Statistics and Its Interface Volume 8 (2015) 457–462

Single-sample SNP detection by empirical Bayes method using next generation sequencing data

Weijie Ding, Qiang Kou, Xueqin Wang*, Qiuya Xu, and Na You*

The rapid development of next generation sequencing technology is changing the way of biological research in many aspects, which has become the most popular platform for the genomic structural variation detection. In this paper, we focus on the single-sample next generation sequencing data analysis, and propose a hierarchical structure to model the dispersion of minor allele frequency in the genome scale. The empirical Bayes method is employed to estimate the hyper-parameters, and the minor allele is identified as a sequencing error or heterozygous allele according to the posterior probabilities. We suggest to leave the ambiguous positions with moderate posterior probabilities ungenotyped for better genotype-call error control. The performances of our proposed method are investigated by simulations and a real dataset.

AMS 2000 subject classifications: Primary 62P10.
Keywords and phrases: Next generation sequencing, Single-sample, Genotyping, SNP detection, Empirical Bayes method.

1. INTRODUCTION

The development of next-generation sequencing (NGS) technology provides us an ultimately efficient way to learn about the genomics at nucleotide level, which nowadays has been widely used for DNA structural variation detection ([6, 9, 11]). In order to identify the difference between the sample genome and a reference genome, the sample is split into thousands of millions of small pieces for sequencing, and then the base-called pieces, called as reads, are aligned to the reference genome to recover their locations in the original sample genome. The structural variation detection and other well known downstream analysis, such as the genome-wise association study and gene regulation network reconstruction, are then inferred based on the alignment data.

To this date, there has been many statistical methods proposed for the structural variation identification, in particular for the single nucleotide polymorphism (SNP) detection such as SOAPsnp ([4]), SAMtools ([3, 2]) and GATK ([1]), for both of single-sample and multi-sample analysis. The Bayes model is widely used among these methods, which estimates the sequencing error rate based on the base-calling quality and mapping quality scores and then calls the genotype according to the posterior probabilities with some pre-specified prior distribution. For a particular base on the aligned read, GATK estimates the sequencing error rate by $10^{-0.1Q}$, where $Q = \min\{B, M\}$, and B and M respectively indicate the base-calling quality and mapping quality scores of that base. The problem with this implementation is, besides the arbitrary integration formula Q, the sequencing error rate determined by the quality scores may be not precise. It's known that not only the chemical synthesis and removal during sequencing, but also the NGS data analysis pipeline are complex processes. Many factors during these processes may contribute to the sequencing errors but not be well considered and reflected by the quality scores, e.g., base-calling estimation bias, repetitive regions and sample preparation errors ([13]). Besides, GATK sets the prior distribution to be fixed for all applications, which also limits the flexibility and accuracy of GATK genotype-calls.

In order to better estimate the sequencing error rate, many articles, e.g., [5, 12, 8, 7, 14], proposed to use the binomial model to account for the variations in the observed alleles. However, almost all of such methods were proposed and programmed for the multi-sample data analysis, where the alleles on different positions were assumed to be independent, and the empirical Bayes method was employed to borrow information across different samples while estimating the sequencing error rate for each position. There are only a few work ([8]) modelling the sequencing error rates across different positions on the same genome. Although theoretically it can be adapted in the single-sample case, their program exactly sets the minimum number of samples as 2.

In many situations, the single-sample genotyping and SNP calling are of great interest, for instance in cancer genome analysis or rare variants analysis. In this paper, we focus on the single-sample genotyping and SNP calling, and propose a hierarchical model to describe the dispersion of minor allele frequency in the genome scale. An empirical Bayes method is employed to combine information across different positions while estimating the hyper-parameters. The genotype at each position is assigned according to the posterior probabilities and SNPs are identified consequently. For the ambiguous positions with moder-

*Corresponding authors of the project.

ate posterior probabilities, we suggest to leave them ungenotyped rather than make any decisions. The proposed method is implemented as R functions and wrapped in an R package `ebSNP`, which is publicly available at http://cran.r-project.org/web/packages/ebSNP/index.html.

The paper is structured as following. In section 2, an empirical Bayes method for genotyping and SNP calling is introduced. Its performances are investigated via simulations and a real dataset in section 3. A short discussion is given in section 4.

2. METHODS

For a diploid genome, at each position, the genotype is either homozygous or heterozygous. If it is homozygous, then only one of four nucleotides could be observed among the alleles if there is no sequencing errors, and two if it is heterozygous. But due to the sequencing errors, other nucleotides also may appear at each position. Let $n_{i\Delta}$ denote the number of alleles with nucleotide Δ covering position i, $i = 1, 2, \ldots, n$, $\Delta \in \{A, C, G, T\}$, and $G_i = 0$ or 1 indicate the genotype at position i is homozygous or heterozygous respectively.

At each position, we only consider the first two most frequent alleles, e.g., the major allele and the minor allele, and discard the rest. The reason for this is, if $G_i = 0$, only the major allele is related to the genotype and the others come out as sequencing errors, while if $G_i = 1$, then both of the major and minor alleles are genotype-related and the rest are sequencing errors. Note that the sequencing error rate is supposed to be far smaller than 1/2, while the heterozygous genotype is expected to show two alleles with equal probabilities, so for the purpose of genotyping and SNP calling, we only need to focus on the first two most frequent alleles and identify whether the minor allele comes out due to the sequencing error or heterozygous genotype.

We sort n_{iA}, n_{iC}, n_{iG}, n_{iT} in the descent order and denote the first two by n_{i1} and n_{i2} respectively, where $n_{i1} \geq n_{i2}$. Let Δ_{ij} indicate the nucleotide corresponding to the allele frequency n_{ij}, $j = 1$ and 2. The coverage at position i is adjusted to be $N_i = n_{i1} + n_{i2}$. Note that $n_{i1} \geq N_i/2$. Given $G_i = 0$, n_{i1} is genotype-related, therefore we assume

$$n_{i1} \sim \text{Binom}(N_i, 1 - p_i),$$

where p_i is the probability of observing a sequencing error at position i. While given $G_i = 1$, n_{i1} is the order statistic $\max(X, N_i - X)$, where $X \sim \text{Binom}(N_i, 1/2)$.

Furthermore, we assume p_i follows a Beta(α, β) distribution across the genome positions, where α and β are unknown parameters and will be estimated by the empirical Bayes method. Denoted by $\pi_0 = P(G_i = 0)$, the probability of a particular position bearing a homozygous genotype, the model can be summarized as,

$$\begin{aligned}
&G_i \sim \text{Binom}(1, 1 - \pi_0),\\
&P(n_{i1}|N_i, G_i = 0) = \binom{N_i}{n_{i1}} \frac{B(N_i - n_{i1} + \alpha, n_{i1} + \beta)}{B(\alpha, \beta)},\\
&P(n_{i1}|N_i, G_i = 1) = 2\binom{N_i}{n_{i1}} \Big(\frac{1}{2}\Big)^{N_i},
\end{aligned}$$

where $B(a, b)$ is the Beta function with parameter a and b. Taking G_i as missing values, an EM algorithm is implemented to get the estimates for unknown parameters. The complete log-likelihood

$$\begin{aligned}
l = \sum_{i=1}^{n} \Big\{ & I(G_i = 0)\Big(\log P(n_{i1}|N_i, G_i = 0) + \log \pi_0 \Big)\\
& + I(G_i = 1)\Big(\log P(n_{i1}|N_i, G_i = 1) + \log(1 - \pi_0) \Big) \Big\},
\end{aligned}$$

where $I(\cdot)$ is the indicator function. Given the parameter estimates from the kth iteration $\alpha^{(k)}$, $\beta^{(k)}$ and $\pi_0^{(k)}$, the $(k+1)$th E-step calculates

$$\begin{aligned}
g_{i0}^{(k+1)} &= E\Big(I(G_i = 0)|N_i, n_{i1}, \alpha^{(k)}, \beta^{(k)}, \pi_0^{(k)} \Big)\\
&= \frac{P(n_{i1}|N_i, G_i = 0, \alpha^{(k)}, \beta^{(k)})\pi_0^{(k)}}{P(n_{i1}|N_i, G_i = 0, \alpha^{(k)}, \beta^{(k)})\pi_0^{(k)} + 2\binom{N_i}{n_{i1}}(\frac{1}{2})^{N_i}(1 - \pi_0^{(k)})},
\end{aligned}$$

and the $(k+1)$th M-step updates parameters as

$$\begin{aligned}
&\pi_0^{(k+1)} = \frac{\sum_{i=1}^{n} g_{i0}^{(k+1)}}{n},\\
&(\alpha^{(k+1)}, \beta^{(k+1)})\\
&\quad = \underset{(\alpha,\beta)}{\text{argmax}} \sum_{i=1}^{n} g_{i0}^{(k+1)} \log \frac{B(N_i - n_{i1} + \alpha, n_{i1} + \beta)}{B(\alpha, \beta)}.
\end{aligned}$$

Given initial values $\alpha^{(0)}$, $\beta^{(0)}$ and $\pi_0^{(0)}$, the above E-step and M-step iterate until convergence, and then we get the estimates for parameters, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_0$.

Whether the genotype at position i is homozygous or heterozygous is determined by

$$g_i = 1 - E\Big(I(G_i = 0)|N_i, n_{i1}, \hat{\alpha}, \hat{\beta}, \hat{\pi}_0 \Big).$$

If it's homozygous, then the genotype at position i is $\Delta_{i1}\Delta_{i1}$, and if it's heterozygous, the genotype is $\Delta_{i1}\Delta_{i2}$. In general, as long as $g_i < 0.5$, G_i could be recognized as 0, and 1 otherwise. But as known, if g_i is nearby 0.5, for example, 0.45, although it still could be taken as a homozygote, there also has non-negligible probability to be called as a heterozygote. Inferring either genotype is dangerous. Therefore, in such situations, we suggest the researcher set the reasonable thresholds to guard the accuracy, e.g., when $g_i < T_1$, G_i is set to be 0 and $g_i \geq T_2$, is set to be 1, while $T_1 \leq g_i < T_2$, is left ungenotyped without any conclusion being made, marked as NA for instance.

Table 1. Parameter settings and estimates in 6 simulated experiments. Standard deviation in parenthesis

Exp.	N	π_0	α	β	$\hat{\pi}_0$	$\hat{\alpha}$	$\hat{\beta}$	r_1	r_2
1	20	0.80	1	10	0.79(0.0049)	1.14(0.04)	12.25(0.5)	1.27	1.74
2	20	0.95	2	100	0.95(0.0021)	2.04(0.27)	102.32(13.79)	2.99	3.73
3	40	0.80	1	10	0.8(0.0045)	1.04(0.02)	10.74(0.31)	1.56	1.98
4	40	0.95	2	100	0.95(0.0021)	2.03(0.15)	101.83(7.59)	4.41	4.77
5	80	0.80	1	10	0.8(0.0044)	1.02(0.02)	10.37(0.23)	1.85	2.18
6	80	0.95	100	4000	0.95(0.002)	93.69(31.58)	3748.86(1269.14)	Inf	Inf

3. ANALYSIS RESULTS

3.1 Simulations

A series of simulations are done to investigate the performances of our proposed method. By setting $n = 10{,}000$, we vary N_i, π_0, α and β in different simulation experiments. For simplicity, within each experiment, we set the coverage at each position N_i as a constant N and make 100 replicates to get the means and standard deviations of parameter estimates. In order to evaluate the genotype-call accuracy, we define GE_1 as the proportion of positions that are homozygous but identified as heterozygotes, GE_2 as the proportion of positions that are heterozygous but recognized as homozygotes, and GE as the proportion of total mis-genotyped positions.

Firstly, the parameter estimation and genotype-call accuracy of our proposed method are investigated, where there are 6 experiments generated as listed in Table 1. Since we only identify the homozygosity or heterozygosity of the genotype in our proposed method, we present $r = -\log_{10}(GE_1 + GE_2)$ with different thresholds T_1 and T_2 in Table 1 to evaluate the actual genotype-call error, where r_1 was calculated when $T_1 = T_2 = 0.5$ and r_2 was obtained when $T_1 = 1 - T_2 = 0.1$. As seen in Table 1, π_0 is well estimated in each of those 6 experiments, and as the coverage goes up, the estimation biases of $\hat{\alpha}$ and $\hat{\beta}$ decrease and their standard deviations become smaller. Meanwhile, the loose threshold can help avoid mis-genotyping at the ambiguous positions and improve the genotyping accuracy, demonstrated by larger r_2 than r_1 as shown in Table 1. We include experiment 6 with extremely large α and β values in this simulation, to reflect the quite small sequencing error rate which may occur in real applications, but possibly have some arithmetic underflow problem in running our EM algorithm. It happens due to the large value underflow in the Beta function, but doesn't have the significant effect on the computation of g_i, since as α or β goes to be larger, the Beta(α, β) density becomes more skewed, and after some extent, the skewness will not change dramatically as α or β changing. The seemingly large estimation bias of $\hat{\alpha}$ or $\hat{\beta}$ is still acceptable. As seen from experiment 6, although α and β were not estimated precisely, they still performed perfectly for genotyping.

The second simulation is conducted to compare the performances of our proposed method to that of GATK, where 8 experiments with different parameter settings are generated, as shown in Figure 1. In the GATK default setting, the nucleotides with mapping quality less than 10 or base-calling quality less than 17 are excluded for analysis, resulting in the sequencing error rate ranging from 10^{-1} to 10^{-4}. In order to apply GATK, we assume Q is a constant across all of the positions and equals 10, 20, 30 or 40 respectively in each experiment. Note that the original GATK assigns the prior probability for the reference homozygote to be $1-3\epsilon/2$ and each of the other nine possible genotypes to be $\epsilon/6$, where $\epsilon = 0.001$. Since we only consider the two most frequent alleles here, we modify the prior as $P(\text{reference homozygote}) = 1-3\epsilon/2$, $P(\text{non-reference homozygote}) = \epsilon/2$ and $P(\text{heterozygote}) = \epsilon$. In Figure 1, besides the GATK and our proposed method, the Bayes classifier ([5]) with the true sequencing error rates and prior distribution that is the same as GATK but $\epsilon = 1 - \pi_0$ are also included for comparison.

Our proposed method outperforms GATK with less genotype-call errors being caused. As shown in Figure 1, with $T_1 = T_2 = 0.5$, our method achieves the comparable or even much higher $-\log_{10} GE$ than GATK with four different Q values. While the thresholds are set to be $T_1 = 1 - T_2 = 0.1$, $-\log_{10} GE$ increases and is consistently higher than that of GATK with different Q values across 8 experiments. The distance between the sequencing error rate which is estimated by Q and the truth determines the genotype-call precision of GATK. As Q becomes larger, the sequencing error rate that GATK can tolerate becomes smaller, which makes it more likely to call the homozygote to be heterozygous. It's shown in Figure 1, $-\log_{10} GE_1$ decreases as Q increasing from 10 to 40. On the other hand, if GATK assigns a smaller Q comparing to the true sequencing error rate, it may call the heterozygote to be homozygous, as seen that $-\log_{10} GE_2$ increases as Q increasing. Unfortunately, as mentioned previously, the Q formula and quality scores may not be able to well reflect the true sequencing error rate, resulting in genotype-call errors using GATK. Different from GATK, our method estimates the sequencing error rates and π_0 from the alignment data, showing less dependence on Q formula and quality scores. Across 8 experiments in Figure 1, $-\log_{10} GE$ from our method is much closer to that of Bayes classifier with true parameters than $-\log_{10} GE$ from GATK.

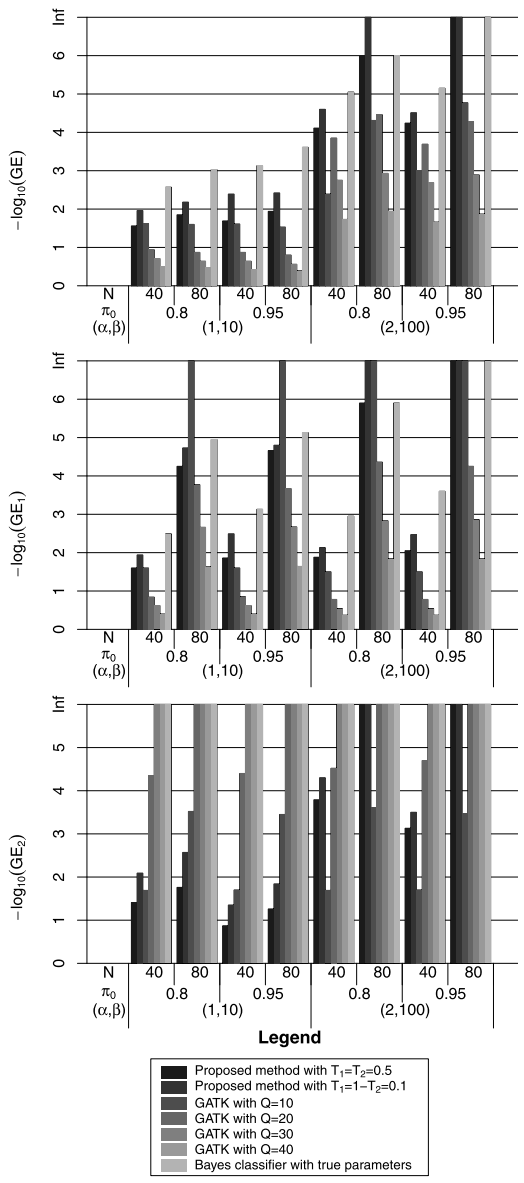

Figure 1. Observed genotype-call error rates in the second simulation, which are presented in each experiment in the order of our proposed method with threshold $T_1 = T_2 = 0.5$, proposed method with threshold $T_1 = 1 - T_2 = 0.1$, GATK with $Q = 10$, 20, 30 and 40, and Bayes classifier with true parameters, as shown in the legend.

3.2 A real dataset

The genome of subject NA12878 ([10]) is well studied in the 1000 Genome Project, whose alignment data could be publicly downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/pilot2_high_cov_GRCh37_bams/data/NA12878/alignment/NA12878.chrom20.ILLUMINA.bwa.CEU.high_coverage.20100311.bam. For illustration, we only retrieved the data on the first 3,000,000 positions of chromosome 20 for genotyping and SNP calling.

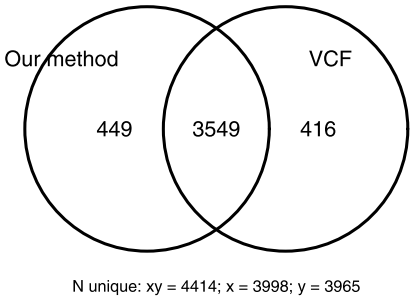

Figure 2. Overlap of SNPs identified by our method and recorded in the VCF file.

The frequencies of four alleles A, C, G and T at each position were generated, and only the first two most frequent alleles and their frequencies were saved for analysis. The alleles frequency data were firstly screened to filter out the positions who only carry one kind of allele, because they don't contribute to the sequencing error rate estimation and it's no doubt to genotype them as $\Delta_{i1}\Delta_{i1}$. Then, the data on the rest of the 946,042 positions were analyzed using our proposed method, resulting in $\hat{\pi}_0 = 0.9963$, $\hat{\alpha} = 122.32$ and $\hat{\beta} = 3798.49$.

Given $T_1 = T_2 = 0.5$, among 946,042 positions, 942,087 (99.58%) were genotyped to be homozygous and 3,955 (0.42%) were heterozygous. While given $T_1 = 0.01$ and $T_2 = 0.99$, there were 3,277 positions left ungenotyped. Among them, 1,342 (40.95%) have coverage lower than 10 and the minor allele frequencies of the rest are between 17.39% and 18.18%, which is seemingly larger than the usual sequencing error rate and also far lower than 0.5 of heterozygous genotype, so we prefer genotyping them as NA, instead of either homozygote or heterozygote. With $T_1 = 0.01$ and $T_2 = 0.99$, 3,388 positions which were genotyped as heterozygote, and 610 positions which were genotyped as homozygote but with the major alleles different from the reference alleles, were identified as SNPs by our procedure.

The SNPs of subject NA12878 found by 1000 genome project using GATK were summarized in a VCF file, which could be downloaded from ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/NIST_IntegratedCalls_12datasets_130517_HetHomVarPASS_VQSRv2.15.vcf.gz. It is shown that, within the region of 1 to 3,000,000 on chromosome 20, there are 5,102 SNPs, among which 3,965 belong to the 946,042 positions which we genotyped. Their overlap with SNPs identified by our methods are shown in Figure 2. Among 416 GATK exclusive SNPs, 352 (84.62%) are recorded as heterozygotes in the VCF file, while we called 206 (49.51%) as homozygous reference genotypes and 210 as NA. According to the simulation results that the overestimated sequencing error rates in GATK may lead the homozygotes to be called as heterozygotes, it seems that our proposed method controls the false heterozygous SNP calls, comparing to GATK. For

SNPs which are exclusively found by our method, although 422(93.98%) were genotyped as heterozygote, except 10 with coverage lower than 10, the minor allele frequencies on the rest 412 positions varied from 18.67% up to 50%, highly showing the heterozygous characteristics, especially 3/4 of 412 minor allele frequencies higher than 25.71% and 1/2 higher than 30%. In addition, our 27 exclusive homozygous SNPs have the non-reference allele frequencies from 80% up to 97.44%, with coverage from 5 to 77.

4. DISCUSSION

The development of NGS technology makes it possible to study the genome structure efficiently, but as mentioned earlier, the data analysis pipeline results in a variety of sequencing errors in the alignment data. How to measure the sequencing error rate to identify true mutations from sequencing errors is challenging, especially when there is only one single-sample data. The multiple-sample methods gave us a clue to use the empirical Bayes method to combine information across samples to estimate the sequencing error rate. Actually, borrowing information is more necessary in single-sample case, since the data resource is more limited. In this paper, we proposed an empirical Bayes algorithm to borrow information across different positions to measure the sequencing error rate in the genome scale, and then used it for genotyping and SNP detection.

We chose the Beta distribution to model the sequencing error rate since its domain is between 0 and 1. Other distributions, the mixture models as [8], or the heuristic algorithms which could learn the distribution shape from data may be alternative choices. Besides the parameters in the Beta distribution, we also estimated the homozygous probability π_0 from data, instead of using some arbitrarily fixed prior probabilities as previous Bayes models, which is more flexible and adaptive in real applications.

ACKNOWLEDGMENTS

Na You's research is partially supported by NSFC (11301554), RFDP (20120171120006) and Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (34000-3211702). Xueqin Wang's research is partially supported by NCET (12-0559), NSFC (11001280), RFDP (20110171110037) and National Program on Key Basic Research Project (2012CB517900).

Received 1 October 2013

REFERENCES

[1] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

[2] Li, H. (2010). Mathematical notes on samtools algorithms. http://lh3lh3.users.sourceforge.net/download/samtools.pdf.

[3] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map (sam) format and samtools. *Bioinformatics*.

[4] Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19(6):1124–1132.

[5] Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S., and Morris, R. W. (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, 26(22):2803–2810.

[6] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46.

[7] Muralidharan, O., Natsoulis, G., Bell, J., Ji, H., and Zhang, N. (2012). Detecting mutations in mixed sample sequencing data using empirical bayes. *Annals of Applied Statitics*, 6(3):1047–1067. MR3012520

[8] Muralidharan, O., Natsoulis, G., Bell, J., Newburger, D., Xu, H., Kela, I., Ji, H., and Zhang, N. (2012). A cross-sample statistical model for snp detection in short-read sequencing data. *Nucleic Acids Research*, 40(1):e5.

[9] Snyder, M., Du, J., and Gerstein, M. (2010). Personal genome sequencing: current approaches and challenges. *Genes & Development*, 24(5):423–431.

[10] The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

[11] The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

[12] Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*, 39(19):e132. MR2865755

[13] You, N., Murillo, G., Su, X., Zeng, X., Xu, J., Ning, K., Zhang, S., Zhu, J., and Cui, X. (2012). SNP calling using genotype model selection on high-throughput sequencing data. *Bioinformatics*, 28(5):643–650. MR3130947

[14] Zhao, Z., Wang, W., and Wei, Z. (2013). An empirical bayes testing procedure for detecting variants in analysis of next generation sequencing data. *Annals of Applied Statitics*, Available online http://www.imstat.org/aoas/next_issue.html. MR3161720

Weijie Ding
E-mail address: jasonding92@gmail.com

Qiang Kou
E-mail address: kouqiang@mail2.sysu.edu.cn

Xueqin Wang
School of mathematics & Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China

South China Research Center of Statistics
Sun Yat-sen University
Guangzhou, 510275
China
E-mail address: wangxq88@mail.sysu.edu.cn

Qiuya Xu
E-mail address: louise.aiesec@gmail.com

Na You
School of mathematics & Computational Science
Sun Yat-Sen University
Guangzhou, 510275
China

South China Research Center of Statistics
Sun Yat-sen University
Guangzhou, 510275
China
E-mail address: youn@mail.sysu.edu.cn