BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2498–2500
doi:10.1093/bioinformatics/btu338

Genome analysis Advance Access publication May 14, 2014

FaSD-somatic: a fast and accurate somatic SNV detection algorithm for cancer genome sequencing data

Weixin Wang[1,2,†*], Panwen Wang[1,2,†], Feng Xu[1,2,†], Ruibang Luo[3,4], Maria Pik Wong[5], Tak-Wah Lam[3,4] and Junwen Wang[1,2,6]

[1]Department of Biochemistry, LKS Faculty of Medicine, Hong Kong SAR, [2]Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, Guangdong 518057, [3]HKU-BGI Bioinformatics Algorithms and Core Technology Research Laboratory, [4]Department of Computer Science, [5]Department of Pathology and [6]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

Associate Editor: John Hancock

ABSTRACT

Summary: Recent advances in high-throughput sequencing technologies have enabled us to sequence large number of cancer samples to reveal novel insights into oncogenetic mechanisms. However, the presence of intratumoral heterogeneity, normal cell contamination and insufficient sequencing depth, together pose a challenge for detecting somatic mutations. Here we propose a fast and an accurate somatic single-nucleotide variations (SNVs) detection program, FaSD-somatic. The performance of FaSD-somatic is extensively assessed on various types of cancer against several state-of-the-art somatic SNV detection programs. Benchmarked by somatic SNVs from either existing databases or *de novo* higher-depth sequencing data, FaSD-somatic has the best overall performance. Furthermore, FaSD-somatic is efficient, it finishes somatic SNV calling within 14 h on 50X whole genome sequencing data in paired samples.

Availability and implementation: The program, datasets and supplementary files are available at http://jjwanglab.org/FaSD-somatic/.

Contact: wangdatou2009@gmail.com.

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

Received on January 21, 2014; revised on April 22, 2014; accepted on May 9, 2014

1 INTRODUCTION

Next-generation sequencing and third-generation sequencing are widely used to detect genetic mutations from tumor tissues and their paired normal tissues (from the same individual), providing an immense power for understanding the cause of cancer. Besides the inherited germ line mutations, the newly acquired somatic mutations constitute the majority of genetic mutations that cause cancer (Hua *et al.*, 2013; Stratton *et al.*, 2009).

Increasing evidences show that tumoral heterogeneity exists not only in intertumoral level but also in intratumoral level (Gerlinger *et al.*, 2012; Swanton, 2012). When the sequencing depth is insufficient, the minor clones, which may harbor the meaningful positively selected mutations, could be easily concealed by the dominant clones. Furthermore, owing to the presence of non-tumor cells in the bulk of tumor, the proportion of effective covered reads with somatic mutations will decrease, therefore making somatic mutation detection more difficult. Finally, the germ line variants can outnumber somatic variants by several orders of magnitudes (Pleasance *et al.*, 2010); this further complicates the difficulty of somatic mutation detection.

Efforts have been made by both biologists and bioinformaticians ato overcome those challenges, but the methods to call somatic single-nucleotide variations (SNVs) directly from sequencing data with accuracy and efficiency are still in demand. (Supplementary Section 1).

2 METHODS

2.1 Model of FaSD-somatic

We assume that tumor samples could be modeled as a mixture of a normal clone and several tumor clones. The tumor clones include a dominant and several minor clones. Our model considers not only the dominant tumor clone, where majority of somatic SNVs come from, but also the minor ones with relatively lower frequency of somatic mutations. Because all clones are from the same individual, the dependency of tumor and normal clones could be used to infer genotypes (Larson *et al.*, 2012). We first count genotypes of all possible diploid clones that may exist in tumor sample, then use the FaSD model for SNPs calling (Xu *et al.*, 2012). Finally, we use a joint genotype likelihoods model to analyze the tumor-normal paired sequencing data and infer high-quality somatic SNVs. (Supplementary Sections 2 and 3).

2.2 Implementation

FaSD-somatic is a command line tool, written in C/C++ and can be compiled in any computing environment with g++ support. It takes a variety of input file formats including SAMtools pileup and BAM files. To speed up the program, we tried strategies, such as earning time from space. For example, for the comparison of two letters (reads), an array containing 26 Boolean values was used instead of 26 letters. Then, the comparison between letters became a Boolean value comparison, which saved much time. For the functions to handle the BAM files, a set of utilities written by C that can manipulate the alignments in BAM or SAM files were borrowed from SAMtools (Li *et al.*, 2009). SAMtools source code was modified and integrated into FaSD-somatic, including the conversion from the implicit variable type to explicit to fit the compiler. We also provided a script called runSOMA.pl to run FaSD-somatic in multi-threads, which significantly accelerates the process.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

2.3 Evaluation metrics

Because large-scale experimental validation of somatic and non-somatic variants is impractical, we measured the concordance of the somatic predictions with databases known to be enriched for either somatic or germ line mutations. Catalog of somatic mutations in cancer (COSMIC) v64 (Forbes *et al.*, 2011) was used to generate the gold standard for somatic mutations. The gold standard for non-somatic mutation was built by excluding COSMIC somatic SNVs from curated germ line mutations in the phase 1 of 1000 genomes project.

To test the callers' performance specifically on the shallow depth data, we applied the receiver operating characteristic analysis. The VarScan2's high-confidence (HC) somatic SNVs calling set on the higher-depth lung adenocarcinoma (LUAD) data (~40X) is used as the benchmark. (Supplementary Sections 4 and 5).

3 RESULTS

3.1 Evaluation benchmarked by known databases

As shown in Figure 1A, in the paired tumor and normal LUAD samples with the lowest sequencing depth of 4X each, FaSD-somatic has the highest concordance with the COSMIC validated somatic gold standard among all five somatic SNVs callers (maximum concordance value 0.0079 compared with Varscan2's 0.0060, SomaticSniper's 0.0039, JointSNVmix's 0.0023 and SAMtools' 0.0011). Even if the quality threshold decreases and the number of predictions increases simultaneously, FaSD-somatic still had the highest concordance with somatic mutation benchmark. In the aspect of the concordance with the 1000 Genomes Project validated germ line benchmark, FaSD-somatic is <20% at the beginning, and it gradually decreases to reach the third highest position with a value of 10% (Fig. 1B). We further evaluated the programs for other cancer types. In the 6X glioblastoma multiforme (GBM) (Supplementary Fig. S1) and 50X lung squamous cell carcinoma (LUSC) (Supplementary Fig. S2) samples, similar trends were observed.

3.2 Evaluation benchmarked by *de novo* higher-depth data

First, five software were used to call somatic SNVs on a 4X LUAD dataset, and the results were benchmarked by the Varscan2's HC calling result on the 40X LUAD data sequenced from the same sample because somatic SNV calling at this depth is considered of high quality (Wang *et al.*, 2013). The AUC of FaSD-somatic has a mean value of 0.801, and the 95% non-parametric confidence interval is [0.765, 0.835], which is significantly higher than other software. Then, the 50% subsampled data from benchmark itself was also evaluated. FaSD-somatic outperformed others, especially in the loci with a sequencing depth lower than 10X (Supplementary Figs S3–S5 and Supplementary Tables S1–S3).

3.3 Processing speed

The time taken for these tools to process the data is a major bottleneck for sequencing data analysis. All five programs were tested on a server, with 2.00 GHz Intel(R) Xeon(R) CPU E5-2620 and 64 GB memory. Based on one thread of single core of CPU, FaSD-somatic can finish the whole genome somatic SNVs calling within 4815, 8601 and 49 807 s, respectively, on 4X LUAD, 6X GBM and 50X LUSC datasets, which is 38% faster than SomaticSniper, 62% faster than SAMtools, 113% faster than VarScan2 and 501% faster than JointSNVmix (Supplementary Fig. S6 and Supplementary Table S4).

Funding: Research Grants Council, Hong Kong SAR, China (grant number 781511M); National Natural Science Foundation of China, China (grant number 91229105);

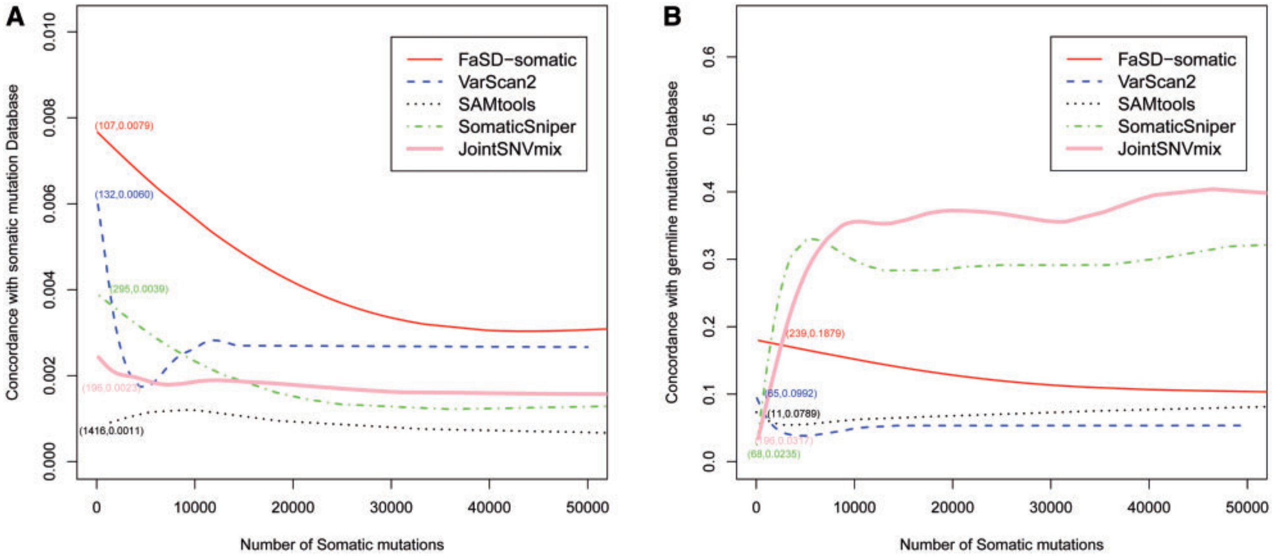

Fig. 1. Somatic (**A**) and germ line (**B**) concordance analyses of paired tumor and normal LUAD samples. The horizontal axis shows the number of predicted somatic SNVs; the vertical axis represents the fraction of those predictions found to be concordance with (A) the merged COSMIC somatic SNVs set and (B) the filtered 1000 genomes germ line SNPs set. A perfect Somatic SNVcalling program should have high concordance in (A) and lower concordance in (B)

Innovation and Technology Fund, Hong Kong SAR, China (grant number GHP/011/012).

Conflict of interest: none declared.

REFERENCES

Forbes,S.A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–950.

Gerlinger,M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, **366**, 883–892.

Hua,X. *et al.* (2013) DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am. J. Hum. Genet.*, **93**, 439–451.

Larson,D.E. *et al.* (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Pleasance,E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–196.

Stratton,M.R. *et al.* (2009) The cancer genome. *Nature*, **458**, 719–724.

Swanton,C. (2012) Intratumor heterogeneity: evolution through space and time. *Cancer Res.*, **72**, 4875–4882.

Wang,W. *et al.* (2013) Assessment of mapping and SNP-detection algorithms for next-generation sequencing data in Cancer Genomics. In: *Next Generation Sequencing in Cancer Research*. Springer, New York, NY, pp. 301–317.

Xu,F. *et al.* (2012) A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat. Commun.*, **3**, 1258.