Pascal User Manual

1 Introduction

Pascal (Pathway scoring algorithm) is a program for calculating gene score and pathway score p-values from GWAS-summary statistics. It makes use of LD information derived from the 1KG-EUR sample by default. It has been tested on Unix and Mac OsX.

2 Requirements

Pascal is mainly written in Java and therefore needs a Java interpreter to run. If the 32-Bit Java virtual machine is used, *Pascal* may run into memory problems (specifically, java VM might not allow to assign the minimal required heap size. You may also get a runtime error if the data set you are analysing is too large). If it is an option, installation of the 64-Bit Java virtual machine should be considered. Pascal also makes use of some native libraries which should be installed automatically during installation. If this fails, it is still possible to use the program but it might run much slower and some options will not be available (see section *options available without native support*). In general, on non-OsX systems, the *gcc* compiler collection has to be installed for successful installation of *Pascal*.

3 Installing *Pascal*

Go to the unzipped folder `PASCAL`. Call the installation script:

```
bash installScript.sh
```

This should compile `openBLAS` and additional fortran libraries locally if neccessary.

4 Running *Pascal*

Pascal is designed to be used as a command-line tool. To see its basic operation go to the folder `PASCAL` and execute the following command:

```
./Pascal --pval=resources/gwas/EUR.CARDIoGRAM_2010_lipids.HDL_ONE.txt --chr=22
```

The following files should be produced in `PascalPackage/output/` path:

```
EUR.CARDIoGRAM_2010_lipids.HDL_ONE.sum.genescores.chr22.txt
```

The gene score results.

`EUR.CARDIoGRAM_2010_lipids.HDL_ONE.PathwaySet--msigBIOCARTA_KEGG_REACTOME--sum.chr22.txt`

The pathway score results. Additionally, the following auxilliary files are produced:

`EUR.CARDIoGRAM_2010_lipids.HDL_ONE.sum.fusion.genescores.chr22.txt`

The fusion gene score results.

`EUR.CARDIoGRAM_2010_lipids.HDL_ONE.sum.fusion.genescores.chr22.txt`

The fusion gene score results.

`EUR.CARDIoGRAM_2010_lipids.HDL_ONE.sum.numSnpError.chr22.txt`

A file containing all genes that contain no SNPs or contain more SNPs than given by the `--maxsnp` option.

`settingsOut.txt`

A file containing the internal settings used by *Pascal*.

5 Command Line Options

Below are a list of command line options that can be provided to the program.

`--pval`

This option has to be set. It gives a path to the a file containing SNP p-values. The file should be a tab-separted two column file where the first column should be each SNPs rs number and the second is its p-value. The rows do not have to be in a particular order.

`--chr`

If this option is supplied *Pascal* only runs over one chromosome. Else it runs over all 22 autosomes. This option should be supplied with a number between 1 and 22 (e.g.: `--chr=22`).

`--up`

Gives the number of base-pairs upstream of the transcription start site that are still counted as belonging to the gene region. The default is 50'000. This option should be supplied with a number (e.g.: `--up=50000`).

`--down`

Gives the number of base-pairs downstream of the gene-body that are still counted as belonging to the gene region. The default is 50'000. This option should be supplied with a number (e.g.: `--down=50000`).

`--maxsnp`

Sets the aximum number of SNPs per gene. If a gene has more SNPs in its region it will not calculate the score. If the option is set to -1, all genes will be computed. The default is 3000. This option should be supplied with a number (e.g.: `--maxsnp=3000`).

`--genescoring`

Chooses the genescoring method. The default is `sum`. This option should be supplied with either max or sum (e.g.: `--genescoring=sum`).

`--runpathway`

Chooses whether *Pascal* should be calculate pathway scores. The default is `off`. This option should be supplied with either on or off (e.g.: `--runpathway=on`).

`--outsuffix`

Adds an additional string to the output file names produced. For example: `--outsuffix=.MYNAME` changes the produced output from

`EUR.CARDIoGRAM_2010_lipids.HDL_ONE.genescores.chr22.txt`

to

`EUR.CARDIoGRAM_2010_lipids.HDL_ONE.MYNAME.genescores.chr22.txt`

`--mergedistance`

Gives the genomic distance in mega-bases that the program uses to fuse nearby genes during the pathway analysis. Only has an effect if `runpathway=on`. The default is 1. This option should be supplied with a number (e.g.: `--mergedistance=1`).

`--mafcutoff`

SNPs with maf below that value in the european samle of 1KG will be ignored. The default is 0.05. This option should be supplied with a number between 0 and 1 (e.g.: `--mafcutoff=0.05`).

`--genesetfile`

Gives the file name to a gmt-file where the gene sets are defined. The default is

`resources/genesets/msigdb/msigBIOCARTA_KEGG_REACTOME.gmt.`

6 Working with custom reference populations

Pascal allows to use custom reference populations instead of the 1KG-EUR sample (For an example about formatting and parameter setting, check the bash script `examples/exampleRunFromTped.sh`). To make use of this option, you have to provide your genotype information in gnu-zipped, space-separated and 1-2-coded `tped`-files split by chromosome. the `tped` format has been popularized by the *plink*-tool for GWAS analysis.
(see `http://pngu.mgh.harvard.edu/~purcell/plink/` especially the commands `--transpose` and `--recode12` to create 1-2-coded `tped`-files). There are two options that have to be specified to use custom reference populations, `custom` `customdir`.

`--customdir`

Gives the path to where the gnu-zipped `tped`-files are stored.

`--custom`

Gives the prefix of a gnuzipped `tped`-files. The total filename per chromosome is 'prefix'.chr'chrNr'.tped.gz.
For example, lets say you created files in the *Pascal* subdirectory `resources/ASN/` that contain the 1KG reference genotypes of the Asian population, and your file for chromsome 1 has the following name:

`ASN.chr1.tped.gz`

Then the you should supply the following flags:

`--custom=ASN`

and

`--customdir=resources/ASN/`

The first time a particular custom referrence population is used, *Pascal* will prepare specially formatted binary files (i.e. java serialized files) within the same folder. If there is a problem during this process (say because thegenoytpe file is not formatted correctly) *Pascal* will produce a corrupted version of these files. You will have to remove these c serialized files before trying again.

7 Options not available without native support

If part of the installation failed, one can still run the program. However, It might be slow. Additionally, max genescoring method might not work. To use the tool in that circumstance, set option `--genescoring=sum`. Also, you might want to update your *gcc* compiler collection and try again.