BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 1 2013, pages 124–125
doi:10.1093/bioinformatics/bts627

Sequence analysis

Advance Access publication October 25, 2012

PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data

Lu Zhang, Jing Zhang, Jing Yang, Dingge Ying, Yu lung Lau and Wanling Yang*

Department of Paediatrics and Adolescent Medicine, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Hong Kong

Associate Editor: Michael Brudno

ABSTRACT

Next-generation sequencing has become a valuable tool for detecting mutations involved in Mendelian diseases. However, it is a challenge to identify the small subset of functionally important mutations from tens of thousands of rare variants in a whole exome/genome. Therefore, we developed a toolkit called PriVar, a systematic prioritization pipeline that takes into consideration calling quality of the variants, their predicted functional impact, known connection of the gene to the disease and the number of mutations in a gene, and inference from linkage analysis.

Availability: Executable jar package is available at http://paed.hku.hk/uploadarea/yangwl/html/software.html.

Contact: yangwl@hkucc.hku.hk

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

Received on May 22, 2012; revised on September 28, 2012; accepted on October 16, 2012

1 INTRODUCTION

Next-generation sequencing (NGS) has become a valuable tool for discovering potentially causal single-nucleotide variations (SNVs) as well as short insertions/deletions (indels) for Mendelian disorders. However, a large number of rare variants (SNVs and indels) exist in an individual, which poses an unprecedented challenge for pinpointing the causal mutations, not to mention a relatively high false rate in discovering rare SNVs/indels. To aid the process in identifying causal mutations from large sets of rare variants, we designed a toolkit named PriVar that provides a comprehensive pipeline for prioritizing coding SNVs and indels for further investigation. Four functional modules are included in this pipeline: variant annotation, call quality summary, prediction of functional impact for all the variants and candidate gene identification. A quality summary was used to help filter out most of the false positive variant calls based on a number of quality control parameters. Annotation of the variants provides essential information on them based on several databases. The functional prediction module predicts potential functional impact of variants based on evaluation of results from multiple existing algorithms. The candidate gene identification module is designed to consolidate information of linkage analysis, dosage of rare mutations in a gene, known connection of a gene with the disease in question and other relevant information. Using the four modules together, PriVar can efficiently reduce the set of potential causal variants to be considered for further studies.

*To whom correspondence should be addressed.

2 METHODS

PriVar is a Java application with four integrated modules including variant annotation, quality summary and variant filtering, prediction of functional impact of variants and candidate gene identification (Supplementary Fig. S1).

2.1 Variant annotation

PriVar uses a number of databases (Supplementary Material) to annotate the variants, such as alternative allele frequency (based on 1000 Genome project, ESP6500 and dbSNP), amino acid changes caused by the variant or a combination of variants (based on UCSC Genome Browser), and potential functional impact of variants and genes [dbNSFP (Liu *et al.*, 2011) and PROSPECTR (Adie *et al.*, 2005)]. PriVar can support both hg18 and hg19 coordinates in input files and annotate multiple samples simultaneously.

2.2 Quality summary and variant filtering

This module is used to keep only the reliable subset of variant calls for further analysis based on various parameters including variant quality, genotype call quality and sequencing depth, which can be applied either separately or jointly. PriVar summarizes the overall quality of exonic, splicing and non-exonic variant calls based on seven criteria, allowing the user to select appropriate quality cutoffs based on the overall quality of the data (Supplementary Material): (i) summary of overall quality; (ii) the concordance with SNP calls if SNP chip data is available; (iii) the proportion of calls that exist in dbSNP; (iv) transition to transversion ratio (Ti/Tv); (v) indel to SNV ratio; (vi) insertion to deletion ratio; and (vii) indel length distribution. When parental data are available, violation of Mendelian inheritance is also examined and presented to users. The false discovery rate of SNV calls are calculated by comparing the observed and expected Ti/Tv ratios (DePristo *et al.*, 2011).

2.3 Prediction of functional impact of variants

PriVar also prioritizes SNVs based on the predicted deleterious scores calculated by results from 10 mainstream mutation and gene evaluation algorithms (Supplementary Material) using logistic regression. Functional impact of amino acid changes is predicted based on both sequence conservation in evolutionary courses and physical chemical property of the reference and substitution amino acids. More information provides better separation between neutral and deleterious SNVs compared with other programs (Supplementary Fig. S2). For indels,

their functional impact is evaluated based on the prediction values of non-synonymous SNVs caused by the indel after a Sidak's correction.

2.4 Candidate gene identification

PriSNV contains six candidate gene selection functions to help choose a subset of genes of high interest for further analysis. The strategies for selection are controlled by the users depending on the data in hand and the assumptions for a particular study. These strategies can be used in conjunction with each other to increase power and to reduce search space.

2.4.1 Linkage-based strategy If multiple family members are affected, the genes within long identity-by-descent (IBD) regions shared by affected family members can be identified. Because of high noise of sequencing data, PriSNV only uses common SNV sites existing in HapMap for corresponding populations and follows a no-violation rule for identification of IBD sharing, while the haplotype allele frequency is also calculated to help evaluate the probability of sharing by chance (Supplementary Material).

2.4.2 Runs of homozygosity For diseases observed in families of consanguineous marriages family, it is highly likely that the causal mutation is in a region of run of homozygosity (ROH). The ROH regions are identified using only common SNVs and should satisfy the same criteria defined for IBD regions, similar to what we reported before (Zhang *et al*., 2011). ROH may also indicate large deletions or uniparental disomy.

2.4.3 'Double-hit' on a gene If recessive inheritance is assumed for a disease, genes carrying homozygous mutations or potential compound heterozygous variants will be identified by PriVar. Because phase is usually unknown, all genes with more than one rare non-synonymous SNV or indel will be highlighted for further analysis (Supplementary Material).

2.4.4 Mutation burden Because of genetic heterogeneity, mutations may be enriched in disease-related genes, although they may occur at different positions. PriSNV prioritizes the genes based on the percentage of patients carrying one or more rare non-synonymous SNVs or indels for a gene (Supplementary Material).

2.4.5 De novo mutations The *de novo* mutations occur during meiosis and only appear in a child but not the parents. If parental data are available and have high coverage and quality on the same position, the rare non-synonymous *de novo* SNVs/indels will be identified as potential causal mutations.

2.4.6 Disease candidate genes Candidate genes are selected based on previous studies on the disease. Well-established relationship between genes and diseases is extracted from Phenopedia database of HuGE Navigator (Yu *et al*., 2008). In addition, customized gene list is also accepted from users to pinpoint the candidate variants from those genes.

3 DATA AND RESULT

PriVar was applied to whole exome sequencing data from a patient affected by early-onset Crohn's disease. Raw reads were aligned to human reference genome hg19 using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) followed by Genome Analysis Toolkit (GATK) data processing pipeline to obtain variant calls (SNV and Indel) (Supplementary Material). Cutoff on sequencing depth was determined based on SNP call concordance between sequencing and SNP chip data, Ti/Tv ratio and indel length distribution (Supplementary Fig. S3). A 10-fold cutoff was determined to be an appropriate criterion for filtering out most of the false variants. A candidate gene-based strategy dramatically reduced the total number of genes to be considered from 1149 to 42, whereas a 'Double-hit'–based strategy filtered out another 32 genes that have only one novel variant (Supplementary Table S3). After removing variants with deleterious score below 0.8, only eight variants and five genes were left. Among them, *IL10RA* was eventually proved to be the disease-causing gene through a combination of genetics and functional characterization (Mao *et al*., 2012). The success of PriVar in defining causal mutation in this example indicates that it could provide significant help to biologists in prioritizing SNV and indel calls in a systematic way to reduce search space for further analysis and experimental verification. Of course it should be cautioned that a causal mutation may be missed from the original variant calls because of various reasons. More effort is needed to incorporate more biological and population genetics information into variant analysis to help with causal mutation identification.

Funding: Research Grant Council of the Hong Kong Government GRF HKU781709M, HKU 784611M (to W.Y.), and HKU 770411M (to Y.L.L.). L.Z. and J.Z. are partially supported by Edward the Sai Kim Hotung Pediatric Education and Research Fund.

REFERENCES

Adie,E.A. *et al*. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC bioinformatics*, **6**, 55.

DePristo,M.A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*., **43**, 491–498.

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

Liu,X. *et al*. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat*., **32**, 894–899.

Mao,H. *et al*. (2012) Exome sequencing identifies novel compound heterozygous mutations of IL-10 receptor 1 in neonatal-onset Crohn's disease. *Genes Immun*., **13**, 437–442.

Yu,W. *et al*. (2008) A navigator for human genome epidemiology. *Nat. Genet*., **40**, 124–125.

Zhang,L. *et al*. (2011) Homozygosity mapping on a single patient: identification of homozygous regions of recent common ancestry by using population data. *Hum. Mutat*., **32**, 345–353.