Bioinformatics, 32(9), 2016, 1423–1426
doi: 10.1093/bioinformatics/btw079
Advance Access Publication Date: 15 February 2016
Applications Note

Genetics and population analysis

RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data

Xiaowei Zhan[1,2,*], Youna Hu[3], Bingshan Li[4], Goncalo R. Abecasis[5,*] and Dajiang J. Liu[6,7,*]

[1]Department of Clinical Science, Quantitative Biomedical Research Center, [2]Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA, [3]A9.Com Inc, Palo Alto, CA 94301, USA, [4]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37240, USA, [5]Center of Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA, [6]Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA and [7]Institute for Personalized Medicine, Penn State College of Medicine, Hershey, PA 17033, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 2, 2015; revised on December 16, 2015; accepted on February 5, 2016

Abstract

Motivation: Next-generation sequencing technologies have enabled the large-scale assessment of the impact of rare and low-frequency genetic variants for complex human diseases. Gene-level association tests are often performed to analyze rare variants, where multiple rare variants in a gene region are analyzed jointly. Applying gene-level association tests to analyze sequence data often requires integrating multiple heterogeneous sources of information (e.g. annotations, functional prediction scores, allele frequencies, genotypes and phenotypes) to determine the optimal analysis unit and prioritize causal variants. Given the complexity and scale of current sequence datasets and bioinformatics databases, there is a compelling need for more efficient software tools to facilitate these analyses. To answer this challenge, we developed RVTESTS, which implements a broad set of rare variant association statistics and supports the analysis of autosomal and X-linked variants for both unrelated and related individuals. RVTESTS also provides useful companion features for annotating sequence variants, integrating bioinformatics databases, performing data quality control and sample selection. We illustrate the advantages of RVTESTS in functionality and efficiency using the 1000 Genomes Project data.

Availability and implementation: RVTESTS is available on Linux, MacOS and Windows. Source code and executable files can be obtained at https://github.com/zhanxw/rvtests

Contact: zhanxw@gmail.com; goncalo@umich.edu; dajiang.liu@outlook.com

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

1 Introduction

Next-generation sequencing and microarray genotyping have enabled the cost-effective interrogation of a full spectrum of sequence variants. There is considerable interest in understanding the functional role of rare and low-frequency variants in the etiology of complex diseases. Efficient software programs for sequence-based association analysis are in great demand. However, several computational challenges must first be addressed: first, to determine the optimal analysis units and prioritize causal variants in sequence-based association analyses, multiple sources of information need to be integrated, including annotations, functional prediction scores and others. For example, simulation studies show that it is beneficial to analyze non-synonymous variants in a gene-level association test (Kryukov *et al.*, 2009) and that the power can be improved by incorporating functional prediction scores, if the scores are correlated with variant causality (Byrnes *et al.*, 2013). An automatic pipeline is needed to integrate this information and facilitate association analyses. Second, whole-genome datasets of many thousands of individuals often contain tens of millions of variants, which can be >100 GB in size even after compression—orders of magnitude larger than a typical array-based GWAS. Developing efficient tools that can scale for these datasets is a critical yet daunting task.

To address the aforementioned challenges and needs, we have developed and distributed the RVTESTS package for efficient and comprehensive analyses of sequence-based associations. RVTESTS offers a great variety of useful features (Table 1):

(1) A comprehensive set of rare variant association tests were implemented for both autosomal and X-linked variants. These tests range from the simple burden test, which assumes rare variants in the gene region have similar effect sizes (Li and Leal, 2008), to sequence-kernel association tests that can handle scenarios where variants in the same gene region affect the phenotype in opposite directions (Wu *et al.*, 2011), e.g. when both hypermorphic and hypomorphic alleles are present. For a comprehensive description of rare variant association analysis methods, the readers may refer to a recent review (Lee *et al.*, 2014).

RVTESTS can perform linear/logistic regression analyses for unrelated samples, as well as efficient linear mixed model analyses of (cryptically) related individuals (with either empirical or pedigree kinships; Lippert *et al.*, 2011). To permit the more accurate assessment of *P*-values for analyzing rare variant association with discrete traits, it also implements the Firth corrected logistic regression analyses for single variant and burden tests (Firth, 1993; Ma *et al.*, 2013). RVTESTS does not implement methods for fine mapping causal variants. Yet, fine mapping studies can be nicely complemented by tools such as pVAAST (Hu *et al.*, 2014). Summary association statistics generated by RVTESTS can be used by tools such as fGWAS (Pickrell, 2014) to prioritize causal variants.

(2) RVTESTS provides a variety of useful companion features, including (i) efficient variant annotation; (ii) integration of various bioinformatics databases; (iii) data quality control (QC); (iv) selection of samples for association analyses; and (v) generation of summary association statistics for meta-analyses of gene-level association tests (Liu *et al.*, 2014).

Software packages that accompany methodology papers are often available (Hu *et al.*, 2014; Wu *et al.*, 2011; Yandell *et al.*, 2011). While they implement useful methods, they typically do not provide sufficient features that are necessary for practical data analyses. It is necessary to develop software tools and pipelines that enable statistical analysis of sequence data which is otherwise cumbersome to perform. A few software packages are available for performing comprehensive sequence-based association analyses, including Variant Association Tools (VAT; Wang *et al.*, 2014) and PSEQ (https://atgu.mgh.harvard.edu/plinkseq/index.shtml). Despite their convenient and useful features, the tools are slower in managing and performing association analyses for large sequence datasets. Moreover, to our knowledge, neither tool supports the linear mixed model analyses of (cryptically) related samples. Instead, RVTESTS is designed to be able to efficiently handle large datasets in a moderately configured computer server. In our benchmarks, for a computer server with 64 GB RAM, RVTESTS can analyze 64 000 individuals with a linear mixed model or >100 000 individuals using a (generalized) linear model.

We evaluated and compared the features for RVTESTS with VAT and PSEQ using the 1000 Genomes Project Phase 3 dataset. We show that RVTESTS is faster than alternative tools and still offers comprehensive support for sequence-based association tests.

2 Methods

We describe the key feature for RVTESTS for gene-level association tests, as well as companion features for annotating sequence variations, performing data QC and supporting meta-analyses.

2.1 Analysis of rare variant associations

RVTESTS takes standard VCF/BCF (for genotypes) and PED (for phenotypes and covariates) files as input. We implement frequently used rare variant association tests, including burden tests, SKAT and variable threshold tests (and many others). These tests allow for the analyses of both autosomal and X-linked genes. We also extend these methods to the analysis of (cryptically) related individuals using Linear Mixed Models (LMMs) (with both empirical and pedigree kinship). To enable the analysis of large datasets with many thousands of individuals genotyped on millions of markers, we incorporate recently developed efficient algorithms for fitting LMMs (see Supplementary Material). Given that rare variant association analyses may be distinctly confounded by population structures, linear mixed model analysis has been identified as a key approach for the proper control of type I errors in rare variant association

Table 1. Main features for RVTESTS

Functionality	Features
Analyses of unrelated samples	Linear regression analysis for continuous traits Logistic regression analysis for binary traits Firth corrected logistic regression for single variant and burden tests of binary traits Commonly used rare variant tests for autosomal and X chromosome genes (Supplementary Table S1 for details)
Analyses of related samples	Linear mixed model (LMM) analysis using pedigree/empirical kinships Commonly used rare variant tests for autosomal and X chromosome genes (Supplementary Table S1 for details).
Variant annotation	Annotate coding variants using various gene definitions Region-based annotation Incorporate numerous bioinformatics databases
Meta-analysis	Generate summary statistics in RAREMETAL format
Integration with R	Summary association statistic files and annotated VCF files can be randomly accessed by SEQMINER (Zhan and Liu, 2015) in R

analyses (Lippert *et al.*, 2011; Listgarten *et al.*, 2013). LMM analyses supported by RVTESTS can be quite valuable.

All supported rare variant tests can analyze both hard genotype calls and imputed genotype dosages. This feature is particularly useful when genotypes are generated by haplotype-based variant caller or imputation algorithms where variant genotypes are represented by their expected values.

To perform gene-level association analyses, variant annotation information is needed. Our tool provides a streamlined annotation pipeline (see 2.2.1 for details) and can also accept manually curated annotations by users or by other annotation software programs [e.g. snpEff (Cingolani *et al.*, 2012), ANNOVAR (Wang *et al.*, 2010)]. This annotation information can be used to determine the analysis units. In RVTESTS, variant groups may be specified for gene-level association tests. Users can also specify the optimal analysis unit inferred by external programs.

2.2 Companion features

2.2.1 Prepare sequence datasets for association analyses

The RVTESTS package incorporates an efficient and powerful variant annotator. To integrate annotation information and other bioinformatics databases, RVTESTS expands VCF files with annotation information and exports resulting datasets, also in VCF format. The integrated dataset is conceptually similar to a master table in a database project, which links sequence variant genotypes to annotation information. The step of preparing sequence data for association analyses is much faster than VAT and PSEQ (Supplementary Table S2).

2.2.2 Data QC, variant and sample selections

A critical step in genetic association analyses and meta-analyses is to examine the quality of the datasets. RVTESTS implements comprehensive features for checking strand flips and reference/alternative allele swaps, among others. It will automatically generate a variety of per-variant QC metrics, including Hardy Weinberg equilibrium *P*-values, variant call rates and global QC metrics, including the transition/transversion ratio. RVTESTS supports selecting a subset of samples for association analyses as well as filtering variants based on generated QC metrics and genotype qualities.

2.2.3 Generating summary association statistics for meta-analysis

RVTESTS can be used to generate summary statistics for meta-analysis. Result files will be automatically compressed and indexed by tabix, which can then be analyzed by RAREMETAL software (Feng *et al.*, 2014; Liu *et al.*, 2014). Summary association statistics and integrated annotation information may also be used with other fine mapping tools to prioritize causal variants.

2.3 Algorithmic optimizations

A series of algorithmic optimizations further improves the performance of RVTESTS, including the capability of directly reading/writing compressed files, the incorporation of OpenMP for parallel processing as well as the use of ordered arrays for storing bioinformatics databases to speed up annotations and queries. More details are described in Supplementary Material Section 4.

3 Results

We summarize supported features for RVTESTS, VAT and PSEQ in Supplementary Table S1. As stated in Methods, RVTESTS uniquely supports the analyses of related individuals, as well as the analyses of X-linked genes. To our knowledge, both VAT and PSEQ in their current implementations can only analyze unrelated samples.

For the shared features of all three tools (e.g. managing sequence datasets and association analyses of unrelated samples), we compared the computational efficiency for data preparation and association analyses using the 1000 Genomes Project Phase 3 dataset. This dataset consists of 2504 individuals genotyped on 84.8 million markers. We simulated continuous phenotypes based on the null model of no genotype–phenotype associations. To prepare sequence data for association analyses, VAT and PSEQ require building databases, while RVTESTS requires annotating sequence variants. VAT and PSEQ took 495 and 36 CPU hours, respectively, on a desktop with Intel® Xeon® CPU E5-2650 v2 @ 2.60 GHz to prepare and import data into SQLite databases. As a comparison, annotating the whole-genome datasets required only 17 h for RVTESTS. It is clear that RVTESTS offers considerable improvement in time complexity.

Next, we compared the association analyses results on simulated datasets. We performed single variant tests, as well as simple burden, SKAT and VT tests analyzing rare variants with MAF $< 5\%$. RVTESTS attains a good balance of time and memory usage compared with other tools, e.g. completing the simple burden test took RVTESTS/PSEQ/VAT a 15.7/60.2/374.8 h, and required 1.6/10.0/0.8 GB of RAM (Supplementary Table S3).

Finally, we evaluated time and memory complexity for linear mixed model analyses of 2504 individuals—a unique feature of RVTESTS. The analyses using single variant, simple burden, SKAT and VT tests took 17/19/1514/856 CPU hours. On a workstation with 64 GB RAM, our tool can handle datasets with 64 000 genotyped individuals. Given that the LMM analyses have linear complexity with respect to the number of genes and quadratic complexity with respect to the number of samples (after spectral decompositions), it is clear that RVTESTS can handle large datasets genotyped on tens of millions of markers (Supplementary Table S4).

4 Conclusion

In summary, RVTESTS is an efficient and comprehensive tool for sequence-based association analyses. As the cost of sequencing continues to decrease (e.g. a milestone of whole-genome sequencing for $1000 has been recently attained), population-based genetic studies will soon reach another scale. The speed and scalability advantage of RVTESTS will be even more critical when applied to these ultra-large scale datasets. We envision that our program will continue to make valuable contributions to the study of rare genetic variants in complex human diseases.

Funding

This work was supported by National Institutes of Health grants 1R01HG008983-01, 1R21DA040177-01 and 1R01DA037904-01A1 to X.Z. and D.J.L., R01HG006857 to B.L. and R01HG007022 to G.R.A.

References

Byrnes,A.E. *et al.* (2013) The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. *Genet. Epidemiol.*, **37**, 666–674.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Feng, S. *et al*. (2014) RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics*, **30**, 2828–2829.

Firth,D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Hu,H. *et al*. (2014) A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nat. Biotechnol*., **32**, 663–669.

Kryukov,G.V. *et al*. (2009) Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci USA*, **106**, 3871–3876.

Lee,S. *et al*. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet*., **95**, 5–23.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet*., **83**, 311–321.

Lippert,C. *et al*. (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

Listgarten,J. *et al*. (2013) FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat. Genet*., **45**, 470–471.

Liu,D.J. *et al*. (2014) Meta-analysis of gene-level tests for rare variant association. *Nat. Genet*., **46**, 200–204.

Ma,C. *et al*. (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol*., **37**, 539–550.

Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet*., **94**, 559–573.

Wang,G.T. *et al*. (2014) Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. *Am. J. Hum. Genet*., **94**, 770–783.

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

Wu,M.C. *et al*. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet*., **89**, 82–93.

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

Zhan, X. and Liu,D.J. (2015) SEQMINER: an R-package to facilitate the functional interpretation of sequence-based associations. *Genet. Epidemiol*., **39**, 619–623.