HHS Public Access

Author manuscript

IEEE Int Workshop Genomic Signal Process Stat. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

IEEE Int Workshop Genomic Signal Process Stat. 2012 December ; 2012: 42–45. doi:10.1109/GENSIPS.2012.6507722.

A Bayesian Model for SNP Discovery Based on Next-Generation Sequencing Data

Yanxun Xu[†],
Department of Statistics, Rice University Houston, TX

Xiaofeng Zheng[†],
Department of Bioinformatics and Computational Biology, The University of Texas, MD Anderson Cancer Center Houston, TX

Yuan Yuan[†],
Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine Houston, TX

Marcos R Estecio,
Department of Biochemistry and Molecular Biology, The University of Texas, MD Anderson Cancer Center Houston, TX

Jean-Pierre Issa,
Fels Institute for Cancer Research and Molecular Biology, Temple University Philadelphia, PA

Yuan Ji, and
CCRI, NorthShore University HealthSystem Chicago, IL

Shoudan Liang
Department of Bioinformatics and Computational Biology, The University of Texas, MD Anderson Cancer Center Houston, TX

Yuan Ji: yji@northshore.org; Shoudan Liang: shoudan@mdanderson.org

Abstract

A single-nucleotide polymorphism (SNP) is a single base change in the DNA sequence and is the most common polymorphism. Since some SNPs have a major influence on disease susceptibility, detecting SNPs plays an important role in biomedical research. To take fully advantage of the next-generation sequencing (NGS) technology and detect SNP more effectively, we propose a Bayesian approach that computes a posterior probability of hidden nucleotide variations at each covered genomic position. The position with higher posterior probability of hidden nucleotide variation has a higher chance to be a SNP. We apply the proposed method to detect SNPs in two cell lines: the prostate cancer cell line PC3 and the embryonic stem cell line H1. A comparison between our results with dbSNP database shows a high ratio of overlap (>95%). The positions that are called only under our model but not in dbSNP may serve as candidates for new SNPs.

Correspondence to: Yuan Ji, `yji@northshore.org`; Shoudan Liang, `shoudan@mdanderson.org`.

[†]Equal contributors

The Supplementary can be found at: https://sites.google.com/site/yanxunresearch.

I. Introduction

A single-nucleotide polymorphism is a DNA sequence variation when one single nucleotide (A, T, C or G) in the genome is altered, such as an A, is replaced by one of the other three nucleotides C, G, or T. Whether a SNP has functional impact on the individual largely depends on the genomic location (coding region, intron, etc) and whether it leads to amino acid change after translation. Those functional SNPs have great potential to be biomarkers and therapeutic targets and therefore are especially interesting. So far, great endeavors have been made to SNP discovery and recently developed Next-generation Sequencing (NGS) technique has largely facilitated this process.

NGS is an emerging high-throughput technology that produces genome-wide data enabling unprecedented access to comprehensive genetic information. Detection of genome-wide variation is one of the most important applications of NGS. For example, Genome-Wide Association Studies (GWAS) examines the relationships between millions of SNPs and traits. The 1000 Genome Project utilizes NGS technology and aims to produce an extensive public catalog of human genetic variation [1].

Two different types of approaches have been taken to detect SNPs. The first type relies upon amplification of DNA using the polymerase chain reaction (PCR) to reduce the complexity of the genome and works on re-sequencing data from diploid samples. These algorithms examine chromatogram trace files and detect variants by extracting or comparing signals in the peaks of traces. The widely used software using these algorithms include PolyPhred [2], SNPdetector [3], and novoSNP [4]. The second type is based on detecting sequence differences among cloned DNA samples. Two representative software are MAQ [5] and Atlas-SNP2 [6]. MAQ proposes a Bayesian approach to call variants by considering correlation of sampling and error rates at one particular position. Atlas-SNP2 takes into account sequence context in training datasets to sift through large amounts of high-throughput re-sequencing data and pick out genetic variants from ubiquitous sequencing errors. However, most of above software were not designed to handle NGS data. An effective and efficient algorithm for SNP discovery in NGS data will provide valuable information for downstream analysis.

In this study, we propose a Bayesian approach to estimate the probability of mismatch due to SNPs by computing posterior probability of hidden nucleotide variations at each covered genomic position from NGS data. Note that, during read alignment the base-level sequence mismatches can result from either sequencing errors [7] or hidden nucleotide variations such as SNPs and our approach will especially take these into consideration. The main idea is to model the base-level sequencing error rates using observed mismatch profiles. We apply our method to data sets on PC3 and H1 cell lines.

The paper proceeds as follows. We introduce the proposed probability model along with MCMC techniques in Section II. Section III shows the effectiveness of our model by applying to prostate cancer cell line PC3 data and comparing with MAQ. In Section IV, we apply our model to stem cell line H1 data. We conclude with a discussion in Section V.

II. Methodology

A. Probability model

We will apply the proposed model independently to each covered genomic position. This strategy allows us to analyze different positions in parallel, achieving fast computational speed. For a given nucleotide position t on the reference genome, suppose there are K_t unique reads overlapping with it. Denote the set $\mathbf{U}_t=\{U_{kt}\}_{k=1}^{K_t}$ the labels of all the unique reads that overlap with the position t. Note that here "overlap" means partial match. For example, if t = 101 and the length of short read is 35, any overlapping unique read will have a starting position in [67,101]. For a unique read $U_{kt} \in \mathbf{U}_t$, let $\{e_{kt} = 1\}$ and $\{e_{kt} = 0\}$ respectively denote mismatch or perfect match between the unique read U_{kt} and position t on the reference genome. So the mismatch profile at nucleotide position t is $\mathbf{e}_t = \{e_{kt}, k = 1,\ldots, K_t\}$.

To estimate the mismatch probability q_{kt} between the base at position t on the reference genome and the corresponding base of a unique read U_{kt} ,we write q_{kt} as a function of α_{kt}, the sequencing error rate, and β_{kt}, the probability of hidden nucleotide variations. Following the law of addition, denote A the event {*there is a sequencing error*} and B the event that {*there is a hidden nucleotide variation*}, then $q = Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A)Pr(B)$. Specifically, using our notation, let

$$q_{kt}=\alpha_{kt}+\beta_{kt}(1-\alpha_{kt}). \quad (1)$$

In our subsequent data analysis, we fix the values of α_{kt} and estimate β_{kt}. Recall that for the k^{th} unique read that overlaps with genomic position t, α_{kt} is the sequencing error rate. During mapping, a mapping quality character was assigned to each base of the reads. Let Z_c denote the count of bases to which a quality character c was assigned, and Z_{cm} denote the count of aforementioned bases that have mismatches. If the corresponding base of a unique read U_{kt} was assigned quality character c, then the corresponding quality score is defined as $\alpha_{kt} = Z_{cm}/Z_c$. Using the quality scores from millions of unique reads in the data, we can reliably estimate α_{kt} based on the observed sequencing error rates, rather than imposing a prior distribution on the α_{kt}. We write the likelihood contribution from the unique reads at position t, which is given by

$$L(\beta_{kt})=\prod_{k=1}^{K_t} q_{kt}^{e_{kt}}\{1-q_{kt}\}^{1-e_{kt}}. \quad (2)$$

In (2), the only unknown variable is β_{kt}. Given a genomic position t, there are three possible hidden nucleotide variations between the reference genome and aligned unique reads. For example, if the base at position t on the reference genome is A, there are three possible base substitutions: A-T, A-C and A-G. We assume that the probability of the three possible hidden nucleotide variations types to be $\boldsymbol{\beta}_t = \{\beta_{it}, i = 1, 2, 3\}$ for notation simplicity, where $\beta_{1t}, \beta_{2t}, \beta_{3t} > 0$ and $\beta_{1t} + \beta_{2t} + \beta_{3t} < 1$. If read k exhibits a mismatch at position t, i.e. $\{e_{kt} = 1\}$, then $\beta_{kt} \in \boldsymbol{\beta}_t$.

Denoting *Dir*(*a*, *b*, *c*, *d*) a Dirichlet distribution with density $\propto \beta_{1t}^{a-1}\beta_{2t}^{b-1}\beta_{3t}^{c-1}(1-\beta_{1t}-\beta_{2t}-\beta_{3t})^{d-1}$ we assume conditionally conjugate priors for $\boldsymbol{\beta}_t = (\beta_{1t}, \beta_{2t}, \beta_{3t})$ for a given genomic position *t*:

$$\boldsymbol{\beta}_t \sim Dir(a, b, c, d).$$

In the subsequent analysis, we assume $a = b = c = 0.001$ and $d = 1$, since the SNP rate in human genome is about 0.1%.

B. Markov chain Monte Carlo simulations

We augment the parameter space [8] and employ a Gibbs sampler to simulate the unknown parameters $\boldsymbol{\beta}_t$. Let g_t be the base at position *t*, and g_{kt} the base at the corresponding position of k^{th} unique read that overlapped with position *t*. Here, g_t is known from the reference genome. For g_{kt}'s, they take values in {A, C, G, T}. If read *k* exhibits a mismatch at position *t*, i.e. $\{e_{kt} = 1\}$, then

$$g_{kt} \in \{A, C, G, T\} \backslash \{g_t\} = \{C_1, C_2, C_3\},$$

where C_1, C_2, C_3 are the nucleotide types different from that at the position *t* on the reference genome.

Let $S_{it} = \{k : g_{kt} = C_i\}$ and $N_{it} = \|S_{it}\|$ for $i = 1, 2, 3$, where $\|\cdot\|$ denotes the number of elements in one set.

The basic idea is to introduce a latent Bernoulli variable with a conditional distribution defined by

$$u_k|\beta_{kt} \sim Bernoulli\left\{\frac{(1-\alpha_{kt})\beta_{kt}}{\alpha_{kt}+(1-\alpha_{kt})\beta_{kt}}\right\}$$

for $k \in \{S_{1t}, S_{2t}, S_{3t}\}$. With the augmented Bernoulli distribution, we can easily show that $[\boldsymbol{\beta}_t|u_k's]$ follows a Dirichlet distribution

$$Dir\{\sum_{k\in S_{1t}} u_k+a, \sum_{k\in S_{2t}} u_k+b, \sum_{k\in S_{3t}} u_k+c, K_t - N_{1t} - N_{2t} - N_{3t}+d\}. \quad (3)$$

In our analysis, the number of iterations *S* was set to 1,000 with the first 200 iterations as burn-in. The Markov chain converged fast and mixed well.

III. Prostate cancer cell line PC3 data analysis

We apply our method to a prostate cancer cell line PC3 dataset, aiming to detect SNPs. Prostate cancer is the most common cause of death from cancer in men over age 75.

We use Bowtie [9] to map the sequencing data and calculate the coverage with BEDTools [10]. 61.8% of the genome is covered with average sequencing depth of 1.55. And 62.6% of the exon regions are covered with the average sequencing depth of 1.62.

Recall that $\boldsymbol{\beta}_t$ denotes the probability of hidden nucleotide variations at genomic position t on the reference genome, so the position with higher posterior probability of $\boldsymbol{\beta}_t$ will have a higher chance to be a SNP We will call a position a potential SNP if there exists $\beta > \beta_0$, where $\beta \in \boldsymbol{\beta}_t$ and β_0 is a cutoff value. To determine the optimal cutoff, we compare the identified SNP candidates with dbSNP [11] using ANNOVAR [1]. When $\beta_0 = 0.5$, 685,277 candidate SNPs are identified and 634,477 (92.6%) of them are found in dbSNP When $\beta_0 = 0.75$, 198,645 candidate SNPs are identified and 188,989 (95.1%) of them are found in dbSNP.

For comparison, we apply MAQ to the same data. MAQ targets 387,931 candidate SNPs, in which 233,283 (60%) are in dbSNP The proportion of identified SNPs falling into dbSNP can be used as a measure of method reliability. Using above criteria, our method outperforms MAQ since our method not only identifies more SNPs, but also has higher quality.

For each potential SNP, we compute the proportions of one particular nucleotide variation and compare them with posterior probability of this hidden nucleotide variation. For example, if the nucleotide at genomic position t is A, and there are 100 short reads mapped to this position with 10 T's, 15 C's and 75 A's. We can get at this position, the proportion of A-T substitution is 0.1; the proportion of A-C substitution is 0.15. Fig. 1 shows the scatter plot with smoothed densities color representation of posterior probability of hidden nucleotide variations versus the proportions of nucleotide changes for the potential 28,057,824 SNPs. Most of the points are on the 45-degree line, which shows the posterior probability of hidden nucleotide variation has very high correlation with proportion rate. In addition, there exist many points, whose observed proportion is 1 and the corresponding posterior probability ranges from 0.5 to 1. This is due to Bayesian shrinkage since we borrow strength from the sequencing errors.

IV. Stem cell line H1 data analysis

We apply our model to stem cell line H1 data that is almost 20 times larger than PC3 data and has relatively high and uniform sequencing depth. H1 cell line is one of the most extensively studied and characterized stem cell lines. Except for the bisulfite-seq and smRNA-seq, all the sequencing data of H1 cell line available at NCBI before November 2011 are used in our analysis. The complete list of data we used is in Supplementary. The H1 sequencing reads cover 91.3% of the genome with the average sequencing depth of 19.2, while 94.1% of the exon regions are covered with the average sequencing depth of 26.6.

For stem cell line H1 data, with the increase of β_0, the percentage of the identified candidate SNPs overlapping with dbSNP also increases as shown in Fig. 2. When $\beta_0 = 0.75$, 1,212,325 candidate SNPs are identified and 1,159,483 (95.6%) of them are found in dbSNP The functions of these SNPs are summarized in Supplementary. Among the 8,514 ex-onic SNPs,

3918 are nonsynonymous, 4,419 are synonymous, 18 are stopgain, 5 are stoploss and 154 are with unknown function.

From these results, we believe that the candidate SNPs with $\beta \geq 0.75$ are trustworthy, because the percentages of the candidate SNPs in dbSNP are higher than 0.95 for both PC3 and H1 data. To determine whether the cutoff is appropriate, we adopt the method introduced by Newton et al. [12] and Müller et al. [13] to control false discovery rate (FDR). For a given cutoff β_0, we define $\text{FDR}(\beta_0) = \sum_{i=1}^{n}(1-\beta_i)I(\beta_i \leq \beta_0) / \sum_{i=1}^{n} I(\beta_i \leq \beta_0)$ where $I_{(\cdot)}$ is the indicator function. Fig. 3 plots the estimated FDR versus selected positions whose β are larger than 0.75. We can see that the FDR is estimated to be less than 0.1.

Because of the large data size of PC3 data (14 GB) and H1 data (200 GB) after mapping, computation is challenging here. We employ an efficient way to implement the algorithm in C++. The program outputs posterior probability for each position with mismatch(es). The PC3 data analysis is carried out on iMac, which equips with 2.8 GHz Intel Core i7 CPU and 16GB memory. The calculation is done within 3 hours. openMP option is enabled to facilitate the parallel computing by chromosome. The H1 data analysis is carried out on MDACC high performance computing (HPC) cluster, which equips with AMD Opteron(tm) Processor 6128 HE, and 32 GB RAM per node. The calculation is performed for each chromosome in parallel and all the calculations finishes within 20 hours. The code is available upon requested.

V. Conclusion

We propose a Bayesian method to detect SNPs in this paper and apply our method to two data sets: prostate cancer cell line PC3 and stem cell line H1. While the sequencing depth of PC3 is low and uneven and H1 has relatively high and uniform sequencing depth, our method works well for both. Our result will provide a useful reference to common cell lines. The data was originally obtained for epigenetic studies of histone modifications using ChIP-Seq technique. We show that the data originally intended for ChIP-Seq studies can be mined for SNP information. Since there are thousands of ChIP-Seq experiments conducted each year, we expect our method to have a wide range of applications.

The proposed Bayesian SNP calling method utilizes quality score of the sequence reads and mismatch profiles between the unique reads and the reference genome in determining the variants. The method is fast, capable of processing whole genome data at 20-fold average coverage in reasonable amount of time. We show that our method is substantially better than MAQ in that it finds more SNPs with higher quality.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We gratefully acknowledge financial support from the National Cancer Institute through grant NCI 5 K25 CA123344 (Shoudan Liang) and NIH R01 CA132897 (Yuan Ji).

References

1. Siva N. 1000 genomes project. Nature biotechnology. 2008; 26(3):256–256.
2. Stephens M, Sloan J, Robertson P, Scheet P, Nickerson D. Automating sequence-based detection and genotyping of snps from diploid samples. Nature genetics. 2006; 38(3):375–381. [PubMed: 16493422]
3. Zhang J, Wheeler D, Yakub I, Wei S, Sood R, Rowe W, Liu P, Gibbs R, Buetow K. Snpdetector: a software tool for sensitive and accurate snp detection. PLoS computational biology. 2005; 1(5):e53. [PubMed: 16261194]
4. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Van Broeckhoven C, De Rijk P. novosnp, a novel computational tool for sequence variation discovery. Genome research. 2005; 15(3):436–442. [PubMed: 15741513]
5. Li H, Ruan J, Durbin R. Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome research. 2008; 18(11):1851–1858. [PubMed: 18714091]
6. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski E, Liu Y, Weinstock G, Wheeler D, Gibbs R, et al. A snp discovery method to assess variant allele probability from next-generation resequencing data. Genome research. 2010; 20(2):273–280. [PubMed: 20019143]
7. Ji Y, Xu Y, Zhang Q, Tsui K, Yuan Y, Norris C Jr, Liang S, Liang H. Bm-map: Bayesian mapping of multireads for next-generation sequencing data. Biometrics. 2011
8. Tanner M, Wong W. The calculation of posterior distributions by data augmentation. Journal of the American statistical Association. 1987:528–540.
9. Langmead B, Trapnell C, Pop M, Salzberg S, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol. 2009; 10(3):R25. [PubMed: 19261174]
10. Quinlan A, Hall I. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. [PubMed: 20110278]
11. Sherry S, Ward M, Kholodov M, Baker J, Phan L, Smigielski E, Sirotkin K. dbsnp: the ncbi database of genetic variation. Nucleic acids research. 2001; 29(1):308–311. [PubMed: 11125122]
12. Newton M, Noueriry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric heirarchical mixture model. Biostatistics. 2004; 5:155–176. [PubMed: 15054023]
13. Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing. Journal of the American Statistical Association. 2004; 99(468):990–1001.

The proposed Gibbs sampler is as follows:

- Step 1: Let, $\beta_{it}^{(1)}=\varepsilon$, where ε is an arbitrary small probability close to zero for i = 1, 2, 3.
- Step 2: In the s-th iteration, sample $u_k^{(s)}$ from

$$Bernoulli\left\{\frac{(1-\alpha_{kt})\beta_{kt}}{\alpha_{kt}+(1-\alpha_{kt})\beta_{kt}}\right\}.$$

- Step 3: Sample $\boldsymbol{\beta}_t = (\beta_{1t}, \beta_{2t}, \beta_{3t})$ from (3).
- Step 4: Iterate steps 1 to 3 S times, for a large integer S.

For the special case in which $N_{it} = 0$, set $u_k^{(s)}=0$ and $\beta_{it}^{(s)}=\beta_{it}^{(s-1)}$ for i = 1, 2, 3.

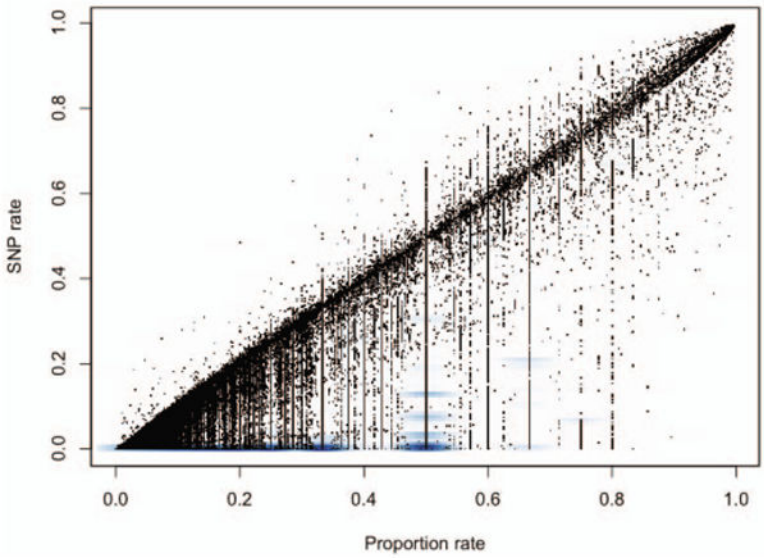

Fig. 1.

(Colored Figure) The smoothed density plot with color representation of posterior probability of hidden nucleotide versus proportion of nucleotide change at potential SNP positions.

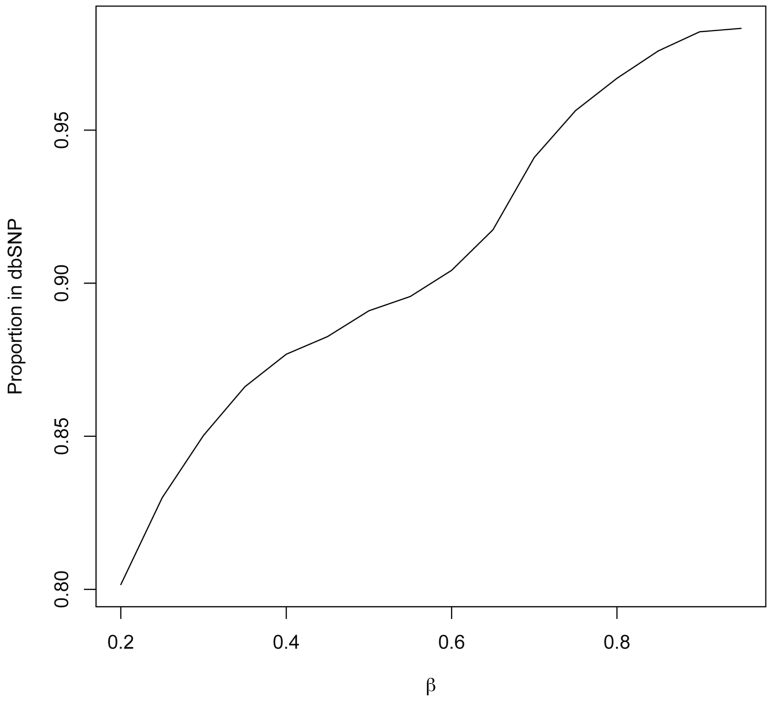

Fig. 2.

The proportion of the identified candidate SNPs overlapping with dbSNP versus the cutoff of β.

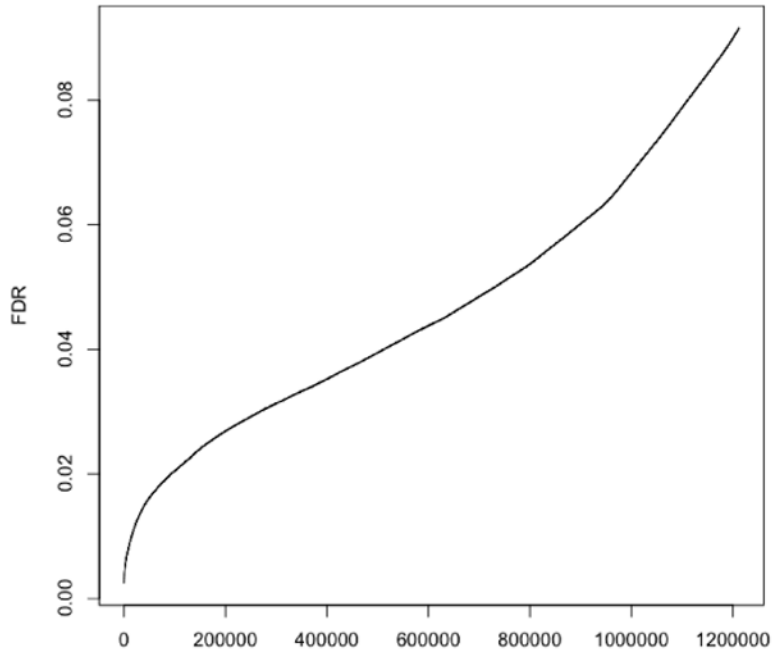

Fig. 3.
Bayesian FDR plot versus selected positions when $\beta_0 = 0.75$.