SNPduo user guide

Background

SNPduo[1] is designed to aid in the discovery of undocumented/misdocumented familial relationships, identical samples, sample mislabeling, and population structure in SNP genotyping datasets. For compatibility with existing software and data formats, snpduo is designed to accept many of the same formats as PLINK[2]. The software supports data with many individuals (thousands in a single experiment) and many markers (1M+). SNPduo summarizes the identity by state (IBS) of alleles between the pairwise comparisons of all samples in a dataset (Table 1). Summaries of the IBS between individuals are useful because they do not rely on calculations of identity by descent (IBD) probability, are computationally tractable, and require no prior familial relationship information. An additional IBS state described by the tool is "IBS2*" (eye-bee-ess two star), where two heterozygous calls are compared to each other.

If you find this software useful, please cite the original article. Submit any questions or bug information to Jonathan Pevsner, PhD at pevsner@kennedykrieger.org. If you find bugs in SNPduo, please send the output log, a description of the problem,

---

[1] Roberson EDO and Pevsner. Visualization of shared genomic regions and meiotic recombination in high-density SNP data. *PLoS One*. 2009;4:e6711. doi: 10.1371/journal.pone.0006711.

[2] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

your operating system, compiler options, and a small dataset capable of reproducing the bug (if possible) along with the email.

Table 1 – Summary of possible IBS states

Individual 1	Individual 2	IBS
AA	AA	IBS2
AB	AA	IBS1
BB	AA	IBS0
NC	AA	NA
AA	AB	IBS1
AB	AB	IBS2/IBS2*
BB	AB	IBS1
NC	AB	NA
AA	BB	IBS0
AB	BB	IBS1
BB	BB	IBS2
NC	BB	NA
AA	NC	NA
AB	NC	NA
BB	NC	NA
NC	NC	NA

Installation

LINUX

The snpduo program is distributed as a compressed archive of source code files. To install, extract the files into an empty directory. Make any desired modifications to the CFLAGS parameter of the Makefile. For example, add the flag "-m64" to make the executable 64-bit compatible, or add the appropriate "-march" or "-mtune" parameters to add in platform specific optimization. After making any additions to the Makefile, run the command "make". The make program will run the appropriate compilation and linking commands, generating an executable named "snpduo". Copy the executable to your search path (usually /usr/bin or /usr/local/bin) and snpduo is ready to use.

WINDOWS

Windows compilation in a 32-bit environment requires an appropriate compiler, typically a GNU compiler like g++. Windows compilation typically must be performed manually. Extract the source files into a directory. Navigate to the directory on the command-line. As long as g++ or another appropriate C++ compiler is in your path the program can be compiled using "g++ -o snpduo *.cpp". Any desired compilation flags must be added as well, such as "g++ -O3 -o snpduo *.cpp". The output will be an executable named snpduo. It must be copied to a folder in your system path to be available from anywhere on the command-line.

Input Formats

There are multiple input formats that SNPduo can read. All formats should be white-space delimited (space, tab, or mixed) and end in newline '\n' characters. Lines ending only in '\r' WILL NOT read properly, though '\r\n' will. This may require altering file format on Macs.

PED/MAP

The ped/map format requires two separate files: a map file with information for each marker on a separate line, and a ped file with information about each individual on a line. The order of markers in the ped file and in the map file MUST be the same.

The .map file contains the information about the location of markers in the dataset. The required columns are chromosome (1-26, X,Y,M acceptable), RSID or SNP ID, centiMorgan location, and physical location. Only the chromosome and position are really used in SNPduo. The other columns are imported for PLINK compatibility. The

map file can contain only the chromosome, RSID and physical position if the "--map3" switch is specified at run time.

The .ped file contains information on individuals in the study, their relationships, and their markers. The required fields are family id, individual id, mother id, father id, sex, and disease status. Absent father and mother information should each be reported as 0 (and thus you shouldn't use 0 as an individual ID). Sex should be specified as 0=unknown, 1=male, 2=female. Disease status should be specified as an integer as well. After the disease status should be the genotypes. Individual SNP alleles can be separated by white space or not, i.e. an "AA" genotype can be represented as "AA" or "A A". Uncalled markers have the genotype "00" (zero-zero), and should be coded as either "00" or "0 0" depending on the overall file format. NC, NoCall, or other variations will not work.

TFAM/TPED

The tfam and tped files are slightly different, transposed files that contain the same information that would be contained in a map/ped file combination. A tfam file contains individual information on familial relationships and disease information where one line represents the information for one individual. A tped file combines the map information and marker information.

A tfam file contains the first six columns of a ped file (family id, individual id, father id, mother id, sex, disease status). The same coding requirements apply to the tfam file as did to the ped file. One individual should be included per line.

The first four columns of a tped file are the same as a map file (chromosome, rsid, cM position, physical position). The following columns are whitespace separated

genotypes. Whitespace consists of tabs and spaces in this context. Individual alleles can be grouped into a genotype or separated by whitespace also. The information for one marker for all individuals is on one line. The cM position can again be left out if "--map3" is specified. The order of individual genotypes in the tped file MUST be in the same order as the individuals in the tfam file.

Usage

DATA INPUT AND RECODING

SNPduo can read both the ped/map and tped/tped. Ped/map files with the same basename are specified for input as "--file filenamebase". This will automatically open both the "filenamebase.map" and "filenamebase.ped". The ped and mapfiles can be specified separately using "--ped pedfilename.ped --map mapfilename.map".

Transposed format is read similarly as "--tfile filenamebase", which will automatically read "filenamebase.tped" and "filenamebase.tfam". They also can be specified separately as "--tped tpedfile.tped --tfam tfamfile.tfam".

An alternative input is the ".genome" output file from PLINK. Use "--genome filename.genome" to read in the PLINK genome output IBS data. Expected relationships cannot be derived using this method at this time.

The standard map/ped and tped/tfam files can be generated from each other. To convert a tped/tfam fileset into ped/map use "--tfile myfilenamebase --recode". For coding ped/map to tfam/tped "--file myfilenamebase --recode --transpose" is used instead. In all cases the output separates alleles by whitespace, i.e. "AA" genotype becomes "A A" in the file to maintain compatibility with PLINK. Similarly the genetic position is

always printed. If it was not specified in the input 0 is used at all positions for the genetic distance. The web-based SNPduoWeb[3] visualization tool can use data exported from the command-line SNPduo IBS summarization tool. To export data from ped/map to SNPduoWeb format use "--file myfilenamebase --recode --webDuo". It is critical to note that the conversion to SNPduoWeb format recodes the genotypes to an AA/AB/BB convention, regardless of the input. The 'A' allele is the allele with the highest frequency in the input file, and the 'B' allele has the minor allele frequency in the input file.

ANALYSIS OPTIONS

There are several different analysis options in SNPduo. The output basename for all output files (including recoded ped/map/tped/tfam files) defaults to "snpduo". To specify a different output base name invoke "--out myoutcustombase" at run time.

The snpduo "--counts" option finds the count of each IBS state (IBS0, IBS1, IBS2, IBS2*) for each pairwise comparison to a comma-separated text file with the extension ".count". The summary calculations (described below) can be printed to a file (.summary) as using the "--summary" switch. This summary information is useful for downstream analysis using direct plotting, clustering analysis, or principal components analysis.

Relationship information can be extracted from input files using SNPduo as well. The "--specified" switch uses the pedigree information to infer the specified first-degree relationships and print them to file (.specified). Currently only first degree relationships are reported, though deeper recursion may be ported in later. The summary information can be used to calculated expected relationships using the "--calculated" switch. First the

---

[3] http://pevsnerlab.kennedykrieger.org/SNPduo

summary information is calculated, and then previously empirically determined cutoff points are used to define the most likely relationship. The pairwise comparisons with summary information and inferred relationship are then printed to file (.theoretical). The inference method now makes use of IBS* information in addition to mean and standard deviation of IBS. To run the old method of relationship determination use the "--oldcalculated" switch. However, use of the current method is recommended.

One of the most useful SNPduo functions is the printing of relationships that conflict with the supplied pedigree information into a file (.conflicting). The default "--conflicting" switch will use the current relationship calculation method. If you wish to find conflicting relationships with the deprecated method you MUST use "--oldcalculated --conflicting".

Summary of reported values

SNPduo calculates several summary methods on genotype data, including mean of IBS state, standard deviation of IBS state, counts of IBS0, IBS1, IBS2, and IBS2*. In the following equations IBS_i represents the vector of IBS states for each allele between two individuals. IBS0, IBS1, IBS2, and IBS2* all represent the counts of each type of IBS state between two individuals.

Equation 1 - Mean IBS Calculation

$$MeanIBS = \frac{\sum_{i=0}^{n} IBS_i}{n}$$

Equation 2 - Standard Deviation of IBS Calculation

$$SDIBS = \sqrt{\frac{\sum_{i=0}^{n} (IBS_i - MeanIBS)^2}{n}}$$

Equation 3 - IBS Informative Percent Calculation

$$InformativePercent = \frac{IBS2^* + IBS0}{IBS0 + IBS1 + IBS2}$$

Equation 4 IBS2* Percent Calculation

$$IBS2^* Percent = \frac{IBS2^*}{IBS0 + IBS1 + IBS2}$$

Equation 5 - IBS2* Percent of Informative SNPs Calculation

$$IBS2^* PercentOfInformative = \frac{IBS2^*}{IBS0 + IBS2^*}$$