SPAI: Single Platform for Analyzing Indels

Mohammad Shabbir Hasan, Liqing Zhang

Department of Computer Science
Virginia Tech, Blacksburg, VA 24061, USA.
E-mail: shabbir5@vt.edu, lqzhang@vt.edu

Abstract. SPAI which stand for Single Platform for Analyzing Indels is a workbench which is intended to aid in the research on indel calling. Unlike other existing command line tools which needs expertise from Computer Science (CS) to run them, we emphasis here to create an environment where users from non-CS background can run the existing indel calling tools using a Graphical User Interface (GUI). In addition to that, this interactive tool written in Java also provide several features that include downloading alignment files (BAM files) from the 1000 Genomes project, viewing the alignment files and variant files in a tabular format, cross-validation among the indel calling results from different tools, comparing the results from different indel calling tools with the benchmark dataset, and graphically visualizing the comparison results.

Keywords: Indel Calling; Variant Calling; Next Generation Sequencing; Deep Sequencing.

1 Introduction

Indel corresponds to the insertion or deletion of base pairs in the DNA sequence. It is a common form of genetic variations in human population genome [1]. Indels have been identified to cause diseases such as Cystic fibrosis, Fragile X Syndrome, Trinucleotide repeat disorders, Mendelian disorders, Bloom Syndrome, acute myeloid leukemia, and lung cancer. Moreover, indels in the promoter region explains the differences in gene expression observed among different human and can be used as genetic markers in natural populations [2]. With the advancement in Next Generation Sequencing (NGS), whole genome sequencing is now possible at an individual level and the result of calling indels from the whole genome of an individual can be used to determine the future health of the individual and thus develop customized medical treatment. Therefore, all of these indicate the importance of analyzing indels to develop effective treatment and medicine for human.

With the growing interest on indel calling research, a good number of indel calling tools have been developed so far [4]. Some of the tools include Genome Analysis Tool Kit (GATK) [3], Dindel [4], VarScan [5], Pindel [6], SAMtools [7], and P-Dindel [8]. All of these tools are very popular among the researchers doing indel calling research. Being command line tools, however, most of them require some exper-

tise from Computer Science (CS) to run these tools. Therefore, researchers specially the beginners from non-CS background such as Biology, Chemistry, and Microbiology etc. often find it difficult to execute and explore different features of the tools by changing different parameters through command line. Research on "Graphical User Interface" (GUI) by different usability laboratories have showed that the usability of a product can be significantly increased through a user friendly GUI [9]. Therefore, here we propose SPAI, a GUI based tool for analyzing indels from the scratch in a convenient way.

2 Description

SPAI which is written in Java comes with several features that are described in brief in this section.

2.1 Running different indel calling tools using GUI

SPAI facilitates user to run 5 indel calling tools using the GUI. These 5 tools include GATK [3], Dindel [4], VarScan [5], Pindel [6], and SAMtools [7]. The GUI allows user to run these tools by just clicking buttons instead of writing command lines in the terminal. For advanced users, they can change the default values for different parameters from the GUI as well. Fig. 1 shows the main GUI of SPAI.

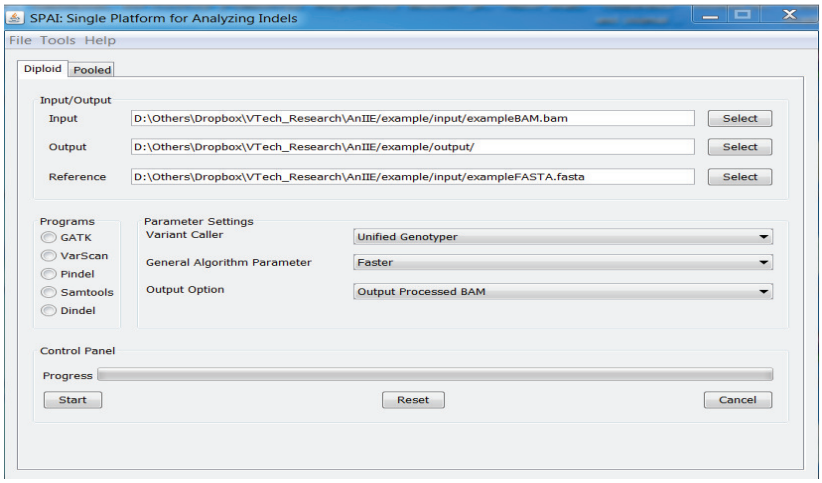

Fig. 1. Main Window of SPAI.

2.2 Downloading Alignment files from the 1000 Genomes Project

To run the indel calling programs, users need to use the Alignment files (in BAM format) as inputs. It is inconvenient to download large BAM files using the browser,

and using command lines to do the same things needs some expertise. Using the GUI, SPAI allows users to download single or multiple BAM files directly from the FTP server of the 1000 Genomes project. User just need to select which BAM file(s) to download and those files are downloaded and stored in the Download folder. Fig. 2 shows the download window of SPAI.

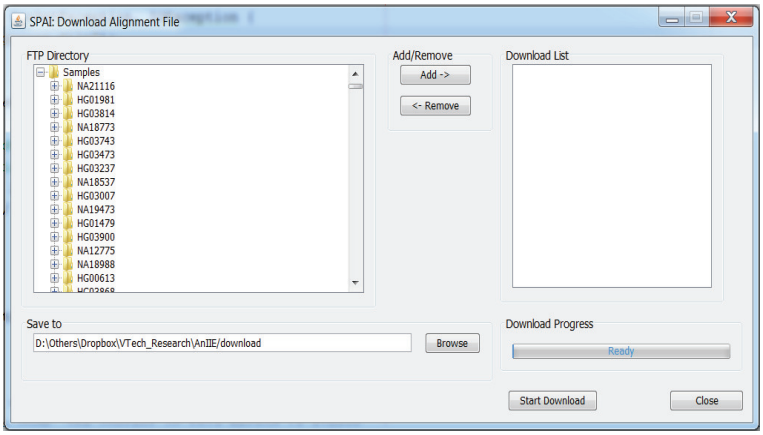

Fig. 2. Download window of SPAI

2.3 Comparing the results with other indel calling tools

User can compare the results of different indel calling tools using SPAI. A benchmark dataset for human genome [10] is given with this tool and SPAI calculates the recall, precision and F-measure for each tool based on that. User can also change the benchmark dataset and provide a different one. This feature is really useful when the users need to assess the performance of any indel calling tool. The results of comparison are shown as graphs which make it easy to interpret.

2.4 Showing the large BAM files and VCF files in a tabular format

The input for most of the indel calling tool is the alignment file in BAM format. BAM format is a binary format, so it can't be opened using a regular text editor. However, it contains several information that the users might find useful. So SPAI uses a tool called BAMSeek [11] that allows user to view large BAM file in a tabular format. Indel calling tools produce the outputs in VCF format which can be opened using a text editor. However, if the file is too large, text editors can't handle them properly. To get rid of this problem, SPAI allows user to view a large VCF file in a tabular format.

2.5 Determining the average depth of coverage

Depth of coverage is the average number of reads that represent a given nucleotide in the sequence. In most cases high depth of coverage is desired for calling indels confidently. SPAI allows users to calculate the depth of coverage of an alignment file.

3 Conclusion and Future plan

SPAI allows users to run indel calling tools without prior knowledge of command line programming. We believe it is a helpful tool for people from non-computer science background. In addition to the current features, in future we plan to add several other features which will be helpful for downstream analysis. For example, determining redundant indels by creating a universal positioning system of the indels will be added to this tool. By facilitating to use the tools developed using computer science algorithms, SPAI creates a bridge among people from different disciplines doing research in Bioinformatics by providing everything necessary for an interdisciplinary research project in a single platform.

References

1. T. R. Bhangale, M. J. Rieder, R. J. Livingston, and D. A. Nickerson, "Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes," *Human molecular genetics,* vol. 14, pp. 59-69, 2005.
2. M. S. Hasan, X. Wu, and L. Zhang, "Performance evaluation of indel calling tools using real short-read data," *Submitted,* 2015.
3. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl*, et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature genetics,* vol. 43, pp. 491-498, 2011.
4. C. A. Albers, G. Lunter, D. G. MacArthur, G. McVean, W. H. Ouwehand, and R. Durbin, "Dindel: accurate indel calls from short-read data," *Genome research,* vol. 21, pp. 961-973, 2011.
5. D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis*, et al.*, "VarScan: variant detection in massively parallel sequencing of individual and pooled samples," *Bioinformatics,* vol. 25, pp. 2283-2285, 2009.
6. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning, "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads," *Bioinformatics,* vol. 25, pp. 2865-2871, 2009.
7. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer*, et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics,* vol. 25, pp. 2078-2079, 2009.
8. M. S. Hasan and L. Zhang, "P-Dindel: a multi thread based tool for calling indels from short reads," *Submitted,* 2015.
9. W. O. Galitz, *The essential guide to user interface design: an introduction to GUI design principles and techniques*: John Wiley & Sons, 2007.
10. R. E. Mills, W. S. Pittard, J. M. Mullaney, U. Farooq, T. H. Creasy, A. A. Mahurkar*, et al.*, "Natural genetic variation caused by small insertions and deletions in the human genome," *Genome research,* vol. 21, pp. 830-839, 2011.
11. BAMSeek. Available: https://code.google.com/p/bamseek/