Bioinformatics, 34(16), 2018, 2851–2853
doi: 10.1093/bioinformatics/bty180
Advance Access Publication Date: 27 March 2018
Applications Note

Genetics and population analysis

ONETOOL for the analysis of family-based big data

Yeunjoo E. Song[1,†], Sungyoung Lee[2,†], Kyungtaek Park[2], Robert C. Elston[1], Hyeon-Jong Yang[3,4,*] and Sungho Won[2,5,6,*]

[1]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA, [2]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, Korea, [3]SCH Biomedical Informatics Research Unit, Soonchunhyang University Hospital, Seoul, Korea, [4]Department of Pediatrics, Soonchunhyang University Hospital, Soonchunhyang University College of Medicine, Seoul, Korea, [5]Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul, Korea and [6]Institute of Health and Environment, Seoul National University, Seoul, Korea

*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.

Associate Editor: Oliver Stegle

Received on July 1, 2017; revised on February 26, 2018; editorial decision on March 22, 2018; accepted on March 26, 2018

Abstract

Motivation: Despite the need for separate tools to analyze family-based data, there are only a handful of tools optimized for family-based big data compared to the number of tools available for analyzing population-based data.

Results: ONETOOL implements the properties of well-known existing family data analysis tools and recently developed methods in a computationally efficient manner, and so is suitable for analyzing the vast amount of variant data available from sequencing family members, providing a rich choice of analysis methods for big data on families.

Availability and implementation: ONETOOL is freely available from http://healthstat.snu.ac.kr/software/onetool/.

Contact: won1@snu.ac.kr or pedyang@schmc.ac.kr

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

1 Introduction

The importance of family-based designs has been repeatedly stressed for analyses with sequence data because of the genetic homogeneity between family members (Bailey-Wilson and Wilson, 2011; Wijsman, 2012). Family study designs provide not only the enrichment of genetic loci containing rare variants, but also methods to control for genetic heterogeneity and population stratification.

Family-based data have different properties from population-based data owing to the genetic relatedness among family members and Mendelian transmission. These well-known features have allowed family-based designs to play a key role in the history of genetic analysis, but they also limit the use of many available tools designed for the analysis of population-based data. Despite the need for separate tools to analyze family-based data, there is only a handful of tools available for family-based data, especially the big data from sequencing which comprise the readily available form of data for genetic and genomic analyses these days.

As for the population-based data analysis tools, the most common family-based big data analysis tools aim to filter/rank/QC/annotate variants (Supplementary Table 1). These tools are optimized to be used with the vast amount of variant data from sequencing but lack the choices of essential analyses needed to test and infer the valid results regarding the relation between traits of interest against the common and/or rare variants. Therefore, users need to use the separate tools individually. And, though there exists a handful of family-based imputation tools, there is a clear lack of family-based association analysis tools that can analyze the dosage files directly as an input.

Table 1. Summary of available family-based analyses in ONETOOL

Main	Sub-category	Detail
InfoQC analysis	Variant information	F_{ST}, Ts/Tv ratio, MAF, HWE, PCA
	Sample information	Het, Het/Hom
	Pedigree information	Description and summary, plot, relative pairs
	Error detection	Mendelian error
	Relatedness matrix	Kinship, IBS, GRM
Trait analysis	Familial aggregation	Correlation
	Heritability	Based on Kinship, IBS, GRM
	Segregation analysis	Mode of inheritance
Linkage analysis	Model-based	Two-point, utilizing segregation analysis
	Model-free	Multipoint, modeling LD
Association analysis	Single variant	Score test, TDT/SDT, MQLS, FQLS, EQLS, GEMMA
	Gene-based	Collapsing, PEDCMC, FAMVT, FARVAT, FBSKAT, PEDGENE, RVTDT
	Genotype probability & dosage data	Scoretest, EQLS, GEMMA, CMC, PEDCMC, FAMVT, FARVAT, PEDGENE

There are many very well-known family data analysis tools available from the era of linkage analysis and genome-wide association studies (GWAS). Among these, S.A.G.E. 6.4 (http://darwin.cwru.edu/sage/) and Merlin (Abecasis *et al.*, 2002) are still used a lot by many researchers. PLINK (Purcell *et al.*, 2007) is one of most popularly used tool for GWAS. It is a part of many standard pipeline tools, therefore, the PLINK input file formats are the standard format for many sequence data analysis tools. However, since it is designed mainly for population-based case-control data, the analysis options are very limited for use with family data.

All three tools have their pros and cons. In this work, we introduce a novel comprehensive tool that combines the good features of these existing tools and many newly developed family-based association analysis methods along with a novel feature to analyze the dosage data, providing in a computationally efficient manner a rich choice of analysis options to use for the vast amount of variant data coming from the sequencing of families. This provides a convenience and time-saver that enables a researcher to perform many of the family-based genetic and genomic analyses using one tool, i.e. ONETOOL, instead of hopping among many different tools to accomplish a family data analysis project.

2 Features

ONETOOL provides four main analyses: informatics and quality control (InfoQC), trait analysis, linkage analysis and association analysis with both genotype data and dosage data (in Table 1 and also see the Supplementary Table 2 for details).

2.1 InfoQC, trait and linkage analysis

Family data requires additional error checking and filtering that also consider the family structure, so that the family structures are maintained. ONETOOL provides the proper methods to deal this complexity of family data and the downstream analyses as an integrated tool. Moreover, ONETOOL's options for the variant-wise InfoQC and filtering are similar with those in Plink, but they are implemented in a computationally optimized way providing more speed and efficiency. It also provides visualization of family data as done utilizing the R package *kinship2* to generate a plot (Sinnwell *et al.*, 2014).

As shown in Supplementary Table 1, not many tools are available for trait analysis nor are optimized to work with the current pipeline of family big data. ONETOOL fulfills this gap by integrating the tools for familial aggregation or correlation, narrow-sense heritability estimation and segregation analysis.

With ONETOOL, both types of linkage analyses, model-based linkage and model-free linkage accounting for linkage disequilibrium, can be done directly with the current genomic data files.

2.2 Association analysis

Depending on the types of trait data (binary or continuous), family data (random or ascertained, trio or general) and variant data (common or rare) in hand, the different family-based association analyses provide the best estimates in terms of both power and type 1 error. Many times, a complex disease data analysis project involves not just a phenotype but a set of multiple phenotypes with different characteristics. It also involves analyzing a set of different types of genetic data.

By combining many different types of association methods developed for specific cases into an integrated tool with a common interface, ONETOOL enables more seemingly harmonized family-based association analyses. In Supplementary Tables 3 and 4, we summarized the proper timing to use for the family-based association test available in ONETOOL.

2.3 Imputation and dosage data

ONETOOL provides an option to impute the missing genotypes. Expected missing genotypes for typed variants are imputed based on the familial relationship, and if phenotypes of any subjects with missing genotypes are available, genotypes imputed with family members' genotypes can improve statistical power (see Supplementary Material for details).

ONETOOL also enables the family-based association analysis with dosage data and genotype probability. See the Supplementary Table 5 for the supported dosage and genotype probability file formats from several popular imputation tools.

3 Discussion

ONETOOL enables a researcher to perform many of the family-based genetic and genomic analyses in a computationally efficient manner. It provides a convenience and time-saver with a rich choice of analysis options available, both existing and novel. ONETOOL supports various types of data input files includes the dosage and genotype probability files from several imputation tools. Using two different family datasets, we show, in Table 2, the performance of ONETOOL and the time savings by running several analyses at

Table 2. Efficiency of the integrated analyses in ONETOOL

Analyses	Run type	Data1	Data2
InfoQC+Trait	separate run	2.21s	55.83s
	ONETOOL	0.74s	44.09s
InfoQC+Trait+single-variant	separate run	2.41s	58.74s
	ONETOOL	0.83s	54.09s
InfoQC+Trait+gene-based	separate run	2.47s	59.20s
	ONETOOL	0.85s	54.76s

once compare to the separate run for each component (see Supplementary Material for details).

Funding

This work was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant no. HC15C1302); and the Bio-Synergy Research Project (NRF-2017M3A9C4065964) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

Conflict of Interest: none declared.

References

Abecasis,G.R. *et al*. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow tree. *Nat. Genet*., **30**, 97–101.

Bailey-Wilson,J.E. and Wilson,A.F. (2011) Linkage analysis in the next generation sequencing era. *Hum. Hered*., **72**, 228–236.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Sinnwell,J.P. *et al*. (2014) The kinship2 R package for pedigree data. *Hum. Hered*., **78**, 91–93.

Wijsman,E.M. (2012) The role of large pedigrees in an era of high-throughput sequencing. *Hum. Genet*., **131**, 1555–1563.