Report

PowerTrim: An Automated Decision Support Algorithm for Preprocessing Family-Based Genetic Data

Tricia A. Thornton[1,2] and Jonathan L. Haines[2]

[1]Neuroscience Graduate Program, Vanderbilt Brain Institute, Department of Biomedical Informatics, and [2]Program in Human Genetics, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Vanderbilt University, Nashville

Statistical genetics software packages for linkage analysis have their own unique constraints on the size and shape of the pedigrees they can process. As a result, researchers are often forced to exclude from analysis some individuals in a given family. Existing procedures for reducing pedigree size to fit computational constraints use arbitrary rules and are not interactive. However, judicious evaluation of which subject(s) to remove to minimize loss of information involves consideration of many factors, including informativeness owing to position in pedigree, availability of genotypic information, and quality of phenotypic information. Thus, automation of this task would be of significant benefit. We designed an interactive algorithm (PowerTrim) that provides the user access to detailed information with which to make informed decisions. In addition, PowerTrim checks for transcriptional and data-entry errors, which can be very time-consuming to localize manually.

Statistical genetics software packages for linkage analysis have their own unique constraints on the size and shape of the pedigrees they can process (Goedken et al. 2000; Ott 2000). As a result, researchers are often forced to exclude from analysis some individuals in a given family. Existing procedures for reducing pedigree size to fit computational constraints use arbitrary rules and are not interactive. However, judicious evaluation of which subject(s) to remove involves consideration of many factors, including informativeness owing to position in pedigree, availability of genotypic information, and quality of phenotypic information. Thus, automation of this task would be of significant benefit.

We designed an interactive algorithm called "PowerTrim," developed in Perl, with four goals in mind. The first goal is to identify and report preprocessing errors in pedigree data in a user-friendly manner. The following is a list of preprocessing errors that PowerTrim identifies in its initial data evaluation and writes to an error file:

1. Individuals with only one parent identified;
2. Females referenced as a father and males referenced as a mother;
3. Invalid sex identifier;
4. Duplicate individuals;
5. Unexpected missing data; and
6. Individuals who are not connected to the respective proband.

The second goal is to "trim" from the pedigree those individuals who provide no information at all to a linkage analysis of specified markers. Following are the conditions under which individuals or whole families are considered noninformative by PowerTrim:

1. Families consisting of only a parent and a child (User may decline to trim such a family.);
2. Individuals who have no parents and no children;
3. Individuals who are unaffected and ungenotyped;
4. Individuals who are ungenotyped, have no genotyped descendants, and have either an affected parent or an affected sibling;
5. Founder individuals who are unaffected and ungenotyped and have only one child; and
6. Individuals or groups of individuals within a family who are unconnected to the proband. (User may remove or make of them a new family.)

The third goal of PowerTrim is to recommend how a pedigree might be further trimmed to maximize the power of the linkage analysis while conforming to the constraints

Received December 11, 2002; accepted for publication January 28, 2003; electronically published April 3, 2003.

Address for correspondence and reprints: Dr. Jonathan L. Haines, Program in Human Genetics, Vanderbilt University Medical Center, 519 Light Hall, Nashville, TN 37232-0700. E-mail: jonathan@phg.mc.vanderbilt.edu

0002-9297/2003/7205-0020$15.00

of the target software. To this end, PowerTrim iteratively examines the relative contribution of each individual within each family to the power of the linkage analysis, using information on data completeness and familial relation. For example, it reports the number and percentage of markers genotyped by individual and by family and allows batch trimming of individuals within some or all families who fall below a user-specified threshold. Following are the types of data reported on command to the user, for help in deciding who should be removed next from a given pedigree:

1. Incompatibility with GENEHUNTER-Plus, Allegro, and Merlin (families that exceed bit limits);
2. Number of affected individuals by family;
3. Number and percent of markers genotyped by individual (for all individuals or only those from families above the user-specified bit size);
4. Number and percent of individuals with ⩾1 marker genotyped by family; and
5. Distribution of individuals within each family by percent of markers genotyped.

PowerTrim offers the following options for pedigree trimming:

1. User specifies an individual or family for removal.
2. User specifies a lower limit for the percent of markers genotyped per individual, such that all individuals falling below that limit are removed.
3. User specifies both a lower limit (as in option 2) and one or more families in which to apply that limit, such that all individuals in those families falling below that limit are removed.

PowerTrim also provides interactive decision support, offering recommendations on the basis of known data constraints of the programs GENEHUNTER-Plus (Kong 1997), Allegro (Gudbjartsson 2000), and Merlin (Abecasis 2002). Each of these programs sets different limits on the number of bits the pedigree may contain, recognizing that the space and time complexity of the analysis can increase exponentially with the size of the pedigree. Existing trim options in these programs are not interactive or user directed, and they make arbitrary choices. PowerTrim calculates the pedigree bit size and then allows the user to choose which individuals to trim, while updating and reporting on which pedigrees continue to exceed the limits of one or more of the analysis programs.

The fourth goal of PowerTrim is to automate the export of data into the standard linkage .ped format and to produce a summary file with descriptive statistics of the trimmed pedigrees. If changes are made to any pedigree in the original .ped file, the user is prompted to enter a new file name, to which the trimmed pedigree data is written without disturbance of the original file. In addition, all activity of the PowerTrim program during any given session is written to a log file, should further examination be desired.

PowerTrim is a very useful tool for one piece of the genetic data analysis puzzle. PowerTrim aids users in quickly identifying preprocessing errors and in trimming uninformative individuals, potentially reducing the time required to run data analysis programs. What is most important, however, is that users can reduce the amount of time spent manually trimming pedigrees that cannot be processed "as is" in GENEHUNTER-Plus, Allegro, or Merlin. This is done while still maximizing the power of the analysis.

PowerTrim is made freely available to academic users. Please visit the software page of the Program in Human Genetics Web site for more information.

Electronic-Database Information

The URL for data presented herein is as follows:

Program in Human Genetics, http://phg.mc.vanderbilt.edu/ (for PowerTrim)

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Goedken R, Crowe R, Deng Z, Fyer AJ, Haghighi V, Heiman G, Hodge SE, Knowles JA, Vieland VJ, Wang K, Weissman MM (2000) Drawbacks of GENEHUNTER for larger pedigrees: application to panic disorder. Am J Med Genet 96: 781–783

Gudbjartsson DF, Jonasson K, Frigge M, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12–13

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188

Ott J, Hoh J (2000) Statistical approaches to gene mapping. Am J Hum Genet 67:289–294