BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 16 2005, pages 3445–3447
doi:10.1093/bioinformatics/bti529

Genetics and population analysis

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.

The first step in any analysis is the validation of input files. At this stage, common data-format errors such as missing or extraneous columns are reported. Next, the reported family structures are validated to ensure that all connecting individuals are present and that sex-codes are consistent for the various individuals. If desired, large pedigrees can be trimmed to remove uninformative individuals with no phenotype or genotype data, or separated into disconnected family units. A brief summary of the number of pedigrees, individuals and a distribution of individuals per family is produced. This information can be graphically summarized (Fig. 1A is an example summarizing the distribution of family sizes in one large dataset) and, optionally, includes counts for various types of relative pairs which can be further broken down by sex. Individuals with no phenotype or genotype information can be automatically removed and a new set of input files generated. PEDSTATS readily accepts files prepared for other packages we have developed, including those prepared for linkage analyses with Merlin (Abecasis *et al*., 2002), association analyses with QTDT (Abecasis *et al*., 2000) and relationship inference with GRR (Abecasis *et al*., 2001). Other popular formats, such as those used by the LINKAGE package (Lathrop *et al*., 1985) and by MENDEL and related tools (Lange *et al*., 1988) are also accommodated.

When verifying genetic marker data, PEDSTATS reports basic statistics like heterozygosity and genotyping completeness and can produce graphical summaries of allele and genotype frequencies. After automatic grouping of rare alleles, conformance of observed genotypes with Hardy–Weinberg equilibrium can be checked with a χ^2 test for multi-allelic markers or an exact test for bi-allelic markers (Wigginton *et al*., 2005). Results of Hardy–Weinberg tests, including an exact distribution for the number of heterozygotes in the sample, can be presented graphically (e.g. Fig. 1B). Mendelian inheritance checks for both autosomal and X-linked marker data are also carried out using a genotype elimination algorithm that finds all inconsistencies in pedigrees without loops (Lange and Goradia, 1987; O'Connell and Weeks, 1999). Verifying Mendelian consistency prior to analysis of genetic marker data can be a crucial step (Lange and Goradia, 1987; O'Connell and Weeks, 1998), since most genetic analysis programs do not model genotyping error explicitly (for an exception, see Sobel *et al*., 2002).

For quantitative traits and covariates, PEDSTATS reports the range, mean and variance of the trait distribution along with the correlation between siblings. Several graphics, including histograms of the overall trait distribution and comparisons of distributions between

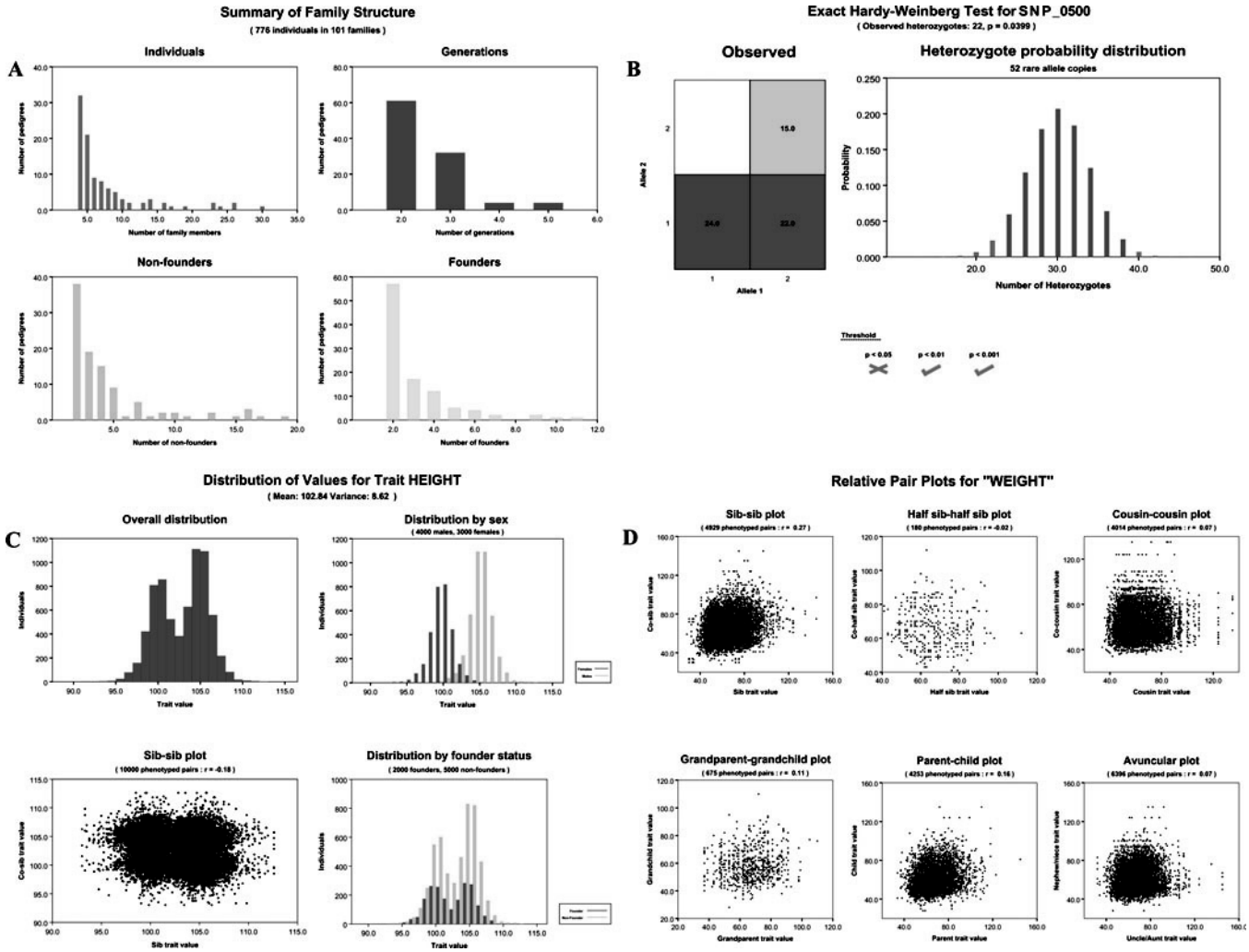

Fig. 1. Examples of available graphical output. (**A**) provides information on the distribution of family sizes; (**B**) summarizes the observed genotype distribution and the exact distribution of heterozygotes conditional on observed allele counts; (**C**) provides information on the distribution of a quantitative trait; and (**D**) summarizes relative pair correlations. More detailed descriptions and examples are available on our website.

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

Abecasis,G.R. *et al*. (2000) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.*, **66**, 279–292.

Abecasis,G.R. *et al*. (2001) GRR: graphical representation of relationship errors. *Bioinformatics*, **17**, 742–743.

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

Allison,D.B. *et al*. (1999) Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.*, **65**, 531–544.

Elston,R., Bailey-Wilson,J., Bonney,G., Tran,L., Keats,B. and Wilson,A. (2004) SAGE Statistical Analysis for Genetic Epidemiology, Version 5.0.

Lange,K. and Goradia,T.M. (1987) An algorithm for automatic genotype elimination. *Am. J. Hum. Genet.*, **40**, 250–256.

Lange,K. *et al*. (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet. Epidemiol.*, **5**, 471–472.

Lathrop,G.M. *et al*. (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.*, **37**, 482–498.

Mukhopadhyay,N. *et al*. (2005) Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics*, **21**, 2556–2557.

O'Connell,J.R. and Weeks,D.E. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, **63**, 259–266.

O'Connell,J.R. and Weeks,D.E. (1999) An optimal algorithm for automatic genotype elimination. *Am. J. Hum. Genet.*, **65**, 1733–1740.

Risch,N. (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.*, **46**, 229–241.

Sobel,E. *et al*. (2002) Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.*, **70**, 496–508.

Whittemore,A.S. and Halpern,J. (1994) A class of tests for linkage using affected pedigree members. *Biometrics*, **50**, 118–127.

Wigginton,J.E. *et al*. (2005) A note on exact tests of Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.*, **76**, 887–893.