Contents

1. Installation:

Please run **"PreMagnetConfig.sh"** before running MAGNET.

Currently we provide configuration options for "SLURM" or "PBS" cluster management system with at least 128GB of memory. Please change in the configuration file if you have any other job management system. At the end of the "PreMagnetConfig.sh" your job script based on any one of these systems will be initialized.

To change number of nodes, memory, time or other parameters please update in the "Thresholds.config" file in configuration folder. This file also contains information which stages of MAGNET the user wants to perform. Please enter Yes/No to "PerformStage1", "PerformStage2" and "PerformStage3" variables.

2. Prerequisites

MAGNET can be run on any 64-bit x86 Unix based system on a SLURM/PBS node management system. MAGNET can be freely downloaded from the following website https://github.com/SheenYo/MAGNET. Please specify a directory where MAGNET will run.

The pipeline uses the following tools which are needed to be installed prior to running i.e. **R>=3.5**, **Perl**, **Python** and **unzip** utility for linux. Make sure these software are in your PATH variable (bash profile). Other tools which would be used for running MAGNET are listed below. In case, these are not installed the pipeline will install it automatically:

- **GO_Elite**
- **liftOver**
- **Magma**
- **Minimac3**
- **Plink v1.9**
- **SHAPEIT v2.727**

Once you have run the **"PreMagnetConfig.sh"** file, missing tools will be installed and configured automatically in the configuration files.

Before running MAGNET please make sure that the genotype file should be in "ACGT/1234 plink format" only

3. Stage 1: Genotype QC (Stage1_GenoQC.sh)

3.1 Input files:

- ***SamplesToQC***

The user needs to provide plink formatted (.bed, .bim, and .fam) input files which are required to be ACGT/1234 allele coded, otherwise the program will exit. The program checks for their existence and exits if the files are not found. By default the path is set to our "Example data", please change the path where your actual study data is available in plink format.

- ***AffectedInds***

Plink normally refers '-9' as the unaffected individuals and '2' as affected individuals, in case your affected statuses are otherwise defined please provide a list of affected individuals and its respective path in the configuration file.

- ***DuplicatedIds:***

By default, MAGNET will remove one of the duplicated individuals, if you wish to keep specific individuals please provide the list of duplicate individuals in config file in the variable "DuplicatedIds".

3.2 Input parameters:

- ***installationDir***

By default the directory where the program is installed is considered as the installation directory, all further output folders will be created within this directory. This directory can be changed in the ToolsConfig e.g. installationDir=/home/user/mypath

This section of the program consists of the following defined variables that can be changed by the user in the MAGNET/ConfigFiles/Thresholds.config

- ***GENO***

Refers to plink missing genotype call rate threshold. By default, we have set it to 0.05, all variants exceeding this threshold will be omitted from the analysis.

- ***HWE***

Defines the threshold which filters out all variants that have Hardy-Weinberg equilibrium exact test p-value below it, which is by default set to 10e-8. For a detailed overview please see the publication Yousaf et al 2018 (https://doi.org/10.1101/336776).

- ***MAF***

Filter variants with a minor allele frequency less than a specific threshold. By default, it is set to 0.02. User can change it based on sample size and study design.

- ***MIND***

Excludes all samples which have missing genotype greater than a specific threshold, here it is by default set as 0.05.

- ***MEFam and MESNP***

These options are used for family-based data only. Filters individuals and/or markers based on the mendel error rate.

MEFam: By default, discards all the families with more than 1% Mendel errors (considering all SNPs).

MESNP: By default, discards all the SNPs with more than 10% Mendel error rate based on number of trios.

3.3 Reference Files

Please install the following reference file in the “RefData” folder:

- ***Hapmapfiles:*** Hapmap genotype file for ethnicity plots can be downloaded from http://zzz.bwh.harvard.edu/plink/dist/hapmap_r23a.zip

```
cd RefData

wget http://zzz.bwh.harvard.edu/plink/dist/hapmap_r23a.zip

unzip hapmap_r23a.zip

wget http://zzz.bwh.harvard.edu/plink/dist/hapmap.pop

echo -e "FID IID Population" | cat - hapmap.pop > Hapmap_siteinfo.txt
```

- ***Siteinfo:*** The file “Hapmap_siteinfo.txt” will be provided as the site information file.

3.4 Output

The program reports the quality of user provided data by generating plink based reports, results from these reports are generated as plots at the end of the analysis:

- ***QC1_report.imiss***

A list reporting sample based missingness, where F_MISS column details the missing call rate.

- ***QC1_report.lmiss***

A list reporting variant based missingness, where F_MISS column details the missing call rate.

- ***PreQC_hardy.hwe***

A statistic generated showing the Hardy-Weinberg exact test with p-values.

- ***PreQC_AlleleFreq.frq***

Lists all the SNPs with minor allele frequencies

- ***PreQC_Inbreeding.het***

Generates the inbreeding coefficient estimates

- ***QC2_Sexcheck.sexcheck***

Reports sex determined based on X chromosome data and provides a detail account of individuals for whom the reported sex in the PED file does not match the estimated sex in the genotype data. By default all individuals with reported sex discrepancies will be removed, in case you wish to include them in the analysis then please correct these gender inconsistencies and rerun the analysis.

- ***FinalQC_Study***

At the end the user gets a clean quality checked plink data file, which can be used for the next stage.

4. Stage 2 Imputation (Stage2_Imputation.sh)

Perform imputation of affected individuals for chromosomes 1-22, using 1000 Genomes data phase 3 as reference panel.
Convert the files for minimac into VCF

4.1 Input Files

- ***QualityCheckedFile*** Final quality check file (FinalQC_Study.bed, FinalQC_Study.bim, and FinalQC_Study.fam). For standalone imputation analysis, please provide path of your QC plink formatted file (bed, bim, fam).

- ***AffectedInds:***

In case your sample consists of unaffected and affected individuals, please provide list of family and individual IDs of affected individuals only. Since GWAS analysis here is based on affected individuals only.

4.2 Input parameters:

- ***LIFTOVER:***

A variable requiring information if LIFTOVER should be performed to change the genome build. If your data is already in hg19 format, then please answer as “No”.

- ***ChainToChoose:***

By default, Chain file from hg18 to hg19 is selected to update the genome build. In case your genome build is not hg18 please refer to following section.

- ***ExtractAffected:***

If the individuals provided in the data are already only affected individuals then set the variable as "No", but if you want MAGNET to extract it then by default it will look in the *.fam file and look at the phenotype value column (column no.6) to select “2” as affected and “1” as non-affected (as per plink standards)

- ***thresholdImp:***

Imputation quality threshold value of Rsq, default to 0.3.

- ***chunkSize:*** Number of SNPs to be present in each SNP raw data file, by default each SNP raw data file consists of 5000 SNPs

4.2 Reference Files

Please download the following reference files in “RefData” folder

- ***Hg19SNPs:*** Hg19 SNPs file containing SNPs rsids, starting and ending bp position, chromosome number. The user can individually download the files and merge them using the following linux code:

```
for ((i =1;i<=22;i++))
do
wget    ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b150_GRCh37p13/BED/bed_chr_"$i".bed.gz
done
```

*chmod 770**

gunzip bed_chr_*.bed.gz

cat bed_chr_*.bed>SNPs_all.bed

Since the 1000 genomes dataset is based on the UCSC hg10 genome build, the study dataset needs to be updated to this genome build. All chain files are provided in the reference data, which you can also download using the following links with respect to your data genome built. In case your genome build differs from hg18 (by default) please change the variable "ChainToChoose" in the "Tools.config".

Chain files can be downloaded from the provided links and can be saved in the RefData folder. In case these files are saved at any other location please update the path accordingly in the Tools.config.

- ***Chain16To19:*** liftOver chain files to convert hg16 SNPs to hg19 annotation

  http://hgdownload.cse.ucsc.edu/goldenPath/hg16/liftOver/hg16ToHg19.over.chain.gz

- ***Chain17To19:*** liftOver chain files to convert hg17 SNPs to hg19 annotation http://hgdownload.cse.ucsc.edu/goldenPath/hg17/liftOver/hg17ToHg19.over.chain.gz
- ***Chain38To19:*** liftOver chain files to convert hg38 SNPs to hg19 annotation http://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz
- ***Chain18To19:*** liftOver chain files to convert hg18 SNPs to hg19 annotation

  http://hgdownload.soe.ucsc.edu/goldenPath/hg18/liftOver/hg18ToHg19.over.chain.gz

- ***MapFile:*** Genetic map for 1000GP_Phase3 consisting of three columns containing the physical position (bp), the recombination rate (cM/Mb) and the genetic position (cM). The file can be downloaded from http://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3/

  If you choose to save the file at any other destination than RefData folder, please update the path for "MapFile" variable in Tools.config.

The following three Shapeit reference files can be downloaded from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.tgz.

Unzip the folder using the following command:

gunzip 1000GP_Phase3.tgz

The folder will contain files in the following three formats:

- ***ShapeitRefHaps:*** The file consists of SNPs and the haplotypes where each line corresponds to single SNP consisting information about the two alleles of a SNP by each haplotype of an individual about the chromosome number, SNP id, SNP position, and the first and second allele.
- ***ShapeitRefLegend:*** The file describes the SNPs, where the columns correspond to SNP id, SNP position, first and second allele.
- ***ShapeitRefSample:*** The file contains reference individuals information i.e. Individual ID, population, group and sex.

If you choose to save these files in any other directory, please update the path in ToolsConfigFile.

- ***AnnotationFile:*** The following lines of code will help you get the dbSNP 142 files:

  (i) First create a directory named “AnnotationFiles_SNP151”.

  ```
  mkdir AnnotationFiles_SNP151
  ```

  (ii) Download the latest SNP build from UCSC genome browser using the following commands.

  ```
  cd AnnotationFiles_SNP151

  wget http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/snp151.txt.gz
  ```

  (iii) Unzip the file

  ```
  gunzip -k snp151.txt.gz
  ```

  (iv)Extract chr, bp and snp id

  ```
  awk '{print $2 " " $4 " " $5}' snp151.txt> Annotation_file.txt
  ```

  (v) Divide base on chromosome number

  ```
  for((i=1;i<=22;i++))

   do

   grep "chr"$i" """ Annotation_file.txt>Annotation_chr"$i".txt

  wait

  awk -F, '{NF=4}1' OFS="\t" Annotation_chr"$i".txt > Annotation_chr"$i"_extraC.txt

  wait

  sed -e 's/chr//g'  Annotation_chr"$i"_extraC.txt|awk '{print $1 " " $2 " " $3 " " $4 " " $1":"$2}'>Comp_chr"$i".txt

  wait
  ```

```
echo $'chr bp_pos snp_name chr:bp' | cat - Comp_chr"$i".txt>Comp_chr"$i"_wHead.txt

done
```

- This folder consists of files annotated based on SNP151 genome build separated based on chromosomes, each file consists of chromosome number, base pair position, snp name and chromsome:base pair.

4.3 Output Files

- ***Gwas.Chr (1-22)_Study.Imputed.Output.dose****

Contains allele dosages for imputed and genotyped SNPs

- ***Gwas.Chr(1-22)_Study.Imputed.Output.erate***

Contains estimated error rate for every imputed and genotype marker.

- ***Gwas.Chr(1-22)_Study.Imputed.Output.info***

Contains information on both genotyped and impute SNPs

- ***Gwas.Chr(1-22)_Study.Imputed.Output.m3vcf.gz***

Output file in vcf format

- ***Gwas.Chr(1-22)_Study.Imputed.Output.rec***

Recombination file, contains switch error rate per interval

- ***Gwas.Chr(1-22)_Study.Imputed.Output.hapDose.gz***

Contains dosage for each haplotype separately

- ***Gwas.Chr(1-22)_Study.Imputed.Output.hapLabel.gz***
- ***Merged_FinalQC_SNPs_Data (bed, bim, fam):***

Plink formatted files consisting of Rsq >0.3 filtered and biallelic SNPs

- ***Data_SNPfile (1-22)***

Raw files with minor alleles each consisting of 5000 imputed SNPs.

5. Stage 3 Genome wide association study

5.1 Input file

- ***SamplesToQC***

In case the user wants to only perform GWAS analysis, then the plink files set for the variable SamplesToQC will be used.

- ***Data_SNPfile*.raw***

Raw files consisting of 5000 SNPs each

5.2 Input parameters:

- ***Phenotype_Status***

For Regression analysis select the individuals phenotype status that will be selected for regression. As per plink the valid values are -9, 2 and 1.

- ***ImputeData***

A variable to know if the imputed data from stage 2 will be used, in case the user performed imputation using MAGNET then set this variable to “Yes” else “No”.

- ***Pheno:*** Name of the phenotype to be analysed, required for naming of the results
- ***ColManhattan:*** Color to be used for manhattan plot in contrast, default is black color.

Further, the user needs to specify five arguments in the ConfigFiles/Config.R i.e:

- ***windowSize:*** The default window size is 5kb upstream and downstream of gene
- ***Covars:*** Name of covariates included in the regression model, please note that the names should correspond to those in the phenotype file provided, e.g by default “Sex,Age” are the two covariates provided, these names are same as in the phenotype file. In case you have other covariates, or the name is written differently than please update in the Thresholds.config file.
- ***Fixed:*** Names of random covariates included in the regression model, please note that the names should correspond to those in the phenotype file provided as explained above.
- ***snpsOfInterest:*** If you are looking for specific set of SNPs of interest than please provide their names in the “Thresholds.config” file e.g. "rs377398625","rs557375998","rs10736578”
- ***MagmaN.*** Sample size (Number ofindividuals) for which Magma would be conducted
- ***MagmaPERMP:*** Permuted p-value for MAGMA analysis, by default set to 0.05

5.3 Reference Files

Please make sure to provide all these files in "RefData" folder

- ***phenofile:*** Please provide a reference file named ""phenofile" consisting of Family ID, Individual ID and phenotype of interest
- ***MagmaRef*** 1000 genomes plink formatted reference file to be downloaded from MAGMA website: https://ctg.cncr.nl/software/MAGMA/ref_data/g1000_eur.zip
- ***MagmaSNPloc:*** The file contains Chromsome, SNP identifier, Position in morgans or centimorgans (mostly dummy coded as "0") of the reference 1000 genome dataset.
- ***MagmaGeneloc:*** The file consist location information about genes in NCBI37 detailing Entrez gene id, chromosome id, start and stop position, strand information and gene. The file can be downloaded using the following link:

  https://ctg.cncr.nl/software/MAGMA/aux_files/NCBI37.3.zip

5.4 Output files

- ***Results file***: For each chunk regression output will be saved in this file consisting of beta, se (stand error), t, p-value and adjusted p-value
- ***Summary file:*** For each chunk regression output summary will be saved in this file
- ***Magma Result file:*** File consisting of complete magma gene lists will be provided in the Magma_pheno.genes_tabseparated.txt file.
- ***Pheno.SignificantGenes:*** List of significant genes resulting from MAGMA analysis below the default threshold of empirical p-value of 0.05

  Please recheck the prg*.out files to see if there are any warnings reported for the regression.

6. Stage 4 Enrichment

6.1 Input file

- ***GenelistProvided:*** If the user is performing only Stage 4 then please provide the path for gene list else the significant genes from MAGMA analysis will be selected for the analysis provided in the variable ***GenelistFromStage3***
- ***Pheno.SignificantGenes:*** List of significant genes resulting from MAGMA analysis that are below the default threshold of empirical p-value of 0.05
- ***Magma Result file:*** File consisting of complete magma gene lists will be provided in the Magma_pheno.genes_tabseparated.txt file.

R Pre-requisites

Requires R-packages "org.Hs.eg.db, annotate, WGCNA, igraph and purrr". These will be automatically downloaded upon script runtime.

Python Pre-requisites

Please make sure that you have Python 2.7 or higher installed along with the following dependencies:

setuptools

Tkinter

scipy

If you are an ubuntu user, please use the following commands to install:

```
sudo apt-get update
sudo apt-get upgrade
sudo apt-get install python-setuptools
sudo apt-get install python-tk
sudo apt-get install python-pip
sudo apt-get upgrade
sudo pip install --upgrade pip
sudo apt-get install python-tk
sudo apt-get update
```

6.2 Input parameters

- ***GOeliteSpecies:*** Since the pipeline is based for human data only, by default the option is set to “Hs” for homosapiens.
- ***GOElitefolder:*** Name of folder where GO elite inputs are stored
- ***Outputfolder:*** Name of output folder where results of downstream analysis will be stored
- ***MAGNETHome:*** Name of the main directory where MAGNET is installed

6.3 Reference Files

- ***Kang Universe:*** All genes from Kang dataset
- ***Kang genes:*** All genes which are in Kang modules
- ***KangData:*** File consisting of Kang expression data

6.4 Output files

- ***All_MagmaGenes_Symbols.txt:*** All MAGMA ENTREZ genes mapped with their gene symbols
- ***All_genesMAGMA_Imputed_Kang.txt:*** All genes which are present in MAGMA and are present in Kang dataset as well.
- ***phenoEnrichment_ouput.txt:*** Output of enrichment analysis
- ***phenoGO_ElitePlot.pdf:*** Plots of top ten GO-terms
- ***phenoKEGG.pdf:*** Plots of top ten KEGG pathways
- ***phenoEnriched_Modules.pdf:*** Gene network plots of enriched modules
- ***Heatmap_EnrichedModules.pdf:*** Heatmap of enriched modules
- ***Cross sectional and sagittal Brain views:*** Eigen gene values of module genes in Kang dataset would also be presented with respect to brain specific region