BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 3 2014, pages 442–444
doi:10.1093/bioinformatics/btt682

Genetics and population analysis

Advance Access publication December 12, 2013

PhenoMan: phenotypic data exploration, selection, management and quality control for association studies of rare and common variants

Biao Li, Gao Wang and Suzanne M. Leal*

Center for Statistical Genetics, Department of Molecular and Human Genetics, One Baylor Plaza 700D, Baylor College of Medicine, Houston, TX 77030, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Next-generation sequencing and other high-throughput technology advances have promoted great interest in detecting associations between complex traits and genetic variants. Phenotype selection, quality control (QC) and control of confounders are crucial and can have a great impact on the ability to detect associations. Although there are programs to perform association analyses, e.g. PLINK and GenABEL, they cannot be used for comprehensive management and QC of phenotype data. To address this need PhenoMan was developed: to select individuals based on multiple phenotype criteria or population membership; control for missing covariate data; remove related individuals, duplicate samples and individuals with incorrect sex specification; recode primary traits and covariates; transform data; remove or winsorize outliers; select covariates for analysis; and create residuals. To ensure consistency and harmonization between analyses, a report is generated for every dataset. Summary statistics are also provided in graphical or text format. PhenoMan can be used for selection and manipulation of quantitative, disease and control data.

Summary: Phenoman is freeware that provides approaches for efficient exploration and management of phenotype data. Proper QC of phenotypes before proceeding to the association analysis is critical to ensure control of type I and II errors, reliable effect estimates and consistent results between studies. PhenoMan is highly beneficial for the preparation of qualitative and quantitative trait data for association studies using new datasets as well as those obtained from public repositories.

Availability and implementation: code.google.com/p/phenoman

Contact: sleal@bcm.edu

Received on August 27, 2013; revised on November 12, 2013; accepted on November 15, 2013

1 INTRODUCTION

In recent years, next-generation sequencing technologies have generated vast amounts of exome and genome sequence data to be used for the detection of complex trait associations. Although association analyses play a crucial role in elucidating the genetic etiology of complex traits, type I error can be increased and power may be profoundly weakened by phenotypic outliers, related individuals, missing phenotypic values, population stratification, inclusion or exclusion of covariates to control for confounders and so forth (de Bakker *et al.*, 2008). Additionally, often between studies analysis of phenotypes are not consistent, e.g. phenotype selection, inclusion of covariates, treatment of outliers. It is important that all analysis steps e.g. determining which study subjects and covariates to include in the analysis are fully documented, so results are reproducible and additionally analyses are consistent between datasets, which is a key for both meta-analysis and replication studies (Chang *et al.*, 2006). Although programs such as PLINK (Purcell *et al.*, 2007) and GenABEL (Aulchenko *et al.*, 2007) can be used for association analysis, they cannot be used for comprehensive management and quality control (QC) of phenotype data. Additionally, R (www.r-project.org) could be used for data exploration, but it is not readily suitable to perform all necessary tasks to prepare phenotype data for association analysis; additionally the user must have expertise in writing R scripts. The PhenoMan software was developed to provide a universally powerful framework with simple commands, to integrate both quantitative and qualitative phenotype data exploration, QC and sample selection into a unified program. PhenoMan, an interactive command-line driven program, is written in Python and R.

2 DESCRIPTION

2.1 Sample selection

PhenoMan can select study subjects not only based on a single trait but also using any phenotypic information of interest. For both case-control and quantitative traits, individuals can be selected according to multiple qualitative and quantitative phenotypic attributes and inclusion and exclusion criteria, e.g. type 2 diabetes status, age, body mass index, lipid profiles and so forth. These criteria can also be used to define case-control status. Selection can also be made using extremes for one or more quantitative traits.

2.2 Removal of individuals

If genotype or sequence data are available, samples can be selected for removal if their reported sex is inconsistent with genetic data, they are a cryptic duplicate or related to another study subject or they belong to a population other than the one understudy. Genetic association analyses are usually performed

*To whom correspondence should be addressed.

separately for each population, e.g. Europeans and Asians. Although self-reported race information is usually available, these data are not always reliable. PhenoMan uses components from multidimensionality scaling to determine population membership and can eliminate those individuals who belong to other ethnic groups, e.g. for a study of African-Americans remove European-American individuals.

2.3 Recoding of variables

PhenoMan can be used to recode phenotype information, e.g. to dichotomize quantitative traits/covariates and to perform dummy coding of covariates that are not ordinal, e.g. medication usage. Also information from multiple covariates can be combined, e.g. waist and hip measurements can be used to determine hip-to-waist ratios. Additionally missing values can be substituted with the average trait value.

2.4 Transformation of quantitative traits and handling of outliers

PhenoMan can perform various types of data transformation on quantitative traits, including log, scaling, standardization, normalization and Gaussian quantile normalization. PhenoMan can be used to detect quantitative trait outliers by visualization of the density and with statistics such as range, mean, median and standard deviation. Identified outliers can be either removed or winsorized. Usually values that are not biologically feasible are removed because they are likely an error, whereas all other outliers are winsorized.

2.5 Selection of covariates

Potential confounding factors should be controlled for in the analysis (Stitziel *et al*., 2011). PhenoMan can aid in the selection of covariates, which are potential confounders by performing either forward or backward selection or using Akaike information criterion. For quantitative and case-control data, these covariates can be included in the regression model. For quantitative traits, it is also possible to use PhenoMan to obtain residuals that can then be analyzed.

2.6 Data summary and outputs

PhenoMan generates a new phenotype file containing only those individuals that were selected for analysis and the desired covariates. PhenoMan also generates a summary file that contains statistics on the selected traits and covariates, e.g. mean, median and standard deviation, number of females, percent of missing values. This summary information can also be generated as graphical output.

2.7 PhenoMan workflow

To prepare phenotypic data for analysis, we suggest using the workflow shown in Figure 1.

2.8 Software overview/functionality

PhenoMan consists of a large variety of modules to handle most phenotype manipulation, testing and cleaning related tasks. PhenoMan uses and outputs standard format text files that are also used by dbGaP (Mailman *et al*., 2007) and PLINK. On completion of the phenotype data exploration, selection and QC, PhenoMan can create a checklist/workflow of all critical decision points by saving executed commands and arguments. This checklist can be used to reproduce results and also to ensure analysis is comparable between studies. Additionally, for ease of use relevant commands can be extracted for the analysis of additional datasets.

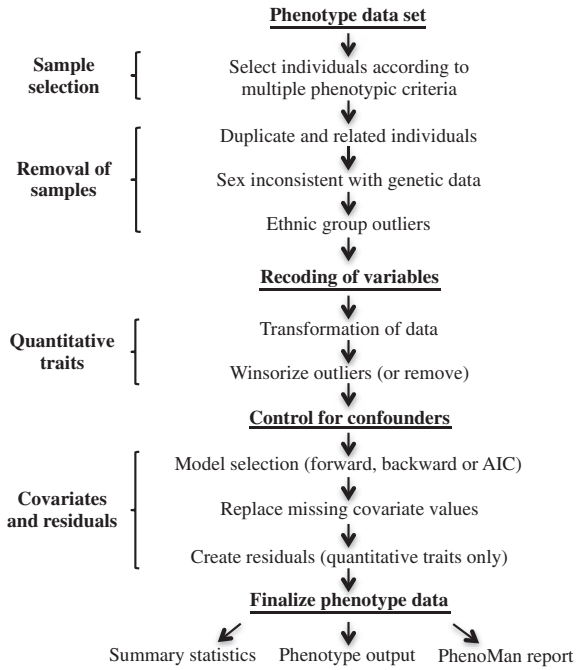

Fig. 1. Schematic workflow for PhenoMan

3 DISCUSSION

For association studies, QC can be time-consuming, but it is necessary before the onset of data analysis. Proper QC of phenotypes before proceeding to association analysis is critical to ensure control of type I and II errors, reliable effect estimates and consistent results between studies. PhenoMan is designed to provide a convenient and comparable manner to obtain analysis-ready phenotype data. Performing this important process contains two steps: (i) exploration—use statistics and graphics to overview the data and (ii) solutions—correct/transform, select and/or remove identified 'erroneous' entities. Ultimately, we expect that PhenoMan will enhance the analysis of association data.

Funding: National Institute of Health (grants numbers HL102926 and MD005964).

Conflict of Interest: none declared.

REFERENCES

Aulchenko,Y.S. *et al*. (2007) GenABEL: an R library for genome-wide association analysis. *Bininformatics*, **23**, 1294–1295.

Chang,Y-P.C. *et al.* (2006) The impact of data quality on the identification of complex disease genes: experience from the family blood pressure program. *Euro. J. Hum. Genet.*, **14**, 469–477.

de Bakker,P.I.W. *et al.* (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*, **17**, 122–128.

Mailman,M.D. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.

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.

Stitziel,N.O. *et al.* (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.*, **12**, 227.