BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 2 2012, pages 277–278
doi:10.1093/bioinformatics/btr612

Genome analysis

Advance Access publication November 15, 2011

TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data

Yan W. Asmann†, Sumit Middha†, Asif Hossain*, Saurabh Baheti, Ying Li, High-Seng Chai, Zhifu Sun, Patrick H. Duffy, Ahmed A. Hadad, Asha Nair, Xiaoyu Liu, Yuji Zhang, Eric W. Klee, Krishna R. Kalari and Jean-Pierre A. Kocher*

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo College of Medicine, Rochester, MN, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: TREAT (Targeted RE-sequencing Annotation Tool) is a tool for facile navigation and mining of the variants from both targeted resequencing and whole exome sequencing. It provides a rich integration of publicly available as well as in-house developed annotations and visualizations for variants, variant-hosting genes and host-gene pathways.

Availability and implementation: TREAT is freely available to non-commercial users as either a stand-alone annotation and visualization tool, or as a comprehensive workflow integrating sequencing alignment and variant calling. The executables, instructions and the Amazon Cloud Images of TREAT can be downloaded at the website: http://ndc.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm

Contact: Hossain.Asif@mayo.edu; Kocher.JeanPierre@mayo.edu

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

Received on March 7, 2011; revised on September 10, 2011; accepted on October 31, 2011

1 INTRODUCTION

Next-generation sequencing offers the promise of scientific discovery with the challenge of results interpretation (Schuster, 2008). One experiment such as exome sequencing can generate tens of thousands of single nucleotide variants (SNVs) and small insertions or deletions (INDELs), which must be elucidated in the search for disease associated mutations (Ansorge, 2009; Metzker, 2010). Whole exome sequencing is an application of NGS that has been successfully used to identify disease-associated variants in several monogenic disorders (Gilissen *et al.*, 2010; Lupski *et al.,* 2010; Ng *et al.*, 2009, 2010) and complex diseases (Bonnefond *et al.*, 2010; Harbour *et al.,* 2010). While these studies demonstrated the power of NGS, they also highlighted the challenge of efficiently sifting through thousands of variants to identify a subset that is potentially clinically relevant. Bioinformatics solutions are beginning to be released that address this challenge and facilitate filtering and interpretation of human sequence variation data (Nix *et al.*, 2010; Sana *et al.*, 2011; Shetty *et al.*, 2010; Wang *et al.*, 2010). We developed TREAT to extend the functionality of these tools and directly integrated structured and sortable formats with embedded hyperlinks to sequence alignment, gene specificity and gene pathway visualizations. In addition, to enable broad accessibility, we have fully deployed TREAT to the Amazon Cloud. TREAT is optionally offered as part of a complete workflow for exome or targeted sequencing, providing users with a convenient method for integrated sequence alignment, mutation detection and results interpretation. We believe this tool offers investigators with an accessible and convenient method for annotating and visualizing sequencing data and a means of efficiently identifying variants of interest.

2 METHODS AND RESULTS

2.1 Variant annotation

TREAT provides four categories of variant annotations (Supplementary Figure S1): (i) the general variant annotations which provide the physical locations, and the dbSNP IDs and allele frequencies of known variants from HapMap and 1000 Genome Pilot Project in Caucasian (CEU), Yoruban (YRI) and East Asian (CHB/JPT) populations; (ii) sample-specific read depths supporting A, C, G, T bases at each variant position, and the quality scores for base calls and read mappings. These annotations are only available when the users choose to use TREAT for read alignment and variant calling; (iii) publically available annotations from SIFT (Kumar *et al.*, 2009) and SeattleSeq (http://gvs.gs.washington.edu/SeattleSeqAnnotation/) that include variant classifications (synonymous, missense, non-sense, frame-shift, etc.) and the predictions of the functional impact of the variants from SIFT and PolyPhen2 and (iv) in-house developed novel annotations including the tissue expression specificity measures for variant-hosting genes (detailed in Supplementary Data S2), and the identification of variants adjacent to exon–intron boundaries that potentially disrupt known splice-sites. An additional novel function of TREAT is the hyperlinks of each variant hosting gene to its associated KEGG pathway(s) (http://www.genome.jp/kegg) and Gene Ontology terms (http://www.geneontology.org/).

2.2 Reporting and visualization

TREAT automatically creates output in one easy-to-navigate HTML page, which provides the project description, QC reports, target coverage and sequencing depth information, descriptions of the annotations provided by TREAT, and links to the SNV and INDEL reports. The Microsoft Excel formatted SNV and INDEL reports provide row-based synopses of per-variant annotation. Each variant is hyperlinked to Integrative Genomics

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

Viewer (IGV) (Robinson *et al*., 2011) for the visualization of read alignments and variant calling information at the variant position. The functions of the variant hosting genes are illustrated via hyperlinks to the KEGG pathways and Gene Ontology terms, and the tissue expression specificity graph.

2.3 Access

TREAT is deployed in two formats, a standalone annotation application and an integrated version for an end-to-end analysis of exome or targeted sequencing data. The standalone annotation tool takes the list of called variants as input files and allows users the flexibility of generating the variants using alignment and variant calling tools of their own choosing. The integrated version accepts either FASTQ or BAM files as input files and carries out sequence alignment using BWA (Li and Durbin, 2009) or Bowie (Langmead *et al*., 2009), local sequence re-alignment (GATK; McKenna *et al.,* 2010) and variant calling (GATK or SNVMix; Goya *et al.*, 2010), which provides users with a convenient solution to their informatics needs. Both TREAT versions can be downloaded for local runs, or can be launched on the Amazon Elastic Compute Cloud (EC2) (http://en.wikipedia.org/wiki/Amazon_Elastic_Compute_Cloud) using Amazon Machine Images provided at our Website. The Machine Images are loaded with all the open-source tools and necessary annotation files for the direct execution of TREAT. The run time and cost estimate of TREAT Cloud version are provided in the Supplementary Data.

3 DISCUSSIONS

We have developed a bioinformatics tool, TREAT, which addresses the current challenges in analyzing and interpreting targeted and whole exome sequencing data. The annotations provided by TREAT have been carefully evaluated and selected from a pool of available open source tools and databases, and complimented by additional in-house developed annotations (details at the TREAT website). The variant reports in Excel format integrate the visualizations of the sequence alignment at variant positions, pathways and expression specificity of the variant hosting genes via clickable hyperlinks for each reported INDELs and SNVs. In addition, the summary of the targeted resequencing results is stored in a centralized HTML report with links to the TREAT website, the targeted region coverage report and the read QC report, the description of the TREAT workflow, and links to the website of the annotation tools and databases.

For maximum flexibility, two versions of TREAT were implemented: an annotation only version, and a version integrating read alignment, variant calling and annotations. Both versions can be downloaded as local installations or as Amazon Cloud images which makes TREAT available for users with no access to local bioinformatics infrastructures. By targeting all user groups and enabling rapid integration of emerging analytic methods, we believe that TREAT provides a sustainable NGS analytic workflow with wide applicability to the research community.

We plan to continue adding new functionality and features to TREAT to make it a comprehensive tool for targeted and exome analysis. These include the development of an in-house variant database that collects all variants detected from hundreds of individuals with various types of diseases using exome and whole genome sequencing. This database will provide critical annotations whether the observed variants are truly 'novel' or disease specific. In addition, we are in the process of making TREAT applicable to whole genome sequencing data analysis, this would require adding annotation tracks for non-coding regions such as the conservations and regulatory domains.

In summary, the rich set of annotations provided by TREAT, the easy to use, centralized HTML summary report, and the Excel-formatted variant reports with hyperlinked visualization utilities enable the filtering of detected variants based on their functional characteristics, and allow the researchers to navigate, filter and elucidate tens of thousands of variants to focus on potential disease-associated variant(s).

ACKNOWLEDGEMENTS

We want to thank Mrs Shannon McDonnell for her constructive suggestions and feedback during the implementation of this work flow. We are very grateful for the following investigators at Mayo Clinic who allowed us to use their exome sequencing data for the purpose of developing the work flow: Dr Stephen Thibodeau, Dr Ellen Goode and Dr Fergus Couch.

Funding: Development Fund from the Center for Individualized Medicine at Mayo Clinic Rochester MN, a generous gift from James and Donna Barksdale.

Conflict of Interest: none declared.

REFERENCES

Ansorge,W.J. (2009) Next-generation DNA sequencing techniques. *New Biotechnol.*, **25**, 195–203.

Bonnefond,A. *et al.* (2010) Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PloS One*, **5**, e13630.

Gilissen,C. *et al.* (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am. J. Hum. Genet.*, **87**, 418–423.

Goya,R. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.

Harbour,J.W. *et al.* (2010) Frequent mutation of BAP1 in metastasizing uveal melanomas, *Science,* **330**, 1410–1413.

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

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Lupski,J.R. *et al.* (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *New Engl. J. Med.*, **362**, 1181–1191.

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

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nature Rev.*, **11**, 31–46.

Ng,S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes, *Nature*, **461**, 272–276.

Ng,S.B. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.

Nix,D.A. *et al.* (2010) Next generation tools for genomic data generation, distribution, and visualization. *BMC Bioinformatics*, **11**, 455.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Sana,M.E. *et al.* (2011) GAMES identifies and annotates mutations in next-generation sequencing projects. *Bioinformatics*, **27**, 9–13.

Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat. Methods*, **5**, 16–18.

Shetty, A.C. *et al.* (2010) SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics*, **11**, 471.

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