancer transcriptomes, and obtain their genotypes with high-confidence (>0.99) through the Affymetrix SNP 6.0 high-density genotyping arrays. Although there are 9000 genomic loci from each sample and a total of 144 271, we only retrieve the sites overlapped by 16 samples to illustrate our method for multi-sample analysis, consisting of 1715 loci.

According to the genotyping arrays, Goya *et al.* (2010) denote by 1 or 0 to indicate whether or not it is a SNP at each genomic locus of each sample. For comparison, we convert the genotype-calls of our method and Martin *et al.* (2010) from RR to 0 and both of RV and VV to 1, resulting in 3.13% errors using our method among 16 × 1715 positions, which is less than the error rate 3.95% of Martin *et al.* (2010).

Given different α, the detection powers and FDRs of our method, ECM-M and Muralidharan *et al.* (2012a) are listed in Table 4. As similarly shown in the simulations, our results are very close to that of ECM-M, but the detection power of our method is slightly higher. Both of these two methods attain the satisfied detection powers with FDRs being well controlled. Whereas, Muralidharan *et al.* (2012a) present significantly lower detection powers and larger FDRs.

Table 4. The detection power and FDR of our method, ECM-M and Muralidharan *et al.* (2012a) in ovarian cancer transcriptomes data analysis

	Detection power		
	Proposed	ECM-M	Muralidharan *et al.* (2012a)
$\alpha = 0.001$	.84	.82	.50
$\alpha = 0.01$	.86	.84	.59
$\alpha = 0.05$	.88	.86	.72
$\alpha = 0.1$	.89	.86	.76
	FDR		
	Proposed	ECM-M	Muralidharan *et al.* (2012a)
$\alpha = 0.001$	.03	.03	.31
$\alpha = 0.01$	.03	.03	.30
$\alpha = 0.05$	.03	.03	.29
$\alpha = 0.1$	.03	.03	.30

Table 5. The detection power and FDR of our proposed method in the human genome data analysis

α	0.001	0.01	0.05	0.1
Detection power	.87	.87	.88	.88
FDR	.01	.01	.01	.01

4.2 Human genome data

In addition to the ovarian cancer transcriptomes data, we also apply the proposed method to human genome data sequenced from 13 samples (Cleary *et al.*, 2014). These 13 samples include NA12878, NA12877 and their 11 offsprings NA12879, NA12880, NA12881, NA12882, NA12883, NA12884, NA12885, NA12886, NA12887, NA12888 and NA12893. A VCF file containing the SNPs identified by Cleary *et al.* (2014) can be downloaded from the FTP server (ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/RTG_small_variants_01132014/phasing_annotated.vcf.gz). Besides the coverage and the non-reference allele counts, the ground true genotypes which can be treated as the benchmark are included. For illustration, we only retrieve the data on the first chromosome, where there are 437 414 sites overlapped by 13 samples. According to the benchmark genotypes, the majority of 13 × 437 414 positions are true SNPs. The percentages of RR, RV and VV are 28.46, 46.68 and 24.86 %, respectively.

The genotype-call error rates from our proposed method and Martin *et al.* (2010) are 8.92 and 16.77%, respectively. We make much less genotype-call errors. It seems that the heterogeneity exists among different samples. By taking the sample effect into consideration, our method significantly improves the genotype-calls.

The SNPs are called using the proposed method, ECM-M and Muralidharan *et al.* (2012a), respectively. However, both of the last two methods report errors in running locfdr. The FDR method used in locfdr is developed with the assumption that there are only a few SNPs, so fitting the empirical distribution can serve as the null distribution. However, for this dataset, where the positions are SNPs called by Cleary *et al.* (2014) and true SNPs dominate, the fitting is more likely to fail. Even if it has not failed, it is risky to use the fitted distribution as null. In Table 5, we only present the results of our method. Given α ranging from 0.001 to 0.1, we consistently get over 87% detection power with FDR controlled under 0.01.

5 Discussion

In this paper, we propose a novel statistical method for both of genotyping and SNP detection using multi-sample NGS data. Instead of pooling the multi-sample data as single-sample or pooled sequencing data, we build the statistical model to integrate information across different samples and genomic sites, to make the genotype-call and identify SNP at each locus for each sample. This offers our method high detection power for both common and rare SNPs, as illustrated in the simulations and real data analysis.

In the proposed model, the sample effect is a constant across all of the genomic sites. Sometimes, it may be not feasible, since as mentioned previously, the sequencing error rate may be affected by the local DNA contents, repetitive regions, etc. We suggest to cut the whole genome into small-pieced regions and analyze them in the piece-wise manner. This not only can solve the heterogeneity issue in the sample effect, but also helps to implement the computation in a parallel way, which will greatly reduce the computation duration for huge data analysis.

Funding

This work was supported by the National Natural Science Foundation of China [NSFC 11301554 to N.Y. and G.H., NSFC 11371235 to S.W., NSFC 11271383 to X.W.]; National Natural Science Foundation of China for Excellent Young Scholar [11322108 to X.W.]; National Natural Science Foundation of Guangdong, China [2015A030313143 to N.Y.]; Research Fund for the Doctoral Program of Higher Education of China [20120171120006 to N.Y.], and Major State Basic Research Development Program [2012CB517900 to X.W.].

Conflict of Interest: none declared.

References

Cleary,J.G. *et al*. (2014) Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol*., **21**, 405–419.

DePristo,M.A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*., **43**, 491–498.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Am. Stat. Assoc*., **99**, 96–104.

Efron,B. (2007a) Correlation and large-scale simultaneous significance testing. *J. Am. Stat. Assoc*., **102**, 93–103.

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

Efron,B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, 1st edn. Cambridge University Press, New York.

Goya,R. *et al*. (2010) Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.

Le,S.Q. and Durbin,R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res*., **21**, 952–960.

Li,B. and Leal,S.M. (2009) Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PloS Genet*., **5**, e1000481.

Li,H. *et al*. 1000 Genome Project Data Processing Subgroup. (2009a) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al*. (2009b) Snp detection for massively parallel whole-genome resequencing. *Genome Res*., **19**, 1124–1132.

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. *et al*. (2012a) Detecting mutations in mixed sample sequencing data using empirical Bayes. *Ann. Appl. Stat*., **6**, 1047–1067.

Muralidharan,O. *et al*. (2012b) A cross-sample statistical model for snp detection in short-read sequencing data. *Nucleic Acids Res*., **40**, e5.

Murillo,G. *et al*. (2016) Multigems: Detection of snvs from multiple samples using model selection on high-throughput sequencing data. *Bioinformatics*, **32**, 1486–1492.

Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.

Shen,Y. *et al*., (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res*., **20**, 273–280.

Snyder,M. *et al*. (2010) Personal genome sequencing: current approaches and challenges. *Genes Dev*., **24**, 423–431.

You,N. *et al*. (2012) Snp calling using genotype model selection on high-throughput sequencing data. *Bioinformatics*, **28**, 643–650.

Zhao,Z. *et al*. (2013) An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data. *Ann. Appl. Stat*., **7**, 2229–2248.

Zhou,B. (2012) An empirical Bayes mixture model for snp detection in pooled sequencing data. *Bioinformatics*, **28**, 2569–2575.