Documentation for PRESTO 1.0

Brian L. Browning

The University of Auckland
Auckland
New Zealand

October 10, 2007

Contents

1. Introduction

PRESTO is a software package for small and large-scale genetic association analysis. PRESTO performs allelic and genotypic tests of stratified or unstratified data and adjusts for multiple testing using permutation to assess statistical significance. PRESTO computes empirical distributions of order statistics (the k-th order statistic is the k-th largest test statistic) via permutation for one-stage or two-stage genotyping designs. The distributions of order statistics can be used to test whether the top ranked markers have lower p-values than expected by chance and can be used to determine the number of top-ranked markers to test in subsequent studies. PRESTO can also perform an omnibus analysis of both single markers and haplotype clusters identified with the BEAGLE software package [3].

PRESTO is computationally efficient and can analyze millions of markers genotyped on thousands of samples in a few hours of computing time using 1000 permutations of the trait status. PRESTO can also run in parallel on multiple processors if required. PRESTO is written in Java and runs on most computing platforms (e.g. Windows, Linux, Unix, Solaris, Mac).

2. Citing PRESTO

Please check the PRESTO web site for the most up-to-date citation details (`www.stat.auckland.ac.nz/~browning/presto/presto.html`). Until a journal citation is given on the PRESTO web site, please cite:

> B L Browning. PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one- and two-stage genetic association studies [Abstract 2045]. Presented at the annual meeting of The American Society of Human Genetics, October 23-27, 2007, San Diego, CA. Available from `http://www.ashg.org`.

In your published analysis, please report the PRESTO version number.

3. Files in the PRESTO software distribution

PRESTO is freely available and can be downloaded from the PRESTO web site:

`www.stat.auckland.ac.nz/~browning/presto/presto.html`

The PRESTO software distribution includes the following files:

1. **presto.jar**, the executable file for running PRESTO (see Section 5).
2. **presto_1.0.pdf**, the documentation for PRESTO 1.0.
3. A folder named **example** containing files used in the example PRESTO analysis described in Section 7. The **example** folder contains one input file (**example.bgl**), and three output files (**example.log**, **example.null**, and **example.pval**).
4. A folder named **utility** containing a utility program, **unphased2beagle**, that creates a file in BEAGLE format from a data file and a pedigree file in linkage or QTDT format (see Appendix A.1). BEAGLE format is described in Section 4.

4. BEAGLE file format

PRESTO input files must be in BEAGLE format. The data for each variable (genetic marker, affection status, or population stratum) is given in a separate line. The first column is a single characters ("M", "A", or "S") describing the information on each line: "M" for marker allele data, "A" for affection status data, and "S" for population stratum data. The second column gives the name (e.g. marker name, disease name, or the population stratification variable name) of the data given on each line. Columns 3-4 give the data for the two alleles of the first individual, columns 5-6 give the data for the two alleles of the second individual, and so on. Note that in a line of affection status or population stratum data, the affection status or population stratum is given twice for each individual (once for each allele), and the affection status and population stratum for both alleles of an individual must be the same. Affection status should be coded as **1** for unaffected individuals and **2** for affected individuals. Affection status and population stratum (if a population strata line is included) can not be missing for an individual, but genotypes or alleles can be missing. When PRESTO is run, the user specifies the missing allele code (e.g. "?", "0", or "-1").

You can include comment lines in your BEAGLE file. A comment line is any line whose first field is not "M", "A", or "S". Comment lines are ignored by PRESTO and can have any format. To ensure compatibility with later versions of PRESTO, we recommend using the hash character, "`#`", as the first field of a comment. You can use the "`#`" character to comment out data lines of a BEAGLE file or to include additional information (e.g. sample identifiers) in a BEAGLE file.

Below is an example of a BEAGLE file for three individuals genotyped for three markers with missing alleles coded as "?" :

```
# sampleID       1001  1001  1002  1002  1003  1003
A diabetes       1     1     2     2     2     2
S strata         A     A     A     A     B     B
M rs1248696      1     3     3     3     3     3
M rs2289310      2     2     ?     ?     1     2
M rs2289311      4     4     2     2     4     2
```

The first line (`# ...`) is a **comment line** that is ignored by PRESTO. The second line (`A ...`) is an **affection status line** that gives the diabetes affection status for each allele (1 = unaffected, 2 = affected). The third line (`S ...`) is a **stratum line** that gives the population stratum for each allele. The last three lines (`M ...`) are **marker lines** that give marker alleles for the three markers (`rs1248696`, `rs2289310`, and `rs2289311`).

In this example BEAGLE file, the first individual (columns 3-4) is unaffected and belongs to stratum A, the second individual (columns 5-6) is affected and belongs to stratum A, and third individual (columns 7-8) is affected and belongs to stratum B. The first individual has genotypes 1/3, 2/2, 4/4, the second individual has genotypes 3/3, ?/?, 2/2, and the third individual has genotypes 3/3, 1/2, 4/2 for markers rs1248696, rs2289310, and rs2289311 respectively.

BEAGLE format imposes very few constraints on your data files. For instance:

- The fields on each line can be separated by one or more white space characters (e.g. any combination of one or more spaces and tabs).
- There is no limit on the number of alleles. In particular, triallelic SNPs and microsatellite markers can be used. For multi-allelic markers, each allele is tested for association with affection status (with the other alleles grouped together)
- Marker alleles can be any sequence of characters that does not contain white space.

PRESTO performs single marker analysis. Consequently, PRESTO does not require markers in the BEAGLE file to be in chromosomal order. However, PRESTO can be used in conjunction with the BEAGLE program [3] to analyze haplotypic data: use BEAGLE [3] to identify localized haplotype clusters, use the pseudomarker utility program (included in the BEAGLE software distribution) to create a BEAGLE input file whose markers represent the haplotype clusters, and test the resulting BEAGLE input file using PRESTO (for more information see: `www.stat.auckland.ac.nz/~browning/beagle/beagle.html`).

It is also possible to use PRESTO to analyze transmitted and untransmitted haplotypes from trio studies (affected individuals and their parents). Use the `diplotypes=false` option described in Section 5.2, and code the affection status of transmitted alleles as **2** and the affection status of untransmitted alleles as **1**. When using the `diplotypes=false` option, the analysis assumes alleles are independent, and alleles in adjacent columns (e.g. columns 3-4, 5-6, etc.) are not required to have the same affection status or population stratum.

5. Running PRESTO

PRESTO is written in Java and requires a Java interpreter. A Java interpreter is probably already installed on your computer. However, if it is not installed or if it is an old version, the Java interpreter can be downloaded free of charge from the `java.sun.com` web site. You will need to download and install the standard edition (SE) Java Runtime Environment (JRE) 5.0 (or later version).

To run PRESTO, enter the following command at the computer prompt:

(1)
```
java -Xmx600m -jar presto.jar <arguments> file1 file2 ...
```

where `<arguments>` is a space separated list of PRESTO arguments and `file1 file2 ...` is a list of genotype data files in BEAGLE format.

The `-Xmx600m` Java option allocates 600 Mb of memory available to the Java interpreter. This is usually a sufficient amount of memory. However, for an unusually large analysis (say involving more than 10,000 permutations or more than 5,000 individuals), it may be necessary to increase the available memory. For example, if your computer has 2 Gb of memory, you can allocate 1.5 Gb of memory by replacing the `-Xmx600m` Java option with `-Xmx1500m`.

Each PRESTO argument has the format `parameter=value`. There is no white-space between the parameter and "=" character or between the "=" character and the value.

The filenames `file1 file2 ...` can not contain the "=" character. PRESTO permits you to divide your genotype data among multiple files (e.g. one file per chromosome).

At least one input file is required. The input files must be in BEAGLE format which is described in Section 4. PRESTO can read compressed files that are compressed using the gzip algorithm. Any data file that ends in ".gz" is assumed to be compressed using gzip.

All arguments are described below. There are only 3 required arguments: `trait`, `out` and `missing`. Other arguments are optional and have sensible default values. The `diplotypes=false` argument must be specified when analyzing transmitted/untransmitted alleles.

5.1. **Required arguments.**

- `missing=<missing allele code>` where `<missing allele code>` is the character or sequence of characters used to represent a missing allele (e.g. `missing=-1` or `missing=0` or `missing=?`). The `missing` argument is required. If your data set has no missing alleles then set the `missing` parameter to any character or sequence of characters that is not used as an allele.

- `trait=<trait file>` where `<trait file>` is the name of the BEAGLE file (see Section 4) containing the affection status for each allele. The affection status is specified on a single line beginning with the character "A". All alleles must have an affection status specified. If the file specified with the `trait` parameter contains marker data, the file must also be specified in the genotype data file list (`file1 file2 ...` in the command line (1) above). If the file specified with the `trait` argument contains multiple affection status lines (i.e. lines whose first field is "A"), then only the first affection status line will be used. The `trait` argument is required.

- `out=<output file prefix>` where `<output file prefix>` is the prefix for the PRESTO output files. For example, if `out=presto` is specified, the output files will be **presto.log**, **presto.pval**, and **presto.null**. See Section 6 for a description of the output files. The `out` argument is required.

5.2. **Optional arguments.**

- `strata=<strata file>` where `<strata file>` is the name of the BEAGLE file (see Section 4) containing the population stratum for each allele. The population stratum for each allele is specified on a single line beginning with the character "S". If the `strata` argument is used, all alleles must have a population stratum specified. If the file specified with the `strata` parameter contains marker data, the file must also be specified in the genotype data file list (`file1 file2 ...` in the command line (1) above). If the file specified with the `strata` argument contains multiple population stratum lines (i.e. lines whose first field is "S"), then only the first population stratum line will be used. The `strata` argument is optional. If it is omitted, all alleles are assumed to belong to a single population stratum.

- `test=<association tests>` where `<association tests>` is one or more characters from the set {`tardo`} (e.g. `test=t` or `test=tardo`) where

  `t` = allelic trend test
  `a` = allelic test
  `r` = recessive test (groups major allele homozygotes and heterozygotes)
  `d` = dominant test (groups minor allele homozygotes and heterozygotes)
  `o` = overdominant test (groups minor and major allele homozygotes).

  If the allelic trend test is specified (`t`), a stratified allelic trend test is performed [6]. For the remaining tests (`a`, `r`, `d`, or `o`) a 2×2 contingency table is constructed for each stratum and a Cochran-Mantel-Haenszel test with continuity correction is performed [7, 1]. The test statistic for each marker is the maximum χ^2 statistic (maximized over the specified allelic and genotypic tests). If a marker has more than two alleles, each allele is tested for association with affection status by grouping the other alleles. Thus a triallelic marker is tested as if it were three diallelic markers, and will result in 3 test statistics. The `test` argument is optional. The default value is `test=t`. Only the allelic test (`test=t`) is permitted when `diplotypes=false`. See the `diplotypes` argument in this section for more details.

- `seed=<random seed>` where `<random seed>` is an integer seed for the random number generator. The `seed` argument is optional. The default value is `seed=-99999`. The seed for the random number generator determines the sequence of permutations of the trait status. The `seed` parameter can be used to parallelize an analysis as discussed in Section 6.3.

- `nperms=<number of permutations>` where `<number of permutations>` is a nonnegative integer giving the number of permutations of the affection status that will be used. You can skip permutation testing by setting `nperms=0`. The `nperms` argument is optional. The default value is `nperms=1000`. The computational time for permutation testing is linear in the number of permutations. Typically 1,000, or 10,000 permutations are used to determine experiment-wide statistical significance.

- `topranks=<number of order statistics>` where `<number of order statistics>` is a nonnegative integer giving the number order statistics that will be written to the null p-value file (see Section 6.2). For each marker and permutation of the trait status, the test statistic is the maximum χ^2 statistic from the association tests specified with the `test` parameter. For each permutation, the test statistics are sorted in decreasing order, and the largest `topranks` test statistics are written to the null p-value file (.null). The `topranks` argument is optional. The default value is `topranks=1`.

- `threshold=<threshold for 2nd stage>` where `<threshold for 2nd stage>` is a non-negative floating point number giving the minimum first-stage test statistic required for a marker to be genotyped in the second stage of a two-stage genotyping

design. For each permutation and for each diallelic marker, data from one-half of the cases and one-half of the controls is used to calculate the test statistic. If the test statistic is greater than or equal to the `threshold` parameter then the test statistic is recalculated using the data from the entire sample.

The `threshold` argument is used to determine significance levels for two-stage genotyping designs using the first-stage samples as suggested by Frank Dudbridge [4]. The p-value file (.pval) is not produced when the `threshold` parameter is set greater than 0.0.

The test statistics in the null p-value file (.null) can be used to determine significance for any of the top k ranks in your two-stage study where k is equal to the `topranks` parameter. The `threshold` argument is optional. The default value is `threshold=0.0` which corresponds to a one-stage genotyping design.

- `diplotypes=<true/false>` where `<true/false>` is `true` if the alleles from the same individual are always paired in the PRESTO input file so that the third and fourth columns are the alleles for the first individual, the fifth and sixth columns are the alleles for the second individual, and so on, and `<true/false>` is `false` if the alleles are not paired. The `diplotypes` options controls how the trait status is permuted during permutation testing. If `diplotypes=true`, the trait status is permuted for the individuals so that both alleles for each individual have the same permuted trait status. When `diplotypes=false`, the trait status is permuted for the alleles (rather than for the individuals). Only the allelic test can be performed when `diplotypes=false` (see the `test` parameter in this section). The `diplotypes` argument is optional. The default value is `diplotypes=true`.

6. Output files

PRESTO produces three output files: a log file (.log), a file of test statistics and a permutation p-value for each marker (.pval), a file of the maximum test statistics (.null) obtained when testing the markers for association with the permuted trait statuses.

The log file (.log) and the p-value file (.pval) will be useful to all users. The null file (.null) will be useful to some, but not all, users.

6.1. **The log file (.log).** The log file gives a summary of the analysis that includes the PRESTO version number, a description of the command line arguments, a list of the command line arguments for the analysis, the elapsed time for the analysis, and a list of all markers from the p-value file (.pval) with a permutation p-value less than 0.2.

6.2. **The p-value file (.pval).** The p-value file records the allelic and genotypic test statistics and the maximum test statistic for each marker, along with a permutation p-value for the maximum test statistic.

The first line of the p-value file is a header line describing the columns of the file. Each line (except the header line) gives the p-values from testing one marker for association with the trait status. The markers in the PRESTO p-value file are in the order they appear in the input files (see Section 4).

The first few lines of the p-value file look like this:

```
Marker     Allele     trend_X2     max_X2     perm_p_value
m1         0          3.753        3.753      1.000
m8         1          8.217        8.217      0.0462
m12        1          4.053        4.053      1.000
```

The first field on the line is the marker identifier. The second field is the marker allele that is tested. If the marker has more than two alleles, each allele is used to define a diallelic marker by grouping all other alleles as the second allele (see the `test` parameter in Section 5.2 for more details).

After the marker and allele fields, the next columns give the χ^2 statistics for the allelic, recessive, dominant, and overdominant tests. Columns corresponding to tests which were not performed are omitted. The second-to-last column gives the maximum χ^2 statistic for the marker. If only one test is performed, as is the case in the preceding p-value file excerpt, the maximum χ^2 statistic equals the statistic for single test that was performed. The final column gives the permutation p-value.

The permutation p-value is a measure of significance that accounts for multiple testing. For example, if your significance level is $\alpha = 0.05$, and a marker has a permutation p-value $p < 0.05$, the association is significant after accounting for multiple testing. More generally, for a given significance level α ($0 \leq \alpha \leq 1$), the probability of observing one or more markers with a permutation p-value less than or equal to α is less than or equal to α under the null hypothesis that the trait and marker data are independent [2].

Given a set of tests specified with the `test` parameter, the permutation test randomly permutes the trait status and tests the markers for association with the permuted trait status. If `diplotypes=true`, the trait status is permuted for the individuals so that both alleles for each individual have the same permuted trait status. When the data consists of transmitted and untransmitted alleles, the `diplotypes=false` argument (see Section 5.2) must be used so that the trait status is permuted for the alleles rather than for the individuals.

For each permuted trait status the set of tests determined by the `test` parameter is applied to all markers and the largest test statistics are saved and written to the null statistics file (described in Section 6.3). If a marker has a maximum χ^2 statistic $t_{\max}$ (maximized over all tests specified with the `test` parameter) when tested for association with the unpermuted trait status, and if for k out of N permutations of the trait status there exists at least one marker with a maximum χ^2 statistic greater than or equal to $t_{\max}$ when tested for association with the permuted trait status, the permutation p-value for the marker is $(k + 1)/(N + 1)$ [2]. Under the null hypothesis, expect most alleles to have a permutation p-value of 1.000 since the p-value of a single marker is being compared to the minimum p-value from all markers for each permutation of the trait status.

The p-value file is designed to be imported into a spreadsheet or a statistical software package. However, if you want to quickly identify the most significant markers, look in the

output log file (.log). The log file contains a list of all markers with a permutation p-value less than 0.2.

6.3. **The null statistics file (.null).** The null statistics file gives a random sample from the distributions of the largest order statistics. The number of order statistics reported is specified with the `topranks` parameter (see Section 5.2). The null statistics file can be used to determine statistical significance of any statistic that is computed from order statistics (e.g. rank truncated products [5]).

For each permutation of the trait status, the test statistic for each marker is calculated, and the `topranks` largest test statistics are written to the null statistics file. The test statistic for a marker is the largest χ^2 statistic from the tests specified with the `test` parameter described in Section 5.2. The j-th line of the null statistics file gives the `topranks` largest maximum test statistics from the j-th permuted trait status for $j = 1, 2, \ldots, N$ where N equals the value of the `nperms` parameter (see Section 5.2). For each line the test statistics are listed in decreasing order. If `nperms=0` or `topranks=0` the null statistics file will be empty. The sequence of permutations of the trait status is determined by the `seed` argument (see Section 5.2).

PRESTO can be run in parallel to reduce running time. For example, if you use a different random seed but the same input files in each of the parallel runs, the null files from each computing run can be concatenated (e.g. using the unix cat command) to obtain the null file for all permutations of the trait status. Alternatively, you can use the same random seed but different marker files in each parallel run. In this case, the same sequence of permutations of the trait status will be tested for association with the disjoint sets of markers in each parallel run. If the top K order statistics in each computing run are written to the null files, then the null files from each computing run can be pasted side-by-side (e.g. using the unix paste command), and the top K values in each line sorted in decreasing order will give the null file for the set of all markers.

7. An example PRESTO analysis

Files from an example PRESTO analysis are included in this software distribution. The example data are 200 markers for 4000 haplotypes (1000 case and 1000 control individuals). The example data were generated using the Cosi program [8].

The example files are

1. example.bgl BEAGLE file with genotype data (see Section 4)
2. example.log PRESTO output log file (see Section 6.1)
3. example.pval PRESTO output p-value file (see Section 6.2)
4. example.null PRESTO output null p-value file (see Section 6.3)

The PRESTO output files (.log, .pval, .null) are created from the PRESTO input file (example.bgl) using the command:

```
java -Xmx600m -jar presto.jar trait=example.bgl out=example missing=?  test=tdr
topranks=5 example.bgl
```

The PRESTO output log file (example.log) contains the following excerpt that identifies the 4 markers that have permutation p-values less than 0.2.

Marker	Allele	rec_X2	dom_X2	trend_X2	max_X2	perm_p_value
m23058	0	2.541	12.74	13.25	13.25	0.07992
m23612	0	3.441	13.74	14.66	14.66	0.03996
m23903	1	0.06322	12.42	4.128	12.42	0.1279
m24828	0	1.330	12.09	12.21	12.21	0.1449

References

[1] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, New York, second edition, 2002.

[2] J. Besag and P. Clifford. Sequential Monte Carlo p-values. *Biometrika*, 78:301–304, 1991.

[3] B. R. Browning and S. R. Browning. Efficient multilocus association mapping for whole genome association studies using localized haplotype clustering. *Genet Epidemiol*, 31:365–375, 2007.

[4] F. Dudbridge. A note on permutation tests in multistage association scans. *Am J Hum Genet*, 78:1094–1095, 2006.

[5] F. Dudbridge and B. P. C Koeleman. Rank truncated product p-values, with application to genomewide association scans. *Genet Epidemiol*, 25:360–366, 2003.

[6] N. Mantel. Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *J Am Stat Assoc*, 58:690–700, 1963.

[7] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer I*, 22:719–748, 1959.

[8] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human sequence variation. *Genome Research*, 15:1576–1583, 2005.

Appendix A. Utility program for creating files in BEAGLE format

A.1. **unphased2beagle.** The **unphased2beagle** program creates a BEAGLE file from a data file and a pedigree file. The format for the data and pedigree files is similar to linkage and QTDT format. QTDT format is described at
`http://www.sph.umich.edu/csg/abecasis/QTDT/docs/data.html`.

The pedigree file has rows corresponding to individuals and columns corresponding to variables. The first five columns are fixed and give the pedigree identifier, subject identifier, father identifier, mother identifier, and gender. If any of these variables are unknown or undefined, then `0` is used. The pedigree identifier, father identifier, and mother identifier are generally not defined for case-control data. The remaining columns are variable and specified by the data file.

The lines of the data file correspond to the variables in the pedigree file (in the order they appear as columns in the pedigree file, beginning with column 6). Each line of the data file has two fields. The first field is a single character identifying the type of data in the column, and the second field is the identifier for the variable. If the first field is "M" (for marker), the variable is a genotype and corresponds to two columns of the pedigree file; otherwise, the variable corresponds to a single column of the pedigree file. (Note: the "S[n]" code used in QTDT is not supported by `unphased2beagle` since the first column of the data file must be a single character).

The `unphased2beagle` program creates a BEAGLE format file whose first two columns are the first two columns of the data file. The third and fourth columns give the two alleles

for the first individual in the pedigree file, the fifth and sixth columns give the two alleles for the second individual in the pedigree file, and so on.

For example, suppose the data file is

```
A   diabetes
M   rs1248696
M   rs2289311
T   BMI
C   age.of.onset
```

Then the pedigree file will have five fixed columns and seven variable columns (one column each for the `A`, `T`, and `C` variables and two columns each for the two `M` variables).

If the pedigree file associated with the data file is

```
0    1001   0   0   1   1   A  G   T   T   23.0   X
0    1002   0   0   1   2   G  G   T   T   24.0   34.5
0    1003   0   0   2   2   G  G   T   C   25.0   67.8
```

then the following BEAGLE file will be created:

```
A     diabetes     1     1     2     2     2     2
M     rs1248696    A     G     G     G     G     G
M     rs2289311    T     T     T     T     T     C
T     BMI          23.0  23.0  24.0  24.0  25.0  25.0
C     age.of.onset X     X     34.5  34.5  67.8  67.8
```

Recall that lines of the BEAGLE file whose first character is not `A`, `M` or `S` will be treated as comments (see Section 4).

To run the `unphased2beagle` program enter

```
java -Xmx600m -jar unphased2beagle <arguments>
```

where `<arguments>` is a space separated list of arguments and each argument has the format `parameter=value`. There is no white space between the parameter and the "`=`" character or between the "`=`" character and the value. The arguments are

- `pedigree=<pedigree file>` where `<pedigree file>` is the filename of a pedigree file. The `pedigree` argument is required.

- `data=<data file>` where `<data file>` is the filename of a data file. The markers in the data file must be in chromosomal order. The `data` argument is required.

- `beagle=<beagle file>` where `<beagle file>` is the filename of the BEAGLE file that will be created from the the pedigree and data files. The `beagle` argument is required.

- `skip=<number of columns to skip>` where `<number of columns to skip>` is the number of columns of fixed data in the pedigree file. Typically the first 5 columns are fixed; however, if you have retained the subject identifier and gender columns out of the first five columns in the pedigree file, and deleted the pedigree identifier, father identifier, and mother identifier columns, then set `skip=2`. The `skip` argument is optional with default valute `skip=5`.