Bioinformatics Advance Access published June 26, 2016

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw359
Advance Access Publication Date: 10 June 2016
Applications Note

Genome analysis

RareVariantVis: new tool for visualization of causative variants in rare monogenic disorders using whole genome sequencing data

Tomasz Stokowy[1,2,*], Mateusz Garbulowski[3], Torunn Fiskerstrand[1,4], Rita Holdhus[1], Kornel Labun[2], Pawel Sztromwasser[1,2], Christian Gilissen[5], Alexander Hoischen[5], Gunnar Houge[4], Kjell Petersen[2], Inge Jonassen[2] and Vidar M. Steen[1,4]

[1]Department of Clinical Science, University of Bergen, Bergen 5020, Norway, [2]Department of Informatics, Computational Biology Unit, University of Bergen, Bergen 5020, Norway, [3]Department of Informatics, Silesian University of Technology, Gliwice 44-100, Poland, [4]Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen 5021, Norway and [5]Department of Human Genetics, Radboud University Medical Center, Nijmegen 6525, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on March 4, 2016; revised on April 27, 2016; accepted on June 1, 2016

Abstract

Motivation: The search for causative genetic variants in rare diseases of presumed monogenic inheritance has been boosted by the implementation of whole exome (WES) and whole genome (WGS) sequencing. In many cases, WGS seems to be superior to WES, but the analysis and visualization of the vast amounts of data is demanding.

Results: To aid this challenge, we have developed a new tool—RareVariantVis—for analysis of genome sequence data (including non-coding regions) for both germ line and somatic variants. It visualizes variants along their respective chromosomes, providing information about exact chromosomal position, zygosity and frequency, with point-and-click information regarding dbSNP IDs, gene association and variant inheritance. Rare variants as well as de novo variants can be flagged in different colors. We show the performance of the RareVariantVis tool in the Genome in a Bottle WGS data set.

Availability and implementation: https://www.bioconductor.org/packages/3.3/bioc/html/RareVariantVis.html

Contact: tomasz.stokowy@k2.uib.no

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

1 Introduction

Human whole genome sequencing (WGS) is expected to become the standard sequencing method for screening and diagnostic examination of many rare diseases and syndromes with heterogenic or unknown etiology of presumed monogenic inheritance. This standardization is driven by the decreasing cost of WGS, more uniform coverage of the exome in WGS data as compared to whole exome sequencing (WES) data, and a need for better analysis and understanding of non-coding regions in the genome (Deciphering Developmental Disorders Study, 2015; Lelieveld *et al.*, 2015) . The analysis of human WGS data is challenging because of the data size (i.e. 80 GB of compressed raw data and 10–20 GB of variants/annotated data for a 30× genome), technical errors, alignment issues in tandem repeat regions or around centromeres (Szalkowski and

Anisimova, 2013), and variant calling problems (Liu *et al.*, 2013; Tan *et al.*, 2014). Currently, a number of software packages are available for visualization of WGS data, such as the Integrative Genome Viewer, Circos and the UCSC browser. Although these tools can efficiently visualize reads and variants across the genome, they do not provide easily accessible summary of rare, sample characteristic variants. It is therefore still a need for additional open source tools that can be applied by non-experts for the handling and visualization of WGS data in an experimental diagnostics setting, focusing on discovery of variants that may be causative for rare disorders of presumed monogenic inheritance (Gilissen *et al.*, 2014).

In this work, we present a new bioconductor package, RareVariantVis, being a simple tool for the filtering and visualization of human WGS data. RareVariantVis provides genome scale information that may replace microarray-based homozygosity mapping and exome sequencing. We demonstrate the performance of the tool in WGS data from the Genome in a Bottle Ashkenazim Trio sample (Complete Genomics data).

2 Materials and methods

The Ashkenazim Trio WGS samples (Zook *et al.*, 2014), sequenced with the Complete Genomics technology, was used in design of the RareVariantVis tool. This data set is publically available through the Personal Genome Project and the Genome in a Bottle project. Detailed description of original data is available at http://sites.stanford.edu/abms/content/giab-reference-materials-and-data, together with raw files in vcf and fastq formats. In our study, these data are limited to chromosome 21 from mother, father and affected son to satisfy Bioconductor limits for test data size and for simplicity. All necessary vcf file operations were performed using vcftools (Danecek *et al.*, 2011). Details of the data preparation for the RareVariantVis tool are provided in package vignette.

The RareVariantVis package is implemented in R and in part based on the following existing methods: package VariantAnnotation (Obenchain *et al.*, 2014) and package googleVis (Gesmann and de Castillo, 2011). The variant annotation package is used for data import, including vcf file reading. The googleVis package is used for dynamic visualization of rare variants (JavaScript d3 library).

3 Results

The main functionalities of the RareVariantVis package are variant filtering and visualization of rare variants and regions of homozygosity (for overview, see Supplementary Table S1). A unique option in RareVariantVis, as compared to other genome visualization tools, is presentation of rare variants (and clusters of them) characteristic for a particular sample. The chromosomeVis (for a single sample), trioVis (for trio samples) and multipleVis (for multiple samples) functions filter all variants within the sequence of a given chromosome and keep non-synonymous coding ones as default. In addition, filters for dbSNP frequency (default value: <0.01) and sequencing coverage (default threshold for read depth: >10) are applied. Variants that pass these filters are considered rare non-synonymous coding variants. In addition, users can filter variants by adding their own list of variant positions to be excluded.

The chromosomeVis function provides a visualization of all variants along a particular chromosome, with the corresponding frequency of the alternative allele, indicated as small blue dots (Fig. 1). This allele frequency can be computed as the ratio of the number of alternative variant reads to all reads in a particular genomic position. Allelic frequency is expected at 0,5 for heterozygous variants and 1 for homozygous, however due to sequencing errors, copy number variation and odd coverage, the calculated frequency values in practice vary from 0 to 1.

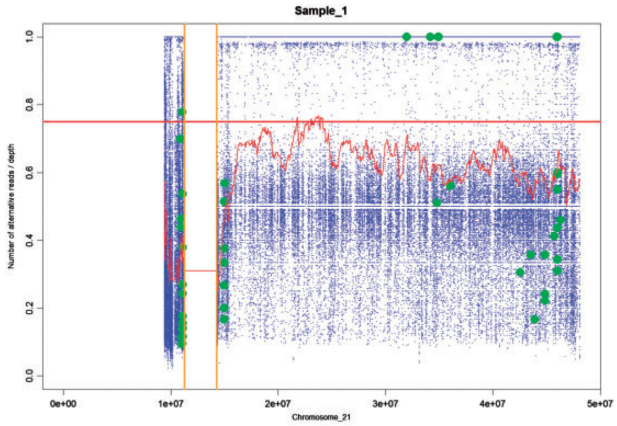

Fig. 1. Visualization of all variants on chromosome 21 from a Genome in a Bottle Ashkenazim Trio—WGS sample. The X axis represents position along the chromosome, whereas the Y axis displays the frequency of the alternative allele for each variant, calculated as the ratio of alternative reads count over the total number of reads. Blue dots represent all variants, both intergenic and intragenic. Green dots depict single nucleotide variants that are classified as rare, non-synonymous coding variants after the filtering. Vertical orange lines provide the position of the centromere. The horizontal red line is a moving average of the mean frequency of the alternative allele of nearby variants (defaults value: 2000 variants). The horizontal red line at an allele frequency of 0.75 represents the suggested cut-off between heterozygous and homozygous status for a given alternative allele

The rare, non-synonymous coding variants that remain after the filtering procedure (see above) are highlighted as green dots, in the example including four that are present in a homozygous state. As shown in Figure 1, the tool also includes visualization of the average frequency of the alternative allele of many nearby variants, displayed as a moving average of a set frame length (default value: 2000 variants). This function can rapidly indicate possible homozygosity regions (cut-off value: 0.75), which is often useful information in the genetic screening of diseases with presumed autosomal recessive inheritance, especially if the region also contains rare variants., as displayed in Supplementary Figure S1. The package may also point at regions with potential false positives that occur mainly due to technical challenges, which can be observed as a clustering of numerous rare, non-synonymous variants in about the same chromosomal position. In Figure 1, there are three such regions around positions 11, 15 and 46 Mb, including two regions in the centromere area that are prone to technical mapping artifacts, since centromeres consist of long tandem repeat stretches (Melters *et al.*, 2013).

The various RareVariantVis functions are also applicable in somatic alterations, for example in analysis of cancer genomes, where information about zygosity is essential. The trioVis function is particularly implemented for discovery of *de novo* variants in trio samples, and can also be useful in autosomal recessive and dominant cases (Supplementary Figs. S2 and S3). The visualization is dynamic, so that pointing at the variants will lead to display of an instant text box with information about the gene, position and inheritance (i.e. whether the variant has been inherited from mother, father or both of parents). The user may also mark and zoom on variants of interest. Potential *de novo* variants are flagged and visualized in red color. Detailed documentation of all functions is available in the Bioconductor vignette.

Acknowledgements

The authors would like to thank Winston Chang for implementation of one package function.

Funding

Bergen Research Foundation grant no. 807964, Meltzer Foundation grant ID 7761, POIG.02.03.01-24-099/13 grant: “GCONiI - Upper-Silesian Center for Scientific Computation”, and the Research Council of Norway (Centres of Excellence funding scheme, grant no. 223273; NORMENT CoE).

Conflict of Interest: none declared.

References

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Deciphering Developmental Disorders Study. (2015) Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, **519**, 223–228.

Gesmann,M. and de Castillo,R. (2011) Using the Google Visualisation API with R. *R Journal*, **3**, 40–44.

Gilissen,C. *et al.* (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature*, **511**, 344–347.

Lelieveld,S.H. *et al.* (2015) Comparison of exome and genome sequencing technologies for the complete capture of protein-coding Regions. *Hum. Mutat.*, **36**, 815–822.

Liu,X. *et al.* (2013) Variant callers for next-generation sequencing data: a comparison study. *PLoS One*, **8**, e75619.

Melters,D.P. *et al.* (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.*, **14**, R10.

Obenchain,V. *et al.* (2014) VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, **30**, 2076–2078.

Szalkowski,A.M. and Anisimova,M. (2013) Graph-based modeling of tandem repeats improves global multiple sequence alignment. *Nucl. Acids Res*, gkt628.

Tan,R. *et al.* (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat.*, **35**, 899–907.

Zook,J.M. *et al.* (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.*, **32**, 246–251.