User Manual for 'proxyGeneLD'

Table of Contents

1. Introduction

This document is intended to show how to run the Perl script 'proxyGeneLD', which is available on https://github.com/Rundmus/ProxyGeneLD, together with two additional Perl programs for preprocessing data files. 'proxyGeneLD' calculates gene-wide adjusted P-value for each gene by reading SNP list files and trait/disease association statistics, which are common outputs of genome-wide association studies.

2. How to use 'proxyGeneLD'

This program is written in 'Perl'. For any issues about the 'Perl' itself please refer to the official Perl website (www.perl.org). The program uses some core modules (File::Spec, File::Basename, Switch, List::Util) which are included in typical installations of Perl, but prior to use these should be confirmed to be present.

2.1. Input and Preprocessed Data files

The program requires not only an input file but also 24 data files that have been preprocessed by the perl programs explained later in chapter 3. The input file is the output of a genome-wide association study with significance levels of phenotype association for each SNP. The file must be written in text and contain at least 2 columns, one for RefSNP accession IDs (rs numbers) of tested SNPs and the other for significance levels (P-values), and be delimited by space, tab, or comma. The

additional data files comprise one file of gene IDs and symbols and 23 files for LD structure in chromosome 1 to 22 and X. The path to the files should be specified by modification of the 2nd and 4th lines of the file "PATH_proxyGeneLD" in the same folder (or directory) where this program is located. Note that the file was created on MS Windows operating system. So, the user of other operating system such as UNIX or MAC, has to make sure every line ends with the system's newline character. One simple way to do is hitting 'delete' and 'newline' keys (e.g. □) at every end of lines when you edit the file. One example of the path files is shown below.

```
#The file of 'refGeneIDnSym' produced by 'preprocess1_GeneIDnSym.pl'
C:\PROJECTS\refGeneIDnSym_UCSC.txt
#Path to the files for HapMap LD created by 'preprocess2_LDfiles.pl'
C:\PROJECTS\HapMap\
```

2.2. Command line

This program requires 4 arguments describing the input file and it can accept 3 more options for different analyses.

```
Usage: >Perl proxyGeneLD (qry) (rs) (P) (d) [r2] [5'] [3']
```

`(qry)` – file name that includes SNP list, output of genome-wide association study.

`(rs)` – column number containing rs numbers of SNPs

`(P)` – column number of P-values

`(d)` – delimiter used in the `(qry)` file (Possible options are 'space', 'tab' and 'comma'.)

`[r2]` – threshold of r^2 that classifies two SNPs are so highly correlated to join proxy cluster (default is 0.8)

`[5']` – 5' promoter region length that is included as the genic region (default is 1000bp)

`[3']` – possible 3'UTR extension from reference transcript (default is 0bp)

2.3. Output files

Three output files are created after each run. Each lists gene-wide P-values, those after moving of genes in LD to bottom, and SNPs that were not included in output for HapMap sample. The file name and format of each file is as described below.

2.3.1. Gene list with adjusted gene-wide P-values

- file name : (input file name of SNPs)_LD(threshold)_(number of genes in this file).out
  e.g. HDLstudy_LD0.8_16934.out
- format : tab delimited text with title line and sorted by adjusted P-values in ascending order.

geneId, symbol, position, rsId, unadjP, adjP, SNP#nearGene, SNP#incl.LD, adjSNP#, length, geneInLD

- position – the largest coverage of the gene (from the closest end of any isoform to pter to qter)
- rsId – the RefSNP accession IDs of the SNP that was selected to represent the best gene-wide P-value considering LD
- SNP#nearGene – the number of SNPs in the input file which are located in the genic region (defined by arguments `[5']` and `[3']` of the program) and included in HapMap output
- SNP#incl.LD – SNP#nearGene plus the number of SNPs of HapMap which are in high LD with any input SNP
- adjSNP# – the number of SNPs used for adjusting gene-wide P-value
- length – average length of all isoforms of the gene
- geneInLD – symbols of the gene over which at least one proxy cluster spans

2.3.2. Gene list after moving genes in LD

- file name : (input file name of SNPs)_LD(threshold)_(number of genes in this file).out**Trim**
  e.g. HDLstudy_LD0.8_16934.outTrim
- format : same as the file in 2.3.1.

2.3.3. non-HapMap SNP list

- file name : (input file name of SNPs)_(number of SNPs in this file).nonHapMap
  e.g. HDLstudy_3015.nonHapMap
- format : tab delimited text without title line including 2 columns of rs numbers of SNPs and P-values that were in the input SNP list but not in HapMap output. The SNP list is sorted by P-value in ascending order.

3. Preprocess

The following preprocess programs should be run in the order as shown below after downloading all necessary data files.

Optionally, the file can be created by the first preprocess Perl program is available at same website as the Perl programs are ("refGeneIDnSym_UCSC.txt"). Preprocessing for LD files (the 2nd program does) requires a number of large files. Altogether, the size of them reaches over 35GB. So,

for users who stick to a threshold of r^2 no less than 0.8 and LD data of HapMap CEU samples, the preprocessed files are provided at the location ("hapmapCEU/hapmapCEU_GeneID_r2_0.8_chr*"). Please note that those files were created on MS Windows. So, the user of the other operating system such as UNIX, MAC has to replace newline characters of those preprocessed files with the ones used in your system.

3.1. preprocess1_GeneIDnSym.pl

```
Usage: >perl preprocess1_GeneIDnSym.pl (refG) (g2sq) (gIfo) [outF=ref...]

     refG - refGene file name from UCSC table browser
     g2sq - 'gene2refseq' file from ftp://ftp.ncbi.nih.gov/gene/DATA/
     gIfo - 'gene_info' file from ftp://ftp.ncbi.nih.gov/gene/DATA/
     outF - output file name (default = .\refGeneIDnSym_UCSC_hg18.txt)
```

This perl script combines the data on genes from 3 files which are available in public databases. 'refGene' file can be downloaded from UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start) by selecting 'RefSeq Genes' in track and 'refGene' in table. The files of 'gene2refseq' and 'gene_info' compressed in gzip format are at NCBI FTP site (ftp://ftp.ncbi.nih.gov/gene/DATA/).

3.2. preprocess2_LDfiles.pl

```
Usage: >perl preprocess2_LDfiles.pl (snpF) (gF) (pth) (pS) [th=.8]

     snpF - hapmapSNP file from UCSC table browser
     gF   - the file that was created by preprocess1_GeneIDnSym.pl
     pth  - the path includes HapMap LD data files
     pS   - population symbol used in HapMap LD data files (e.g. CEU)
     th   - threshold of r2 to trim out lower values than (default 0.8)
```

LD data files for each population of the HapMap project can be downloaded from "http://ftp.hapmap.org/ld_data/". The files are compressed in gzip format. After unzipping the files, their size becomes very large (35GB of all CEU data). So, the first preprocessing perl script trims out the data that would not be used and create files in more compact text form. Note that the threshold of r^2 selected here has to be equal to or smaller than the threshold that would be chosen in the main perl program "proxyGeneLD", because this program is erasing the LD data below the threshold. The UCSC 'hapmapSNP' file can be downloaded by selecting 'HapMap SNPs' in track and 'HapMap SNPs (population symbol)' in table at the website of UCSC table browser.