Pathway_PDT version 1.0

04/23/2013

Usage: pathway_pdt control_file

References

1. A test for linkage and association in general pedigrees: the pedigree disequilibrium test.
Martin ER, Monks SA, Warren LL, Kaplan NL.
Am J Hum Genet. 2000 Jul;67(1):146-54. Epub 2000 May 23

2. Gao X, Starmer J, Martin ER. (2008), A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet. Epidemiol. 32(4):361-9.

Input files

1. A pedigree and a map file. There are 3 distinct input format options.

a. The standard .ped and .map files

The leading 6 columns are

<ped ID> <indiv. ID> <father> <mother> <sex> <affection status,0 = unkown, 1= inaff, 2= aff>

Pedigrees are expected to be consecutive. Pedigree and individual ID numbers are integers. The alleles are coded as 0 for missing data, while valid alleles are 1 and 2. The map file has 3 leading columns which are chromosome, rs-number (or any other identifier string) and base pair. Additional columns are ignored. The columns are separated by one or more spaces. The <tab> character is not an acceptable column separator. The .ped file will look something like:

990007 1 0 0 2 1 1 1 1 2 …

990007 2 0 0 1 1 1 2 1 2 …

990007 4 2 1 2 2 1 1 1 1 …

…

b. .ped and .map using ACGT

As above but the alleles are coded as upper or lower case A,C,G or T. The file will actually be recoded into a format conforming to the one described under item “a”. The resulting file will be named pdt2.ped.

c. Plink .bim, .bed, .fam format

Please consult the Plink pages for more information. Again the input will be recoded into a format conforming to option “a”.

2. A gene range file

One gene per line in no particular order. The file has 4 columns separated by one or more space characters.

<gene name> <chromosome> <start base pair> <end base pair>

It will look something like:

…

USP48 1 22004793 22109688

EPHB2 1 23037330 23241822

PINK1 1 20959947 20978003

…

3. A pathway file

As usual, columns are separated by one or more spaces. The 1st column is the pathway name which can by any string. It is followed by one or more gene names as they appear in the “gene range file” outlined under item #2. Typically each pathway contains somewhere between a few to a few hundred genes. Example:

```
1 SGIP1 ADC CLIC4 NECAP2 AGBL4 SLC45A1 TGFBR3 DBT C1orf21 PRUNE
19  RAD50 MRE11A MLH1 BLM
30  PIGM PIGV PIGB PIGZ ALG1 POMT1 ALG12 ALG8 ALG3 ALG2 SDF2 POMT2 DPM1
…
```

4. Sindex file

This file is optional. It gives the “weight” of each gene depending on the number of SNPs that fall into that gene and the LD structure. For more information please refer to the references (2). The S_index file is in 1-to-1 correspondence with the gene range file. It must have the exact same number of lines in the exact same order of the genes as they occur. The S_index file will have one integer per line. Based on our own simulations using the default (ie. no S_index file at all) will yield slightly higher power than using S_index.

5. A control file

This file contains all parameter settings for Pathway_PDT. The format is a keyword followed by a value or name. The settings do not have to be in any particular order. The <tab> character is unacceptable. Input files are expected to reside in the present working directory or you must provide the full path along with the file name. An example control file would look like:

```
inputformat:                1
plink_bed_bim_fam:          /mihg/users/rchung/WORKDIR/SIM/CAPL/wga/output/capl.bed
/mihg/users/rchung/WORKDIR/SIM/CAPL/wga/output/capl.bim    /mihg/users/rchung/WORKDIR/SIM/CAPL/wga/output/capl.fam

pedigree_file:              ped.new
map_file:                   map.new
outfile:                    pathway_outfile.out
non_marker_fields:          6
max_cpus:                   3
options:                    0
max_fam_size:               10
verbose:                    1
max_permutations:           2000
pdt_average:                0
gene_file:                  gene_pos.new
sindex_file:                sindex.txt
gene_buffer:                20000
pathway_file:               pathway.txt
print_new_map_ped:          0
```

A parameter keyword would be “outfile:” followed by the one or more spaces and the corresponding value of “pathway_outfile.txt”. For any comments start the line with #. If a default setting for the parameter is available it is in [].

A detailed information of the various parameters follows.

inputformat: [1] Refers to the format described above under #1. Use 1 for the standard .ped/.map input. Use 2 for the Plink .bed, .bim, .fam input and use 4 for the ACGT style input.

plink_bed_bim_fam: location of the 3 Plink style input files. This setting is only required if input option 2 is selected.

pedigree_file: The location of the pedigree file. Only needed if input options 1 or 4 are selected.

map_file: The location of the map file. Only needed if input options 1 or 4 are selected.

outfile: Name of the output file.

non_marker_fields: [6] The number of leading columns in a typical .ped file that are not alleles.
Typically that number is 6 unless you have covariate values stores in the file.

max_cpus: [1] Pathway_PDT is multi-threaded. It can use as many computing cores as you have available on your computer. It will not exceed the number specified under this setting. The number you enter under this setting should never exceed the number of cores available on your computer.

options: [3] 0=use all info; 1=triads only; 2=sibs only; 3=only triads if available, otherwise use discordant sib-pairs. In simulations option 3 was slightly more powerful than other options.

max_fam_size: [10] The size of the largest pedigree in the data set. If you select this number too small the program will crash. If you select it too large it will use more memory than necessary.

verbose: [0] 0 = be silent, 1 = print out all kinds of debug information that ends in .dbg.

max_permutation: [2000] It will permute the sign (+/-) of the per pedigree statistic as many times as indicated.

pdt_average: [0] 0 = use the PDT-sum statistic, 1 = use the PDT-average statistic. PDT-sum is slightly more powerful.

gene_file: Name of the gene range file described above.

sindex_file: (optional) s_index file described above. If no file is provided the weight of every gene is set to 1.

gene_buffer: [0] How many base pairs on either side of the gene should be included into the analysis.

pathway_file: The name of the pathway file described above.

print_new_map_ped: [0] Print a new set of .ped and .map files for Pathway_PDT. This might be useful if your input format differs from the one described under “The standard .ped and .map files”.

Output

The non-verbose output consists of 5 files.

1. sig_snps_in_genes.txt

   A list of all SNPs that are in or near (see gene_buffer) a gene.

2. sig_genes.txt

   This file will contain lines like:

   ```
   30 1576 2.322581 4404 3.000000 4333 1.528302
   ```

   The first is the pathway name (30). Next are pairs of gene index (line in gene_file) and score, so (1576, 2.323581), (4404, 3.0) and (4333, 1.528302).

3. permutation_Z-value.txt

1st column is the original score while subsequent columns are the permuted score. The number of columns depends on the number you entered under max_permutations.

4. permutation_info_raw.txt

   This table gives the raw statistic for the original (1st column) and each of the permutations (subsequent columns).

5. outfile

   The file name specified in the control file under “outfile:”. This file contains 4 columns.

   The pathway name, unadjusted permutation p-value, the adjusted p-value and FDR (False Discovery Rate).

   The adjusted p-value is the proportion the original Z-score beat the maximum of each of the permutations taken over all pathways.

Limits and maximums

All limits and maximums are in a file called max.h. For efficiency reasons pathway_pdt does not check those limits but seg faults instead. One of those constants is `const int max_snps_in_gene = 50000` which describes the maximum allowable number of SNPs per gene. Change those constants at your own risk.

Operating System Specifics

Pathway_pdt will run on Windows (binary provided) and Linux (source code and makefile provided). On the Windows OS it requires that the DLL pthreadVC2.dll is either on the path or resides in the same directory as the Pathway_pdt binary.

On Linux Pathway_pdt can be compiled under GNU’s g++ version 4.4.3 or 4.6.2. To compile the Linux version change into the “src” directory and type make.

Example

A small example it provided in the example directory. It has a set of input files, including the control file example.ctrl and a set of expected output files. The example should complete in less than 2 minute. The output might vary slightly since Pathway_pdt uses a random number generator whose sequence of random numbers will vary each time Pathway_pdt is executed.