PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data

Janis E. Wigginton* and Gonçalo R. Abecasis

Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48103, USA

Received on April 28, 2005; revised on June 5, 2005; accepted on June 6, 2005
Advance Access publication June 9, 2005

ABSTRACT

Summary: We describe a tool that produces summary statistics and basic quality assessments for gene-mapping data, accommodating either pedigree or case-control datasets. Our tool can also produce graphic output in the PDF format.

Availability: http://www.sph.umich.edu/csg/abecasis/Pedstats/
download/

Contact: wiggie@umich.edu

Supplementary information: http://www.sph.umich.edu/csg/abecasis/Pedstats/

A crucial first step in the analysis of gene mapping data is the careful description of the available data, including, for example, genotyping completeness and heterozygosities for genetic markers, and distributions and familial correlations for quantitative traits. Although a number of programs now provide some facilities for data checking or summary (Mukhopadhyay et al., 2005; Lange et al., 1988; Elston et al., 2004; O’Connell et al., 1998) complete screening and summary of genetic data frequently involves the use of multiple programs and/or in-house tools. As the scale of the datasets available for analysis increases, this process can become particularly challenging. For example, with the advent of high-throughput single nucleotide polymorphism genotyping technologies, datasets will soon be available that includes genotypes for hundreds of thousands or millions of markers for each individual. In addition, with the focus on uncovering the genetic basis of complex disease, it is likely that collaborative projects will collect samples with hundreds or thousands of phenotypes each measured on thousands of individuals. We have developed PEDSTATS, a freely available utility, for summarizing salient features and performing basic quality checks on gene-mapping data. Our utility can conveniently handle these very large datasets and here we summarize its main features.

PEDSTATS runs on any platform where a modern C++ compiler is available, including those based on the Linux, UNIX, Windows and Mac OS X operating systems. It is a command-line utility that can produce both text output to the console and graphical output to a PDF file. Its major capabilities can be grouped into four areas: (1) checks of input formats and pedigree consistency, (2) checks and descriptions of genetic marker data, (3) checks and descriptions of quantitative traits and covariates and (4) descriptions of discrete traits. We describe each of these in turn below.

*To whom correspondence should be addressed.
males and females can be generated (as illustrated in Fig. 1, Panel C which summarizes the distribution of ‘Height’ in one large dataset). These can be helpful in detecting outliers as well as detecting deviations from approximate normality, which is important for many quantitative trait analyses (Allison et al., 1999). Optionally, correlations for other relative pair types can be calculated and plotted (as illustrated in Fig. 1, Panel D, which summarizes the correlations between ‘Weight’ for different relative pairs) and stratified by sex, if desired. Correlations between relatives can provide information about the overall impact of genes on a particular trait. In the example, it is clear that correlation of the variable ‘Weight’ for first degree relatives (in this case, parent–offspring and sibling pairs) is higher than for more distant relatives (half-sibling, avuncular, grand-parent grand-child and cousin pairs). When an age variable is present, we have implemented checks to ensure that values recorded for each individual are compatible with those of their ancestors, subject to user-specified minimum and maximum generation times.

Finally, for discrete traits, PEDSTATS reports the proportion of phenotyped individuals and provides a breakdown of affected individuals. A summary of affected, unaffected and discordant pairs can also be produced, and may help guide decisions on whether a dataset contains sufficient information for an affected relative pair analysis to be carried out (Risch, 1990; Whittemore and Halpern, 1994). As with the other analysis options, discrete trait reports can be segregated by sex.

In addition to the ability to report statistics separately for different relative pairs and segregate results by sex, PEDSTATS can produce reports for individual families and allows various filters to be applied to input data prior to analysis. For example, all analyses can be restricted to affected individuals (for a specific trait) or to individuals with a minimal amount of genotype data.

We hope our tool will prove valuable to scientists hoping to discern important features of their data, and ease the burdensome task of verifying the consistency and integrity of input formats. Executables, source code and a web-based tutorial that explains input file format,
implementation details and output for various tests are available from our website.

ACKNOWLEDGEMENTS
This work was supported by research grants from the National Human Genome Research Institute and the National Eye Institute.

Conflict of Interest: none declared.

REFERENCES