BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 7 2014, pages 1013–1014
doi:10.1093/bioinformatics/btt655

Sequence analysis

Advance Access publication November 9, 2013

AVIA: an interactive web-server for annotation, visualization and impact analysis of genomic variations

Hue Vuong*, Robert M. Stephens and Natalia Volfovsky†

Advanced Biomedical Computing Center (ABCC), Information Systems Program, SAIC-Frederick Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: The plethora of information that emerges from large-scale genome characterization studies has triggered the development of computational frameworks and tools for efficient analysis, interpretation and visualization of genomic data. Functional annotation of genomic variations and the ability to visualize the data in the context of whole genome and/or multiple genomes has remained a challenging task. We have developed an interactive web-based tool, AVIA (Annotation, Visualization and Impact Analysis), to explore and interpret large sets of genomic variations (single nucleotide variations and insertion/deletions) and to help guide and summarize genomic experiments. The annotation, summary plots and tables are packaged and can be downloaded by the user from the email link provided.

Availability and implementation: http://avia.abcc.ncifcrf.gov.

Contact: vuonghm@mail.nih.gov

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

Received on May 1, 2013; revised on August 26, 2013; accepted on November 7, 2013

1 INTRODUCTION

With the advances of next-generation sequencing technology, many laboratories and research groups generate huge amounts of sequencing data and face common challenges of identification, managing and interpretation of genomic variations. Recent surveys (Gnad *et al*., 2013; Liu *et al*., 2013; Pabinger *et al*., 2013) of software tools supporting basic computational steps of variation analysis present a comprehensive picture of the different options for analysis and help to identify the best practices in the field. The variety of the variation analysis pipelines is indicative of how each research group incorporates its own goals and expertise. Some of these pipelines are available as web-based applications wANNOVAR (Chang and Wang, 2012), VARIANT (Medina *et al*., 2012), Interpretome (Karczewski *et al*., 2012), CRAVAT (Douville *et al*., 2013) and snpEff (Ren *et al*., 2010) and can be used by a wider research community to facilitate genomic research. We have developed an interactive web-based tool, AVIA (Annotation, Visualization and Impact Analysis), to explore and interpret large sets of genomic variations generated by high-throughput genomic experiments such as exome capture, whole-genome sequencing and targeted re-sequencing projects. AVIA was implemented by adopting the ANNOVAR (Wang *et al*., 2010) framework and Circos (Krzywinski *et al*., 2009) visualization options, with an emphasis on interactive annotation and visualization capabilities and a modularized structure that allows facile extension of the pipelines with new tools and databases as they become available.

AVIA provides a basic functional impact assessment of small indels and single nucleotide variations based on their protein-coding capacity and/or position-associated ability to affect known non-coding regulatory elements and genomic features. Several of AVIA's workflows allow functional annotation and variant filtering options, tumor-normal comparisons and identification of population-specific variants. Interactive features of AVIA's interface permit exploratory analysis that can focus on selected gene sets or whole-genome data, with flexible visualization options and resubmission of a previously established workflow for the analysis of new datasets. A comprehensive tutorial posted on the AVIA site (http://avia.abcc.ncifcrf.gov/apps/site/tutorials) can guide users from a 'Quick-start' example to more rigorous analytical options.

2 TOOL DESCRIPTION

AVIA's general workflow is based on the coupling of a comprehensive annotation pipeline with a flexible visualization method (Fig. 1). This overview includes four major steps, starting with the user submitting a variation file, through annotation and visualization modules and ending with reporting of the results.

2.1 Annotation modules and workflows

We leveraged the ANNOVAR framework for assigning functional impact to genomic variations. AVIA expands ANNOVAR's capabilities in three different directions: by annotation content, parallelization of data processing and implementation of new predictive options. First, we extended ANNOVAR's list of reference databases [RefSeq, UCSC, SIFT (Kumar *et al*., 2009), Polyphen2 (Adzhubei *et al*., 2013), Encode (Krupp *et al*., 2012), etc], with additional databases both from recent publications by the scientific community (Supplementary Table S1) and from our in-house-developed databases. The framework was further extended by allowing the addition of third-party annotation databases provided by the user, advancing available filtering options and supplying access to ethnicity-specific-derived subsets of the Complete Genomics (CG) set of 69 genomes. Second, the original framework was adjusted to allow parallel computations of large lists of variations by subdividing the input set of variations into multiple subsets and querying multiple

*To whom correspondence should be addressed.
†Present address: Simons Foundation, New York, NY 10010, USA.

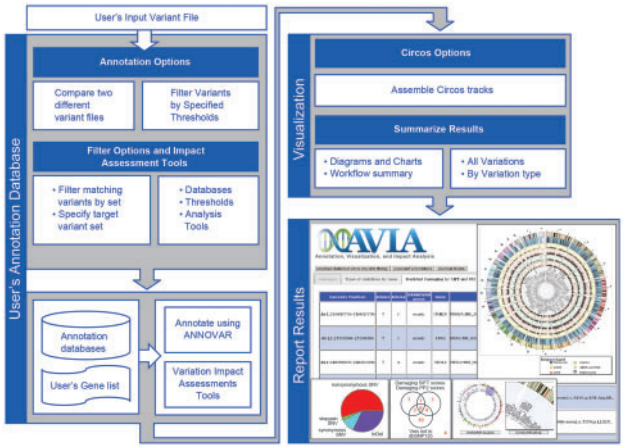

Fig. 1. Modular structure of AVIA workflow. The interface allows user to configure annotation, visualization and reporting options

databases simultaneously. This feature allows for results to be returned to the user in a timely manner (<15 min for 23 million mutations for the CG dataset, depending on server load). Finally, we added the ability to incorporate new analysis tools, thereby expanding the predictive capabilities of the pipeline beyond the static annotation databases ('MiRSNP' workflow).

The pipeline has been divided into workflows reflecting different approaches to the annotation of large lists of genomic variations. The first workflow, 'Feature Annotation,' provides annotation of all variations in the input list according to the databases selected by the user. The second workflow, 'Cascade filtering,' allows the gradual reduction of the number of the resulting variations by filtering all variations identified at each annotation step associated with a user's selected database. If more complex analysis is required, it can be explored using AVIA's 'Advanced workflow,' which allows comparison between variations of two input genomes, making it a useful tool for sample comparisons (tumor versus normal, affected versus unaffected, child versus parent and sample versus sample's ethnicity-specific population genome). The annotation results are reported in tabular format, summary diagrams and as tracks in whole genome circular views generated by the Circos application (example result files are available as Supplementary Table S2).

2.2 Interactive visualization with Circos

AVIA is focused on gene-related impact assessment. Individual Circos tracks showing, e.g. the distribution of genes with variations of specific functional effects such as non-synonymous variations, frame shifts, variable micro RNA target sites or variations in 5'UTR-located G-quadruplexes can be produced. Users can also choose to include additional precomputed genomic tracks from a select list into the Circos images. Alternatively, users may choose to display only part of the data focusing on user-provided gene lists, genomic regions or default sets of genes provided at AVIA's Web site [COSMIC (Forbes *et al.*, 2008)], Kyoto Encyclopedia of Genes and Genomes pathways (Kanehisa *et al.*, 2012). After the selection of the tracks/analysis of interest, users can aggregate the selected tracks in one signature plot: AVIA will produce a main Circos configuration file with the data corresponding to all selected individual tracks, and the new aggregated plot will be generated in several minutes depending on the amount of traffic on the server.

3 CONCLUSION

AVIA provides fast and comprehensive analysis of large sets of genomic variations in a user-friendly accessible manner. The main features of the current version are its interactive options for analysis, visualization and speed. The workflows can be easily configured to address a user's annotation questions and developed workflows can be used in new analyses. Many current software tools are focused on the analysis of the protein-coding variants, leaving most of the other types of variants beyond the scope of the analysis. In the future, we are planning to expand the AVIA's palette of non-coding impacts evaluated by the tool, including identification of both 'gain' and 'loss' of function regulatory variations.

ACKNOWLEDGEMENTS

The authors thank the members of Bioinformatics Support and Scientific Web Programming Groups at ABCC and ABCC's In Silico Research Center of Excellence.

Funding: National Cancer Institute, National Institutes of Health, under contract [HHSN261200800001E].

Conflict of Interest: none declared

REFERENCES

Adzhubei,I. *et al.* (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet.*, **Chapter 7**, Unit 7.20.

Chang,X. and Wang,K. (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.*, **49**, 433–436.

Douville,C. *et al.* (2013) CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics*, **29**, 647–648.

Forbes,S.A. *et al.* (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.*, Chapter 10, Unit 10 11.

Gnad,F. *et al.* (2013) Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, **14 (Suppl. 3)**, S7.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Karczewski,K.J. *et al.* (2012) Interpretome: a freely available, modular, and secure personal genome interpretation engine. *Pac. Symp. Biocomput.*, **2012**, 339–350.

Krupp,M. *et al.* (2012) RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, **28**, 1184–1185.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.

Liu,X. *et al.* (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.*, **34**, E2393–2402.

Medina,I. *et al.* (2012) VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic Acids Res.*, **40**, W54–W58.

Pabinger,S. *et al.* (2013) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.*, [Epub ahead of print, January 21, 2013].

Ren,J. *et al.* (2010) PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Mol. Cell Proteomics*, **9**, 623–634.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.