MethylPCA

Individual Programs

Contents

1. Introduction

The design philosophy of MethylPCA is to build small independent executable components first and then combine them to perform complex tasks. There are two advantages for this design style:

1) Easy to debug. Each component can be debugged independently.

2) Flexible to use. Each component can be used either independently or combined together.

This is the documentation of the independent executable programs or useful R scripts in the bin directory if the user wants to use them individually.

In MethylPCA, the R script pcaStackJobs.R allocating different jobs to different programs, and provide an easy-to-use interface. For this purpose, small executable batch files are generated by pcaStackJobs.R. Then based on the operating system and whether to execute in parallel, these executable batch files are submitted to the cluster or PC.

For some C++ programs, there are some constants in the source code that can be modified if necessary. For example, in the constants.h:

```
// Limit on number of samples
const int max_sample     = 1500;
// Number of sites for the largest chromosome.
const int max_coord = 3000000;
// Limit on number of sites for all the chromosomes used in
// calculating chunks of XXT.
// Increase if actual data has more sites.
// To use memory more efficently you can also lower if there are
// fewer sites.
const int max_coord_total = 6000000;
// Threshold to avoid very small standard deviations
const double sd_threshold     = 1.0e-10;
```

After the changes, the user can re-compile the source codes in the src directory.

2. Description of individual programs

findBlocks.exe

Description: The program findBlocks.exe combines inter-correlated methylation data at adjacent sites into a single block.

Usage: findBlocks.exe methyldir samplefile selfile selected_sites_prefix outputdir min_mean_block_cor min_new_block_cor min_n_new_block_cor chrID

Argument	
methyldir	The directory of the methylation data
samplefile	The sample file containing the sample IDs
selfile	Whether to use selection files to select sites from this chromosome. It has two string values: “yes” or “no”.
selected_sites_prefix	It specifies the path prefix of the selection files. Therefore the name of the selection files should be <*selected_sites_prefix*><*chrID*>.csv, <*chrID*> is the number of the chromosome.
outputdir	The output directory of the created blocks
min_mean_block_cor	This is the threshold for the average inter-block correlation
min_new_block_cor	This is the threshold for correlation of new site vs. sites already in the block
min_n_new_block_cor	The program will stop expanding a block if *min_n_new_block_cor* new sites/blocks have correlation below *min_new_block_cor* with sites already in block
chrID	Specify which chromosome to be processed
Input	The methylation data
Output	The output data have the same directory structure as the input methylation data. See user guide for details.

descriptives.exe

Description: The program descriptives.exe calculates the descriptives (mean and standard deviation) of each location based on the input methylation data.

Usage: descriptives.exe methyldir samplefile desdir chrID

Argument	
methyldir	The directory of the methylation data
samplefile	The sample file containing the sample IDs
desdir	The directory to store the calculated descriptives
chrID	Specify which chromosome to be processed
Input	The organization and file format of the methylation data is described in the user guide. The sample file is a one column file and each row is the sample ID without the header row
Output	The outputs are stored in the directory *desdir*. The file names are des_chr<*chrID*>.csv, where <*chrID*> is the chromosome number specified

calc_residuals_desp.exe

Description: The program calc_residuals_desp.exe regresses out covariates from the methylation data, save the residuals and then calculates the descriptives of the residuals.

Usage: calc_residuals_desp.exe methyldir covariates_file resdir desdir chrID

Argument	
methyldir	The directory of the methylation data
covariates_file	The covariates file containing the covariates to be regressed out
resdir	The directory to store the residual methylation data
desdir	The directory to store the descriptives of the residual methylation data
chrID	Specify which chromosome to be processed
Input	The organization and file format of the methylation data is described in the user guide. The covariates file is a csv file with the first column being the sample IDs and rest columns covariates. The first row contains names for each column.
Output	The calculated residuals are stored in the directory *resdir*. The descriptives are stored in the directory *desdir*. The file names are des_chr<*chrID*>.csv, where <*chrID*> is the chromosome number specified

calc_PCAinput_diag.exe

Description: The program calc_PCAinput_diag.exe calculates a certain diagonal chunk of the inner product matrix XX^T.

Usage: calc_PCAinput_diag.exe methyldir samplefile desdir outputdir start_chr end_chr row_begin row_end

Argument	
methyldir	The directory of the methylation data
samplefile	The sample file containing the sample IDs
desdir	The directory where the calculated descriptives stores
outputdir	The directory of computed chunks of XX^T
start_chr	The starting chromosome number to be processed
end_chr	The ending chromosome number to be processed
row_begin	The beginning row of the diagonal chunk
row_end	The ending row of the diagonal chunk
Input	The organization and file format of the methylation data is described in the user guide. The sample file is a one column file and each row is the sample ID without the header row.
Output	The output diagonal chunk is named as diag_<*row_begin*>_<*row_end*>.csv stored in the *outputdir*. This is a csv file storing the chunk of the matrix XX^T.

calc_PCAinput_rect.exe

Description: The program calc_PCAinput_rect.exe calculates the off diagonal chunks of the inner product matrix XX^T.

Usage: calc_PCAinput_rect.exe methyldir samplefile desdir outputdir start_chr end_chr row_begin row_end col_begin col_end

Argument	
methyldir	The directory of the methylation data
samplefile	The sample file containing the sample IDs
desdir	The directory where the calculated descriptives stores
outputdir	The directory of computed chunks of XX^T
start_chr	The starting chromosome number to be processed
end_chr	The ending chromosome number to be processed
row_begin	The beginning row of the diagonal chunk
row_end	The ending row of the diagonal chunk
col_begin	The beginning column of the rectangular block
col_end	The ending column of the rectangular block
Input	The organization and file format of the methylation data is described in the user guide. The sample file is a one column file and each row is the sample ID without the header row.
Output	The output rectangular chunk is named as rect_*<row_begin>*_*<row_end>*_*<col_begin>*_*<col_end>*.csv stored in the *outputdir*. This is a csv file storing the chunk of the matrix XX^T.

calc_loadings_cov.exe

Description: The program calc_loadings_cov.exe calculates the loadings of PCA based on the covariance matrix

Usage: calc_loadings_cov.exe methyldir desdir workdir label number_of_PCs_loadings chrID

Argument	
methyldir	The directory of the methylation data
desdir	The directory where the calculated descriptives stores
workdir	The directory where eigenvalue and eigenvector files are. This is also the directory to store the loadings under the subdirectory “loadings”
label	The added label in the eigenvalue and eigenvector filenames
number_of_PCs_loadings	The number of loadings to be calculated corresponding to the top PCs
chrID	Specify which chromosome to be processed
Input	The inputs include the files of eigenvalues and eigenvectors (normalized PCs), as well as the methylation data.
Output	Loadings on the chromosomes are stored in the directory <*work_dir*>/loadings/: cov_loadings_chr<i>.csv: The loadings on the chromosome based on the covariance matrix, where <i> is the chromosome number

calc_loadings_cor.exe

Description: The program calc_loadings_cor.exe calculates the loadings of PCA based on the correlation matrix

Usage: calc_loadings_cov.exe methyldir desdir workdir label number_of_PCs_loadings chrID

Argument	
methyldir	The directory of the methylation data
desdir	The directory where the calculated descriptives stores
workdir	The directory where eigenvalue and eigenvector files are. This is also the directory to store the loadings under the subdirectory “loadings”
label	The added label in the eigenvalue and eigenvector filenames
number_of_PCs_loadings	The number of loadings to be calculated corresponding to the top PCs
chrID	Specify which chromosome to be processed
Input	The inputs include the files of eigenvalue and eigenvector (normalized PCs), as well as the methylation data.
Output	Loadings on the chromosomes are stored in the directory <*work_dir*>/loadings/: cor_loadings_chr<i>.csv: The loadings on the chromosome based on the covariance matrix, where <i> is the chromosome number

association_ols.exe

Description: The program association_ols.exe performs the association test adjusting for covariates.

Usage: association_ols.exe methyldir pheno_file outputdir test_label chrID

Argument	
methyldir	The directory of the methylation data
pheno_file	The data file with columns: sample ID, outcome, covariate1, covariate2, …, see the user guide
outputdir	The output directory
test_label	a string that characterizes the analysis and is appended to the output file
chrID	Specify which chromosome to be processed
Input	The methylation data and the phenotype data including outcome and covariates.
Output	The output file is named as association_<test_label>_ chr<chrID>. It outputs the t test statistics and corresponding p-values.

create_PCA_script.R

Description: The R script create_PCA_script.R creates scripts to calculate the chunks of the inner product matrix XX^T.

Usage: Rscript create_PCA_script.R diag_program rect_program methyldir samplefile desdir scriptdir outputdir n_sample n_sample_chunk n_cores n_parallel_proc start_chr end_chr

Argument	
diag_program	program to calculate diagonal chunks
rect_program	program to calculate the rectangular (off-diagonal) chunks
methyldir	The directory of the methylation data
samplefile	The sample file containing the sample IDs
desdir	The directory where the calculated descriptives stores
scriptdir	directory to store the scripts
outputdir	directory to store the computed chunks of XX^T
n_sample	The total number of samples
n_sample_chunk	number of samples allocated per chunk
n_cores	number of cores needed to load the number of samples specified in n_sample_job. Set it to the maximal number of cores in a node to avoid memory competition among cores in a single node
n_parallel_proc	number of jobs submitted to run in parallel
start_chr	The starting chromosome number to be processed
end_chr	The ending chromosome number to be processed
Input	The input are the paths of the programs calculating the chunks of the inner product matrix and related parameters
Output	The outputs of the R code are several scripts that contains commands to calculate the chunks of the XX^T

assemble_PCAmatrix.R

Description: The R script assemble_PCAmatrix.R assembles the chunks of the inner product matrix XX^T together

Usage: Rscript assemble_PCAmatrix.R chunksdir samplefile outputdir n_sample n_sample_chunk

Argument	
chunksdir	The chunks of XX^T to be assembled
samplefile	The sample file containing the sample IDs
outputdir	The output directory of XX^T
n_sample	The total number of samples
n_sample_chunk	number of samples allocated per chunk
Input	The input chunks are generated by either calc_PCAinput_rect.exe or calc_PCAinput_diag.exe. The parameter n_sample and n_sample_job must match the those in the create_PCA_script.R
Output	The output matrix is the full inner product matrix XX^T PCAinput_cov.csv, PCAinput_cor.csv and the matrix storing the diagonal elements sd_PCAinput.csv. See user guide for more details.

outputPC.R

Description: The R script outputPCs.R calculates the PCs and eigenvalues based on the input inner product matrix XX^T.

Usage: Rscript outputPC.R inputXXT pcscores eigenval eigenvec

Argument	
inputXXT	The input file of the inner product matrix XX^T
pcscores	The output file of the PC scores
eigenval	The output file of the eigenvalues
eigenvec	The output file of the normalized PC scores
Input	The input file is a csv file, with the subject IDs on the first row. The rest is a N × N matrix, where N is the number of the subjects.
Output	The output are as follows: 1. pcscores: This is a csv file. The first row and the first column are PC names and subject IDs respectively with the element on the first row and first column an empty string. 2. eigenval: This is a one column file, the first row is the header “Eigenvalues”, and the rest rows are the eigenvalues 3. eigenvec: This is the normalized PC scores with unit length. The format is similar to that of the pcscores.