GWAS for BACTERIA

Sarah Earle, Chieh-Hsi Wu and Daniel J. Wilson (2015)

SNP ANALYSIS

Usage example:

```
Rscript /path/of/SNP_GWAS_MAIN.R -dataFile dataFile.txt
-phylogeny phylogeneticTree.newick
-ref_fa referenceGenome.fasta -ref_gbk referenceGenome.gbk
-prefix outputPrefix –script_dir /gwasSourceCodes/snpGWAS/
-CFML_prefix cfmlOuputPrefix -run_gemma yes –PCA yes -
externalSoftware /path/of/externalSofware.txt
```

dataFile

This specifies the path of a tab-delimited text file containing the data of the isolates. The file contains three columns:

(1) The unique ids of the isolates
(2) The paths to the sequence data files of the isolates in fasta format
(3) The binary phenotype data of the isolates

The first row of the file must be the column names, namely id, filePath, and phenotype. Note that fasta files are compressed to gzip files. In addition these genomes have been mapped to the same reference genome.

An example of the contents in text file required for **dataFile** is shown below.

```
id          filePath                              phenotype
ecol1       /home/data/ecol/ecol1.fasta.gz        0
ecol2       /home/data/ecol/ecol2.fasta.gz        1
```

phylogeny

This specifies the path to a newick file that contains a phylogeny of all the isolates concerned in the GWAS studies. If this input is not provided, then a phylogeny is built from the sequence data provided for **dataFile**. If the number of isolates is less than 100, the phylogeny is built using PhyML, otherwise it is built using RAxML. Although the option of building the phylogenetic tree on the fly is available, it is strongly recommended that the users should provide the **phylogeny** input. This is because it is a good idea to check whether phylogeny has been properly reconstructed. If the phylogeny were built on the fly, there would not be an opportunity to check the phylogeny before resuming the rest of the GWAS analysis. The phylogeny is used for imputing missing values and therefore can influence the results from the GWAS analysis.

ref_fa

This specifies the path to the fasta file of the reference genome used to map the isolates concerned in this study.

ref_gbk
This specifies the path to the genbank file of the reference genome used to map the isolates concerned in this study.

prefix
This specifies the prefix of the output files.

script_dir
This specifies the path to the directory where the other required R scripts are located. Therefore all those R scripts must be in the same directory. The R source codes are all in the same folder (snpGWAS), which is inside the src folder. Therefore it would be easiest not to move the source codes and specify the path to the snpGWAS folder for **script_dir**.

CFML_prefix
This specifies the prefix of the output files produced by ClonalFrameML if the output files of ClonalFrameML have already been produced. The input value of **CFML_prefix** must be different to that of **prefix**. If **CFML_prefix** is not specified, then ClonalFrameML will be run and the output files will have the prefix as specified for **prefix**. The output files of ClonalFrameML are used for imputation.

run_gemma
This specifies whether or not to additionally run the analysis with linear mixed models using GEMMA (yes or no). The default value is no.

PCA
This specifies whether or not to additionally run a PCA analysis using Eigensoft. The default is no.

externalSoftware
This specifies the path to a tab-delimited file containing the name and paths of the external software packages used in the analysis. The tab-delimited file should contain two columns with headers *name* and *path*. The *name* column and *path column* respectively specifies the names and paths of the external software packages and non-R scripts used in the SNP analysis. The *name* column must contain the following:
ClonalFrameML
GEMMA
RAxML
PhyML
EigenSoft
ConvertPhylipToFasta
The *path* column contains the corresponding path to the software package. ClonalFrameML, GEMMA, RAxML, PhyML and EigenSoft are the names of the software packages that must be installed prior to the GWAS analysis.

ConvertPhylipToFasta refers to ConvertPhylipToFasta.class, which is in the same directory as the R scripts (snpGWAS).

KMER ANALYSIS

Usage example:

```
Rscript /path/of/KmerAnalysisMain.R -dataFile dataFile.txt
-srcDir  /gwasSourceCodes/kmerGWAS/ -prefix outputPrefix
-signif 100000  -genFileFormat BAM –minCov 5 –kmerFileDir
/home/kmers/ -nproc 5 –createDB –refSeqFasta1 refSeqFasta1.txt
–refSeqFasta2 refSeqFasta2.txt –ncbiSummary ncbiSummary.txt
–runLMM TRUE -relateMatrix relatednessMatrix.txt –signifLMM 100000
-externalSoftware /path/of/externalSofware.txt
```

or

```
Rscript /path/of/KmerAnalysisMain.R -dataFile dataFile.txt
-srcDir  /gwasSourceCodes/kmerGWAS/ -prefix outputPrefix
-signif 100000  -createKmerFiles TRUE –kmerFileDir /home/kmers/
-nproc 5 –createDB –refSeqFasta1 refSeqFasta1.txt –ncbiSummary
ncbiSummary.txt -db2 /path/to/db2 –runLMM TRUE
-relateMatrix relatednessMatrix.txt –signifLMM 100000
-externalSoftware /path/of/externalSofware.txt
```

dataFile

This specifies the path to a tab-delimited file containing the data of the isolates. The file contains three columns:

(1) The unique ids of the isolates
(2) The paths to the sequence data files of the isolates
(3) The binary phenotype data of the isolates

The first row of the file must be the column names, namely id, filePath, and phenotype. If the user intends to create the kmer files from bam files then the sequence data files would be BAM or FASTA files. The filePath column would therefore contain a list of BAM (or FASTA) file paths.

The following is an example of the contents in the data file required when the user intends to create the kmer files from BAM files.

```
id        filePath                          phenotype
ecol1     /home/data/ecol/ecol1.bam         0
ecol2     /home/data/ecol/ecol2.bam         1
```

If the kmer files have already been created, then the filePath contains a list of paths to gzipped kmer files. The following shows an example of the contents in the data file required when the user already has the kmer files at hand.

```
filePath                          phenotype
/home/data/ecol/ecol1.kmer.txt.gz  0
/home/data/ecol/ecol2.kmer.txt.gz  1
```

In the phenotype column, 1 and 0 respectively denote the presence and absence of the trait of interest.

srcDir
This specifies the path of the directory that contains all the other required R scripts and therefore all those R scripts must be in the same directory. The R source codes are all in the same folder (kmerGWAS), which is inside the src folder. Therefore it would be easiest not to move the R source codes and specify the path to the kmerGWAS folder for **srcDir**.

prefix
This is the prefix of the output files created from the kmer-based GWAS analysis.

signif
This is an integer input, which specifies the number of top significant kmers to be annotated. The default value is 10000.

genFileFormat
This requires a String input, which is the file format of the genomic data. The accepted values are BAM, FASTA and KMER. If the user has specified BAM or FASTA, then the kmer files will be generated for the association tests. If the kmer files have already been generated, then the user should specify KMER for this input. The file format specified here must be consistent with the files specified in the data file (for input **dataFile**). E.g. if the user has specified BAM here, the data file should contain the bam file paths (see **dataFile**).

minCov
This specifies the minimum depth of the kmers. If the genomic data is based on assemblies then this should be 1. The default value is 5.

kmerFileDir
This specifies the path of the directory where the kmer files, that are to be created from bam files, will be located. The default is current working directory.

nproc
This is an integer input, which specifies the number of CPU processors used for creating kmer files and kmer annotation. The default is 1 processor.

createDB
This requires a Boolean input, which specifies whether or not to create the blast databases for annotation. The input value must be either TRUE or FALSE. The default value is TRUE.

Setting createDB to TRUE
With this setting the user can create up to two nucleotide blast databases and convert the NCBI summary file to the required format for kmer annotation. The NCBI summary file must be provided (see more details below on **ncbiSummary**). If the

user would like to create only one blast database, then the input of **refSeqFasta1** must be provided (see more details below on **refSeqFasta1**). If the user would like to create two blast databases, then the inputs of **refSeqFasta1** and **refSeqFasta2** must be provided (see more details below on **refSeqFasta2**).

Setting createDB to FALSE
With this setting the user must specify all the inputs for **db1**, **db2** and **ncbiAnnot**.

refSeqFasta1
This specfies the path of a text file containing a list of paths to all the genome sequences to be used to create the first database for kmer annotation. The genome sequences are in fasta file format.

To create this list, go to http://www.ncbi.nlm.nih.gov/guide/howto/dwn-genome/ and then click on Genomes FTP site. Click on guest when the box pops up, then click on connect. Go into the folder of your species, then download the nucleotide fasta files for your species. The **refSeqFasta1** is a text file with a single column containing paths to all of these fasta files.

refSeqFasta2
This specifies the path of the file that is in the same format as the input for **refSeqFasta1**. This file is used to create the second blast database and the genome sequences used should be different to those used to create the first database.

ncbiSummary
This specifies the path to the summary file downloaded from the NCBI database for the genes of a given species. To create this summary file,

(1) Go to NCBI gene (http://www.ncbi.nlm.nih.gov/gene/).
(2) Put the species of interest in the search bar, e.g. *S. aureus*
(3) When the search has been completed and matches have been loaded, at the top right it should say "Send to:". Click "Send to:".
(4) Click "File".
(5) Choose "Summary (text)".
(6) Click "Create File".

db1
This is the path to the first blast database used for kmer annotation. This must be specified if **createDB** is set to FALSE or **refSeqFasta1** is not specified. The database would have been created using makeblastdb. makeblastdb would have created several files with the same name having different extensions. For **db1**, specify the path of the output files created by makeblastdb but exclude the extensions.

For example, let the following be the paths to the files created by makeblastdb for the first blast database:
/home/db/example1.nhr

/home/db/example1.nin
/home/db/example1.nsq

/home/db/example1 would then be the input for **db1**.

db2
This specifies the path to the second blast database used for kmer annotation. This must be specified if **refSeqFasta2** is not specified. The specification of this option follows **db1**.

ncbiAnnot
This specifies the path to the NCBI annotation file created from the NCBI summary file.

runLMM
This requires a Boolean input, which specifies whether or not to run the analysis with linear mixed model (LMM). The input value must be either TRUE or FALSE. The default value is true.

relateMatrix
This specifies the path to a text file containing the relatedness matrix created from GEMMA. The number of rows and columns of the matrix should be the same as the number of isolates. If **runLMM** is TRUE, then the relatedness matrix must be provided.

signifLMM
This requires an integer input that specifies the number of top significant kmers to be used for the LMM analysis. The default is 100000.

externalSoftware
This specifies the path of a tab delimited file containing two columns with headers *name* and *path*. The *name* column and *path* column respectively specifies the names and paths of the external software packages and non-R scripts used in the analysis.

The *name* column must contain the following:
parallel
samtools
samToFastq
trimmomatic
trimmomaticPE
shuffleSequencesFastq
dsk
blastn
makeblastdb
GEMMA

The *path* column contains the corresponding path to the software package. Table 1 presents the software package to which each name refers.

Table 1: External software packages used in the kmer-based GWAS analysis.

name	**Software package**
parallel	GNU parallel
samtools	SAMtools
samToFastq	Picard command line tool SamToFastq.jar
trimmomatic	Trimmomatic
trimmomaticPE	The Trimmomatic adaptor file combined-PE.fa
shuffleSequencesFastq	shuffleSequences_fastq.pl in the software package velvet
dsk	DSK
blastn	BLAST
makeblastdb	BLAST
GEMMA	GEMMA

All the software packages mentioned in Table 1 must be installed prior to the GWAS analysis.

PrintOutTopXChisq , sortDsk and gwasKmerPattern refers to compiled scripts not re-written in R. These are in the same folder as the R source scripts. Table 2 presents the file names for those non-R scripts.

Table 2: The file name of the non-R scripts.

name	**Software package**
PrintOutTopXChisq	PrintOutTopXChisq.class
sortDsk	sort_dsk
gwasKmerPattern	gwas_kmer_pattern