BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 599–600
doi:10.1093/bioinformatics/btr711

Data and text mining

Advance Access publication December 30, 2011

VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer

Jamie K. Teer[1], Eric D. Green[1], James C. Mullikin[1,]* and Leslie G. Biesecker[1]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: VarSifter is a graphical software tool for desktop computers that allows investigators of varying computational skills to easily and quickly sort, filter, and sift through sequence variation data. A variety of filters and a custom query framework allow filtering based on any combination of sample and annotation information. By simplifying visualization and analyses of exome-scale sequence variation data, this program will help bring the power and promise of massively-parallel DNA sequencing to a broader group of researchers.

Availability and Implementation: VarSifter is written in Java, and is freely available in source and binary versions, along with a User Guide, at http://research.nhgri.nih.gov/software/VarSifter/.

Contact: mullikin@mail.nih.gov

Supplementary Information: Additional figures and methods available online at the journal's website.

Received on October 11, 2011; revised on December 6, 2011; accepted on December 21, 2011

1 INTRODUCTION

Massively-parallel DNA sequencing technologies are rapidly maturing, and are now being applied to a wide variety of projects of different scales and aims. A frequent goal among these projects is to examine inter-individual genomic variation, and to correlate this variation with phenotype. This approach has proven useful in medical sequencing by identifying causative variants in disease (Rios *et al.*, 2010; Sobreira *et al.*, 2010; Wei *et al.*, 2011) (for further examples, see (Mardis and Wilson, 2009; Ng *et al.*, 2010; Teer and Mullikin, 2010)). Although genome sequencing studies are becoming increasingly practical, the use of DNA-capture technologies (e.g., exome capture) extends the number of samples that can be investigated. The large number of sequence variants that are typically identified in such projects can be daunting. For example, a single exome sequencing experiment can lead to the detection of >200 000 variants. While these variant lists can be reduced using custom filtering scripts, or command-line filtering programs like GATK (McKenna *et al.*, 2010), this requires significant bioinformatics knowledge. These large datasets therefore present a challenge for those with limited bioinformatics expertise and resources.

Various programs have been written to visualize massively-parallel DNA sequence reads and their alignments (for review, see (Nielsen *et al.*, 2010)), but are not designed to view or analyze the identified variants. Some options for viewing annotated sequence variants are available (SVA (Ge *et al.*, 2011) or Galaxy (Goecks *et al.*, 2010)), but are designed more as analysis packages with limited viewing options and require dedicated high-performance workstations or servers. Recently, a probabilistic search tool designed to identify causative variants and genes, VAAST, was described (Yandell *et al.*, 2011). Although powerful, this analysis tool is designed to identify the genes most likely to be involved in disease, not to browse variant data.

There is a current need for analysis systems that allow investigators to view and manipulate variants identified using next-generation sequencing technologies, but that do not require high-end computational infrastructure or end-user bioinformatics expertise. Such systems would be designed with genetic variation data in mind, and would allow the user to apply expert knowledge in many different ways when deciding how to filter or prioritize the variants, allowing for further data mining possibilities beyond those offered by other tools with static filtering strategies. Towards that end, we have developed a program (VarSifter) that allows researchers to view exome-scale sequence variation and to perform the sorting, filtering, and searching required to analyze these data for biological relevance.

2 RESULTS

2.1 Rationale

We sought to develop a software tool that would empower individual researchers to readily analyze sequence variation data generated by next-generation DNA sequencing technologies. This system would have a graphical interface, would operate on any desktop-class computer, and would offer common and customizable filtering routines. Many programs have been developed to perform initial alignment, genotype determination, and variant annotation using next-generation sequencing data. We therefore focused our efforts on displaying and manipulating data generated using these programs; this program would be run by end-users to visualize and work with the results of genotype calling and annotation pipelines.

2.2 File formats

VarSifter accommodates Variant Call Format v4.0 (VCF, http://www.1000genomes.org) files, as well as a structured tab-delimited text file that can be imported into common spreadsheet programs (see Supplemental User Guide for details). Both formats can contain annotation information as well as genotypes for many samples.

2.3 Graphical display

Variant data are displayed in several panels using a graphical user interface (Supplemental Fig. S1). Annotation information is

*To whom correspondence should be addressed.

Published by Oxford University Press 2012.

displayed as a table in the Annotation Panel, with each row containing a single variant. All annotation columns are sortable to facilitate variant prioritization. Selecting a variant row displays the sample names, genotypes, quality scores, and read depth for each sample (all sortable) in a Sample Panel, allowing rapid access to sample genotypes.

Currently, there is no dominant standard for describing the position of a variant, particularly for insertions and deletions. We have developed a variant-position numbering system that describes both substitutions and insertions/deletions using the left and right flanking positions and the bases in between (Supplemental Fig. S2). VarSifter displays variant coordinates using this unambiguous system.

2.4 Filtering

A major challenge in analyzing massively-parallel sequencing data is assessing which of the large number of identified variants are biologically important. To reduce the initial list of variants, various filtering strategies (inclusive and exclusive) can be implemented that utilize pre-calculated annotation information. Variant-type filters (e.g., synonymous, non-synonymous, stop, etc.) are generated dynamically from annotations in the data file, allowing flexibility in annotation nomenclature. When family information is pre-generated (see file format description), additional filters can be used to display those variants fitting the desired inheritance model. Variants in genes of interest can be identified by entering the gene name is a search box, and a flexible syntax for text matching (regular expression) is supported. Lists of gene names or genomic regions can also be used for additional filtering.

While these filtering strategies have proven to be useful for many common queries, users may want to filter variants in other ways. We therefore implemented a custom query framework that allows the user to generate a hierarchical set of filtering criteria using any combination of genotype comparisons among samples and annotation filters. Sample comparisons allow the development of filters for alternate inheritance models. Annotation filters allow filtering of custom annotations (including any INFO field in the VCF file format.) Individual tests can be linked together using logical "AND/OR" connections, and the logical connections themselves can be connected (Supplemental Fig. 3). Although the filter logic can be complex, it is intuitively displayed as a network graph using the Java Universal Network/Graph Framework (JUNG) (jung.sourceforge.net). Such custom filtering strategies allow for many possible inheritance models, sample comparisons, or annotation filters.

2.5 Performance

We compared the performance of a widely used spreadsheet-based program (Microsoft Excel®) to that of VarSifter. Although VarSifter offers numerous useful functions for the analysis of exome data that this spreadsheet program does not, it should perform comparably. We compared RAM (random access memory) usage and startup time with VarSifter and Microsoft Excel for Mac 2011 across files of increasing size. VarSifter uses a similar amount of memory compared to the spreadsheet program and requires much less time to load the data (Supplemental Fig. S4). We found that memory usage increased quadratically with increasing sample number, as each additional sample (more columns) generally adds a number of unique variants (more rows). A total of 5.3 GB was required to load all detected variants in 160 exomes.

3 DISCUSSION

We developed VarSifter to support the need of biomedical researchers to visualize and interpret variation data from massively-parallel sequencing platforms. This tool reads pre-annotated variation data from several file formats, and allows users to search, sort, and sift through large sets of variants. Although the program is designed to handle exome-scale data, genome-scale data can be viewed by pre-filtering the data file on individual genomic regions or annotation type. VarSifter has been used successfully to assist in the discovery of biologically important variants linked with human disease (Lindhurst *et al.*, 2011; Sloan *et al.*, 2011; Wei *et al.*, 2011). Overall, this tool will allow groups without dedicated bioinformatics resources to interpret and analyze variation data.

ACKNOWLEDGEMENTS

We thank the following VarSifter beta users for their input: Steve Gonsalves, Jennifer Johnston, Marjorie Lindhurst, and David Ng from the Biesecker lab; Praveen Cherukuri, Pedro Cruz, and Nancy Hansen from the Mullikin Lab; David Adams, Karin Fuentes-Fajardo, Thomas Markello, and Murat Sincan from the NIH Undiagnosed Diseases Program; and Yardena Samuels and Peter Chines. We thank Mark Fredriksen and the NHGRI Bioinformatics and Scientific Programming Core for developing and hosting the VarSifter website.

Funding: This work was supported by the Intramural Research Program of the National Human Genome Research Institute.

Conflict of Interest: none declared.

REFERENCES

Ge, D., *et al.* (2011) SVA: software for annotating and visualizing sequenced human genomes, *Bioinformatics*, **27**, 1998–2000.

Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, *Genome Biol.*, **11**, R86.

Lindhurst, M.J., *et al.* (2011) A Mosaic Activating Mutation in AKT1 Associated with the Proteus Syndrome, *N. Engl. J. Med.*, **365**, 611–619.

Mardis, E.R. and Wilson, R.K. (2009) Cancer genome sequencing: a review, *Hum. Mol. Genet.*, **18**, R163–R168.

McKenna, A., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, **20**, 1297–1303.

Ng, S.B., *et al.* (2010) Massively parallel sequencing and rare disease, *Hum. Mol. Genet.*, **19**, R119–R124.

Nielsen, C.B., *et al.* (2010) Visualizing genomes: techniques and challenges, *Nat. Methods*, **7**, S5–S15.

Rios, J., *et al.* (2010) Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia, *Hum. Mol. Genet.*, **19**, 4313–4318.

Sloan, J.L., *et al.* (2011) Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria, *Nat. Genet.*, **43**, 883–886.

Sobreira, N.L., *et al.* (2010) Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene, *PLoS Genet.*, **6**, e1000991.

Teer, J.K. and Mullikin, J.C. (2010) Exome sequencing: the sweet spot before whole genomes, *Hum. Mol. Genet.*, **19**, R145–R151.

Wei, X., *et al.* (2011) Exome sequencing identifies GRIN2A as frequently mutated in melanoma, *Nat. Genet.*, **43**, 442–446.

Yandell, M., *et al.* (2011) A probabilistic disease-gene finder for personal genomes, *Genome Res.*, **21**, 1529–1542.