Table of Contents

[EXTERNAL TOOLS DEPENDENCIES 2](#_Toc351637277)

[EQTL WORKFLOW 3](#_Toc351637278)

[INPUT FILES 3](#_Toc351637279)

[GENOTYPING DATA 3](#_Toc351637280)

[GENE EXPRESSION 3](#_Toc351637281)

[RUNNING THE WORKFLOW 4](#_Toc351637282)

[SAMPLE EXECUTION 4](#_Toc351637283)

[OUTPUT FILES 4](#_Toc351637284)

[RLM MODULE 4](#_Toc351637285)

[SAMPLE EXECUTION 4](#_Toc351637286)

[INPUT ARGUMENTS 4](#_Toc351637287)

[FOLDCHANGE 4](#_Toc351637288)

[ANNOTATION MODULE 4](#_Toc351637289)

[INPUT ARGUMENTS 4](#_Toc351637290)

[SAMPLE EXECUTION 4](#_Toc351637291)

[OUTPUT FILE 4](#_Toc351637292)

[LD BLOCK MODULE 4](#_Toc351637293)

[SAMPLE EXECUTION 4](#_Toc351637294)

[INPUT ARGUMENTS 4](#_Toc351637295)

[OUTPUT 4](#_Toc351637296)

[You can find out two output files in the temp directory. 4](#_Toc351637297)

[Downloading External Tools (TOOLINFO FILE) 4](#_Toc351637298)

[EXAMPLES 4](#_Toc351637299)

This manual helps the users to run the EQTL workflow. The EQTL workflow is a modular set of scripts that should be run in the following sequence.

1. [EQTL WORKFLOW](#EW) : To run the EQTL , permuted gene expression EQTLs, apply gene expression, apply SNP filtering and generate the pvalue Q-Q plots.
2. [ANNOTATION MODULE](#AM) : To get annotations for the SNPS using the ANNOVAR tool.
3. [LD BLOCK MODULE](#LBM) : To calculate LD blocks from the significant CIS and TRANS EQTLs using PLINK tool.

# EXTERNAL TOOLS DEPENDENCIES

A number of external tools need to be installed.

plink, ANNOVAR & R packages.

We provide a binary whenever the license terms allow us, but you may want to download the latest version and copy them in the bin directory. An install script is provided to install the tools and a script is provided to generate the tool information config file for the scripts to use.

1. [Downloading External Tools (TOOLINFO FILE)](#TOOLINFO)

Abbreviation table:

# EQTL WORKFLOW

Eqtl workflow requires Gene expression and Genotyping input data.

## INPUT FILES

### GENOTYPING DATA

Tped & tfam files

The genotyping input data for the scripts should be plink transpose format.

Transposed fileset contains two text files: one (TPED) containing SNP and genotype information where one row is a SNP; one (TFAM) containing individual and family information, where one row is an individual.

The first 4 columns of a TPED file are the same as a standard 4-column MAP file. Then all genotypes are listed for all individuals for each particular SNP on each line. The TFAM file is just the first six columns of a standard plink PED file.

Example

*<------------- trans.tped ------------->* *<- trans.tfam ->*

1 snp1 0 5000650 A A A C C C A C C C C C 1 1 0 0 1 1

1 snp2 0 5000830 G T G T G G T T G T T T 2 1 0 0 1 1

3 1 0 0 1 1

4 1 0 0 1 2

5 1 0 0 1 2

6 1 0 0 1 2

This kind of format can be convenient to work with when there are very many more SNPs than individuals (i.e. WGAS data).

To read a transposed fileset, using plink

plink --tfile mydata

http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#tr

#### PLINK COMMANDS

1. Convert BED(Binary plink files) to TPED (Transpose plink files)

Plink –bfile <input binary file> --recode --transpose –out <output transpose file>

1. Convert PED/MAP(Ped plink files) to TPED/TFAM (Transpose plink files)

Plink –file <INPUT PED FILE> --recode --transpose –out <OUTPUT TRANSPOSE FILE>

### GENE EXPRESSION

The gene expression input data for the scripts consists of two files

1. Gene expression file: Tab delimited file with first column is the PROBE, ENTREZ or GENE SYMBOL ID and labels from column 2 represent samples. Numeric values should be present from column2.

Example :

Entrezid Sample\_1 Sample\_100 Sample\_102 Sample\_103

1 12.5 21 14.9 10.5

10 8 10.4 7.1 1.7

100 70.1 63 53.5 30.3

1000 1825.1 2247 2165.7 2682.9

1. Gene expression location file: This is location file for the gene expression. It should contain
2. Gene/Probe ID : Column 1 should contain gene or probeset id same as gene expression file.
3. Chrm\_probe : Column 2 should contain chromosome( values same as TPED file (column 1))
4. S1: Probe/Gene start
5. S2: Probe/Gene stop

ID, Gene symbol, Chr, start, end

Example:

geneid chrm\_probe s1 s2

1 19 58858171 58864865

10 10 52559168 52645435

100 10 52559168 52645435

1000 10 52559168 52645435

1. Covariates file (OPTIONAL): This is the optional file. You can use covariates for EQTL analysis.

Row 1 should contain sample id’s just like gene expression file and from row 2 covariates should be included. Covariates should be numeric values.

Example:

id Sample\_1 Sample\_100 Sample\_102 Sample\_103

gender 0 0 1 1

age 36 40 46 65

1. Gene annotation file (OPTIONAL):: This is annotation file for the gene expression. The provided annotations will be attached to the gene summary file. This will make end user to understand the results.

Example:

GeneID GeneName

2315554 TTLL10

2315633 B3GALT6

2315674 SCNN1D

2315739 PUSL1

2315894 VWA1

2315918 ATAD3C

**POINTS TO REMEMBER**

1. Same number of samples and same sample names should be present in both gene expression and genotyping data i.e genotyping tfam file(column 2) and gene expression matrix file (row 1).
2. Samples may be in different order. The workflow will automatically reorder genotyping, gene expression and covariates files to match samples in each file.
3. Please provide unique samples ids in the tfam files (column 2).
4. Convert your platform specific marker Id’s to rsid’s based on platform annotation file.
5. Set mother and father id to missing (plink dataset: 3rd and 4th column in the tfam/fam file, i.e. the columns should be 0 in plink).
6. No duplicate RSid ‘s in the input dataset.
7. Make sure your input dataset doesn’t have duplicate positions for the same chromosome.
8. TPED file should be sorted according to chromosome and position.

### RUNNING THE WORKFLOW

Following are the steps to run the EQTL workflow. Create the config file with below mandatory parameters (optional parameters are in red)

**RUN INFO CONFIG PARAMETERS**

#Plink files for genotypes

tped=/data4/bsi/RandD/Workflow/EQTL\_WORKFLOW/sampledata/193sgenome.tped [Location of the genotyping TPED file]

tfam=/data4/bsi/RandD/Workflow/EQTL\_WORKFLOW/sampledata/193sgenome.tfam [Location of the genotyping TFAMfile]

#presently no files are transferred to output files

output.dir=/data4/bsi/RandD/Workflow/EQTL\_WORKFLOW/OUTPUT [location to the output folder where output files are located]

#Temporary files are created in the tempfolders (currently all output files are created here)

temp.folder=/data4/bsi/RandD/Workflow/temp[location to the TEMP folder where temp files are created and once the EQTL run is successful you can transfer the required files and delete the folder. If you want to run the RLM module, please don’t delete the generated temp folder in the EQTL analysis. RLM module will be using files “SNP.txt”and “expression.txt” in the temp folder.]

#Expression file location

expr.data=/data4/bsi/RandD/Workflow/EQTL\_WORKFLOW/sampledata/GE1.txt [location to the input gene expression file]

#Expresion file na sysmol

expr.na.symbol=NaN [How NA’s are denoted in the gene expression file]

#Genelocation file for the Matrix EQTL

expr.loc=/data4/bsi/RandD/Workflow/temp/chen\_wang/geneloc.txt [Location to the gene expression annotation file]

#Just need summary then set it to YES

only\_summary=no [When you turn on this option the workflow will stop after generating the summary reports for the genotyping and gene expression input files]

#SGE queue name

cluster.queue=1-hour [Queue you want submit for EQTL jobs.If you are not running on SGE then you can fill any value]

#sge memory required

cluster.max.mem=3g[Expected memory taken by your EQTL jobs. If you are not running on SGE then you can fill any value]

#email id  (SGE jobs)

email[=prodduturi.naresh@mayo.edu](mailto:=prodduturi.naresh@mayo.edu) [Fill in user’s full e-mail address]

#Number of permutation required

num.permutation=100[Number of permutation required to calculate the FDR. For no permutations enter 0]

#CIS PVAL cutoff for the EQTL runs

cis.pval=0.05[CIS PVAL cutoff for the EQTL runs]

#TRANS PVAL cutoff for the EQTL runs

trans.pval=0.01[ TRANS PVAL cutoff for the EQTL runs]

#SGE=YES if you are running on the sun grid Engine SGE or SGE=NO

sge=yes[if you are running on the sun grid Engine SGE or SGE=no]

run.dir=PROJECT NAME [INNER\_DIR is the inner directory in the temp folder to differentiate multiple projects]

covariate.data=[PATH TO THE COVARIATES FILE (Optional)]

#Filter option

filtering=yes[If you want to filter the EQTl results(SNP & genes/probes) based on the below filter options]

fil.min.num.gnt=5 [Filter genotype: Minimum number of genotypes in two out of three categories Reference homozygote, Heterozygote, Alternate allele homozygote] (optional: must specify if the filter parameter FILTER=YES)

#genotyping NA max 1

fil.max.na.gnt=0.5 [Filter genotype: Filters marker with NA percentage [Min value: 0 Max value:1] (percentage/100] less than specified] (optional: must specify if the filter parameter FILTER=YES)

#genotyping maf  max 1

fil.maf.cutoff.gnt=0.7[Filter genotype: Filters marker with Minor allele frequency less than specified] (optional: must specify if the filter parameter FILTER=YES)

#filteroption :max gene expression

fil.max.mean.geneexp=10000 [Filter gene expression: Filters probes/genes with mean gene expression upper limit] (optional: must specify if the filter parameter FILTER=YES)

#filteroption : min gene expression

fil.min.mean.geneexp=-10000 [Filter gene expression: Filters probes/genes with mean gene expression lower limit] (optional: must specify if the filter parameter FILTER=YES)

# filteroption :min sd gene expression

fil.min.std.geneexp=0 [Filter gene expression: Filters probes/genes with standard deviation of gene expression lower limit] (optional: must specify if the filter parameter FILTER=YES)

#filter option:max standard deviation gene expression

fil.max.std.geneexp=10000 [Filter gene expression: Filters probes/genes with standard deviation of gene expression upper limit] (optional: must specify if the filter parameter FILTER=YES)

#filter option: geneexpression NA max 1

fil.max.na.geneexp=0.1 [Filter gene expression: Filters probes/genes with NA percentage [Min value: 0 Max value:1] (percentage/100] less than specified] (optional: must specify if the filter parameter FILTER=YES)

cisDist=100000[distance from marker to the gen to consider as cis eQTL]

geneannot.attach.genesumm=/data4/bsi/RandD/Workflow/EQTL\_WORKFLOW/gene\_annot.txt[gene annotation file]

**TOOL INFO CONFIG PARAMETERS**

#PATH TO RSCRIPT EXECUTABLE

path\_rscript=[ PATH TO RSCRIPT EXECUTABLE]

# PATH TO PERL EXECUTABLE

path\_perl=[ PATH TO PERL EXECUTABLE]  
# PATH TO SH EXECUTABLE

path\_sh=[ PATH TO SH EXECUTABLE]  
# PATH TO QSUB EXECUTABLE

path\_qsub=[ PATH TO QSUB EXECUTABLE]

# PATH TO PLINK EXECUTABLE

path\_plink=[ PATH TO PLINK EXECUTABLE]

# PATH TO R PACKAGES LIBRARY

path\_rlib=[ PATH TO DIRECTORY WHERE R PACKAGES ARE INSTALLED]

## SAMPLE EXECUTION

Once you create the config .you can execute the script in the following

|  |
| --- |
| perl $INSTALLDIR/ perl\_workflow\_eqtl.pl -run\_config <run.config > -tool\_config <tool.config> |

## OUTPUT FILES

“Gene\_Summary”-> Gene Expression summary file for the expression. Below are the columns

1. Expr\_ID : unique Probe/Gene ID
2. mean : Gene expression Mean value
3. SD : Gene Expression standard deviation
4. N\_NA : Number of NA’s/Total number of genes/probes

“ PVAL.CIS.gz”-> Below are the columns

1. SNP : SNP ID
2. gene : GENE/PROBE ID
3. t-stat : T Statistic
4. p-value : P Value
5. FDR : False Discovery Rate

“ PVAL.TRANS.gz”-> Below are the columns

1. SNP : SNP ID
2. gene : GENE/PROBE ID
3. t-stat : T Statistic
4. p-value : P Value
5. FDR : False Discovery Rate

“ Table\_nofiltering.txt”-> Below are the columns

1. Pval\_cutoff : P Value range
2. N\_cis\_p1: Number of CIS EQTLs less than Pvalue cutoff in permutation 1

.

.

1. N\_cis\_p(n): Number of CIS EQTLs less than Pvalue cutoff in permutation n
2. N\_cis\_ori : Number of CIS EQTLs less than Pvalue cutoff in the main run
3. N\_trans\_p1: Number of TRANS EQTLs less than Pvalue cutoff in permutation 1

.

.

1. N\_trans\_p(n): Number of TRANS EQTLs less than Pvalue cutoff in permutation n
2. N\_trans\_ori: Number of TRANS EQTLs less than Pvalue cutoff in the main run
3. cis\_FDR: CIS EQTL FDR
4. trans\_FDR: TRANS EQTL FDR

“Tped\_Summary”->Genotype summary file.

1. SNP\_ID : SNP id
2. Chr : Chromosome
3. Pos : Position
4. N\_major : Major Alllele
5. N\_minor : Minor Allele
6. MAF : Minor allele Frequency
7. genotype\_distribution : Number of Major Homozygote\_Number of Heterozygote alleles\_Number of minor Homozygote alleles
8. N\_NA: Number of NA’s/Total number of markers

“Table\_withfiltering.txt”->

1. PVAL-RANGE : P Value range
2. CIS : Number of CIS EQTLs less than Pvalue cutoff in permutation 1

.

.

1. CIS: Number of CIS EQTLs less than Pvalue cutoff in permutation n
2. CIS\_ORI : Number of CIS EQTLs less than Pvalue cutoff in the main run
3. TRANS : Number of TRANS EQTLs less than Pvalue cutoff in permutation 1

.

.

1. TRANS : Number of TRANS EQTLs less than Pvalue cutoff in permutation n
2. TRANS\_ORI : Number of TRANS EQTLs less than Pvalue cutoff in the main run
3. CIS\_FDR : CIS EQTL FDR
4. TRANS\_FDR : TRANS EQTL FDR

“PVAL\_FILTER.CIS.gz”-> Below are the columns

1. SNP : SNP ID
2. gene : GENE/PROBE ID
3. t-stat : T Statistic
4. p-value : P Value
5. FDR : False Discovery Rate

“PVAL\_FILTER.TRANS.gz”-> Below are the columns

1. SNP : SNP ID
2. gene : GENE/PROBE ID
3. t-stat : T Statistic
4. p-value : P Value
5. FDR : False Discovery Rate

“PVAL.CIS.png” -> Q-Q plot for the cis eQTL pvalues

“PVAL\_FILTER.CIS.png”-> Q-Q plot for the filtered cis eQTL pvalues

“PVAL\_FILTER.TRANS.png”-> Q-Q plot for the filtered trans eQTL pvalues

“PVAL.TRANS.png” -> Q-Q plot for the trans eQTL pvalues

# ANNOTATION MODULE

This module will get the annotations for the significant genotyping markers. The Workflow uses the ANNOVAR tool to get the annotations.

Create the config file with below mandatory parameters (optional parameters are in red)

## INPUT ARGUMENTS

**RUNINFO CONFIG FILE ARGUMENTS**

tped =/data4/bsi/RandD/Workflow/EQTL\_WORKFLOW/sampledata/193sgenome.tped [Location of the genotyping TPED file]

tfam=/data4/bsi/RandD/Workflow/EQTL\_WORKFLOW/sampledata/193sgenome.tfam [Location of the genotyping TFAMfile]

temp.folder =/data4/bsi/RandD/Workflow/temp[location to the TEMP folder where temp files are created and once the Annotation is successful you can transfer the required files and delete the folder]

snp.list =/data4/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/temp\_annotate/snp\_list[List of SNP ids for annotations]

annot=geneanno\_\_gene,regionanno\_\_mce46way,regionanno\_\_tfbs,regionanno\_\_band,regionanno\_\_segdup,regionanno\_\_dgvMerged,regionanno\_\_gwascatalog,filter\_\_1000g2012apr,filter\_\_snp135\_\_webfrom,filter\_\_avsift\_\_webfrom,filter\_\_ljb2\_sift\_\_webfrom,filter\_\_esp6500\_all\_\_webfrom,filter\_\_gerp++gt2\_\_webfrom [List of annotations required]

Possible annotations are

(i)geneanno\_\_gene-> This option if for the REFSEQ gene based annotations.

Sample output: 1 161003087 161003087 C T comments: rs1000050, a SNP in Illumina SNP arrays  
1 11326183 11326183 - AT comments: rs35561142, a 2-bp insertion  
1 105293754 105293754 A ATAAA comments: rs10552169, a block substitution  
2 233848107 233848107 A G comments: rs2241880 (T300A), a SNP in the ATG16L1 associated with Crohn's disease  
16 49303427 49303427 C T comments: rs2066844 (R702W), a non-synonymous SNP in NOD2  
13 19661686 19661686 G - comments: rs1801002 (del35G), a frameshift mutation in GJB2, associated with hearing loss  
13 19695176 20003944 0 - comments: a 342kb deletion encompassing GJB6, associated with hearing loss

1. regionanno\_\_mce44way -> ANNOVAR uses phastCons 44-way alignments to annotate variants that fall within conserved genomic regions.

Sample output: mce44way Score=388;Name=lod=50 1 67478546 67478546 G A comments: rs11209026 (R381Q), a SNP in IL23R associated with Crohn's disease  
mce44way Score=421;Name=lod=68 16 49314041 49314041 G C comments: rs2066845 (G908R), a non-synonymous SNP in NOD2  
mce44way Score=392;Name=lod=52 16 49321279 49321279 - C comments: rs2066847 (c.3016\_3017insC), a frameshift SNP in NOD2

1. regionanno\_\_tfbs ->Transcription factor binding site annotation

Sample output: tfbs Score=878;Name=V$FREAC3\_01 13 19661686 19661686 G - comments: rs1801002 (del35G), a frameshift mutation in GJB2, associated with hearing loss  
tfbs Score=1000;Name=V$AML1\_01,V$MZF1\_01 13 19695176 20003944 0 - comments: a 342kb deletion encompassing GJB6, associated with hearing loss

1. regionanno\_\_band-> Identify cytogenetic band for genetic variants

Sample output: band 1q23.3 1 161003087 161003087 C T comments: rs1000050, a SNP in Illumina SNP arrays  
band 1p31.1 1 84647761 84647761 C T comments: rs6576700 or SNP\_A-1780419, a SNP in Affymetrix SNP arrays

1. regionanno\_\_segdup->Segmental Duplications

Sample output: segdup Score=0.996154;Name=1:13065149 1 13133880 13133881 TC - comments: rs59770105, a 2-bp deletion

1. regionanno\_\_dgvMerged ->Database of Genomic Variants. ANNOVAR can annotate deletions and duplications and compare them to previously published variants.

Sample output: dgv Score=0;Name=48150 1 161003087 161003087 C T comments: rs1000050, a SNP in Illumina SNP arrays  
dgv Score=0;Name=2298,31604,38855,48105,0256,3283,6779,0677,30369 1 13133880 13133881 TC - comments: rs59770105, a 2-bp deletion

1. regionanno\_\_gwascatalog ->Identify variants reported in previously published GWAS

Sample output: gwascatalog Score=0;Name=Ankylosing spondylitis,Crohn's disease,Ulcerative colitis,Inflammatory bowel disease 1 67478546 67478546 G A comments: rs11209026 (R381Q), a SNP in IL23R associated with Crohn's disease  
gwascatalog Score=0;Name=Crohn's disease 2 233848107 233848107 T C comments: rs2241880 (T300A), a SNP in the ATG16L1 associated with Crohn's disease

1. filter\_\_avsift\_\_webfrom -> SIFT/PolyPhen functional importance score annotations

Sample output: avsift 0.06 2 233848107 233848107 T C comments: rs2241880 (T300A), a SNP in the ATG16L1 associated with Crohn's disease

1. filter\_\_ljb\_sift \_\_webfrom -> **PolyPhen2, MutationTaster, LRT, PhyloP**

ljb\_pp2 0.97 16 49303427 49303427 C T comments: rs2066844 (R702W), a non-synonymous SNP in NOD2  
ljb\_pp2 0.508 16 49314041 49314041 G C comments: rs2066845 (G908R), a non-synonymous SNP in NOD2

1. filter\_\_esp6500\_all\_\_webfrom -> ESP (exome sequencing project) annotations

Sample output: esp6500\_all 0.046909 1 67478546 67478546 G A comments: rs11209026 (R381Q), a SNP in IL23R associated with Crohn's disease  
esp6500\_all 0.031558 16 49303427 49303427 C T comments: rs2066844 (R702W), a non-synonymous SNP in NOD2  
esp6500\_all 0.010157 16 49314041 49314041 G C comments: rs2066845 (G908R), a non-synonymous SNP in NOD2

1. filter\_\_gerp++gt2\_\_webfrom -> GERP identifies constrained elements in multiple alignments by quantifying substitution deficits

Sample output: gerp++gt2 5.14 1 67478546 67478546 G A comments: rs11209026 (R381Q), a SNP in IL23R associated with Crohn's disease  
gerp++gt2 2.42 16 49303427 49303427 C T comments: rs2066844 (R702W), a non-synonymous SNP in NOD2

[For further information visit http://www.openbioinformatics.org/annovar/]

db=/data4/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/ANNOVAR\_DB [Path to download annovar annotation source files, if exists will not recreate or download]

path\_perl=/usr/local/biotools/perl/5.10.0/bin/perl [Path to perl executable]

BUILD=18 [Build version]

path\_plink= /projects/bsi/gentools/bin/plink[Path to plink]

**TOOLINFO CONFIG FILE ARGUMENTS**

# PATH TO PERL EXECUTABLE

path\_perl=/usr/local/biotools/perl/5.10.0/bin/perl [Path to perl executable]  
# PATH TO PLINK EXECUTABLE

path\_plink= /projects/bsi/gentools/bin/plink[Path to plink]

# PATH TO ANNOVAR PROGRAM DIRECTORY

path\_annovar= /data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/EXTERNALTOOLS/ANNOVAR/annovar[Path to Annovar]

## SAMPLE EXECUTION

Once you create the config .you can execute the script in the following

|  |
| --- |
| perl perl\_eqtl\_workflow\_annotate.pl –run\_config <run.config> -tool.config <tool.config> |

## OUTPUT FILE

1. “FINAL.txt” : First column “chr start\_pos stop\_pos ALL1 ALL2 comments: marker” has chromosome, start position, Stop position, Allele 1, Allele 2,comments.
2. “FINAL.log” : This is the log file generate by the script, you can track the process and errors if occurred.
3. From column two you can find the annotations if there is no annotation the field will be annotated as “NA”.

# LD BLOCK MODULE

To calculate LD blocks from the significant cis and trans eQTLs associated SNPs using PLINK tool.

Create the config file with below mandatory parameters (optional parameters are in red)

## SAMPLE EXECUTION

Once you create the config .you can execute the script in the following

|  |
| --- |
| perl perl perl\_eqtl\_cal\_LD.pl -run\_config <run.config > -tool\_config <tool.config> |

## 

## INPUT ARGUMENTS

Run info config file description

snp.list=/data4/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/CIS\_SIG\_RSIDS[List of SNP ids for annotations]

temp.folder=/data4/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/temp\_LD[location to the TEMP folder which is use in the eQTL run]

run.dir=INNER\_TEST2[location to the inner directory in the TEMP folder which is use in the eQTL run]

ld.window.size =10000 [Preferred LD block length]

ld.window.r2 =0.3 [Preferred LD block r2 value]

output.dir=INNER\_TEST2[location to the output directory which is use in the eQTL run]

Tool info config file description

# PATH TO PLINK EXECUTABLE

path\_plink=[ PATH TO PLINK EXECUTABLE]

## OUTPUT

## You can find out two output files in the output directory.

1. “output.blocks”: lists each block (2 or more SNPs) on a row, starting with an asterisk symbol (\*), for example:

\* rs7527871 rs2840528 rs7545940

\* rs2296442 rs2246732

\* rs10752728 rs897635

\* rs10489588 rs9661525 rs2993510

1. “output.blocks.det”: is similar to the first, but contains some addition information:

Chr Chromosome identifier

LD\_block\_start The start position (base-pair units) of this block

LD\_block\_end The end position (base-pair units) of this block

block\_size\_KB The kilobase distanced spanned by this block

N\_SNP The number of SNPs in this block

SNP\_ID List of SNPs in this block

# Downloading External Tools (TOOLINFO FILE)

Script “install\_tools.sh” will install all the external tools. Make sure there are no errors while installation and manual checkup is required.

Script “make\_tool\_info.sh” creates the tool info file.

Download the following the tools, unzip them and place them in a common directory. Create a tool info config file for the programs to execute.

Here is the sample tool info config file (see below detailed description).

Create main tools directory

Mkdir /home/EQTL\_Sample\_project/EXTERNALTOOLS

Change directory

Cd /home/EQTL\_Sample\_project/EXTERNALTOOLS

1. **PLINK(Version v1.07):** Plink is a free, open-source whole genome association analysis toolset. You can download plink from <http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml> .

Create main plink directory

Mkdir /home/EzImputer\_Sample\_project/EXTERNALTOOLS/PLINK

Change to plink main directory

Cd /home/EzImputer\_Sample\_project/EXTERNALTOOLS/PLINK

Download the plink packge

Wget <http://pngu.mgh.harvard.edu/~purcell/plink/dist/plink-1.07-x86_64.zip>

Uncompress zip files

Copy uncompressed contents to this directory

When you execute the plink executable. You should be able to see

plink

@----------------------------------------------------------@

| PLINK! | v1.07 | 10/Aug/2009 |

|----------------------------------------------------------|

| (C) 2009 Shaun Purcell, GNU General Public License, v2 |

|----------------------------------------------------------|

| For documentation, citation & bug-report instructions: |

| http://pngu.mgh.harvard.edu/purcell/plink/ |

@----------------------------------------------------------@

Web-based version check ( --noweb to skip )

Connecting to web... OK, v1.07 is current

Writing this text to log file [ plink.log ]

Analysis started: Fri Mar 15 10:37:07 2013

Options in effect:

Before frequency and genotyping pruning, there are 0 SNPs

0 founders and 0 non-founders found

0 SNPs failed missingness test ( GENO > 1 )

0 SNPs failed frequency test ( MAF < 0 )

After frequency and genotyping pruning, there are 0 SNPs

ERROR: Stopping as there are no SNPs left for analysis

1. ANNOVAR: ANNOVAR, functional annotation of genetic variants from high-throughput sequencing data, is an efficient command line Perl program to functionally annotate genetic variants from diverse genomes (including human genome hg18, hg19, as well as mouse, worm, fly, yeast and many others).Fill the form at <http://www.openbioinformatics.org/annovar/annovar_download_form.php> and you will receive an email with download instructions.

Create main ANNOVAR directory

Mkdir /home/EzImputer\_Sample\_project/EXTERNALTOOLS/ANNOVAR

Change to ANNOVAR main directory

Cd /home/EzImputer\_Sample\_project/EXTERNALTOOLS/ANNOVAR

Download the ANNOVAR packge (link in the email)

Wget <http://www.openbioinformatics.org/annovar/download/uYl9e52m87/annovar.latest.tar.gz>

Uncompress zip files

tar -zxvf annovar.latest.tar.gz

cd annovar

echo `pwd` (This is your ANNOVAR directory)

When you execute the perl script

perl annotate\_variation.pl

Usage:

annotate\_variation.pl [arguments] <query-file|table-name> <database-location>

Optional arguments:

-h, --help print help message

-m, --man print complete documentation

-v, --verbose use verbose output

1. Dependent R packages needs to be installed

**Before this step make sure ‘R’ is installed on your system**

**Try the command ‘which Rscript’ (add path to R and Rscript to your bashrc profile)**

**You should be able to see the path to Rscript executable**

Create main Rpackages directory

Mkdir /home/EzImputer\_Sample\_project/EXTERNALTOOLS/Rpackages

Change to Rpackages main directory

Cd /home/EzImputer\_Sample\_project/EXTERNALTOOLS/Rpackages

**MATRIX EQTL PACKAGE (VERSION 1.6.2)**

Download Matrix EQTL package at

<http://cran.r-project.org/web/packages/MatrixEQTL/index.html>

wget <http://cran.r-project.org/src/contrib/MatrixEQTL_1.6.2.tar.gz>

**MASS(VERSION** 7.3-23)

Download MASS package at

<http://cran.r-project.org/web/packages/MASS/index.html>

wget <http://cran.r-project.org/src/contrib/MASS_7.3-23.tar.gz>

**Sfsmisc(VERSION** 1.0-23)

Download sfsmisc package at

<http://cran.r-project.org/web/packages/sfsmisc/index.html>

wget <http://cran.r-project.org/src/contrib/sfsmisc_1.0-23.tar.gz>

Create directory “rlib”

mkdir rlib

Install R package in that rlib directory

R CMD INSTALL -l ./rlib/ MASS\_7.3-23.tar.gz

R CMD INSTALL -l ./rlib/ MatrixEQTL\_1.6.2.tar.gz

R CMD INSTALL -l ./rlib/ sfsmisc\_1.0-23.tar.gz

Check the installed r packages in this way

Execute the R script “**check\_Rpackages.R”** in this way (give argument path to rlib directory)

Rscript check\_Rpackages.R /home/EzImputer\_Sample\_project/EXTERNALTOOLS/Rpackage/rlib/

PLINK=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/PLINK /plink

PERL=/usr/local/biotools/perl/5.10.0/bin/perl

QSUB=/home/sge6\_2/bin/lx24-amd64/qsub

SH=/bin/bash

# EXAMPLES

Here are some example config files. One tool info config for all the scripts

**TOOLINFO CONFIG FILE**

path\_plink=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW//EXTERNALTOOLS/PLINK/plink-1.07-i686/plink

path\_perl=/usr/local/biotools/perl/5.10.0/bin/perl

path\_sh=/bin/sh

path\_qsub=/home/oge/ge2011.11/bin/linux-x64/qsub

path\_rscript=/usr/local/biotools/r/R-2.15.0/bin/Rscript

path\_rlib=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW//EXTERNALTOOLS/Rpackages/rlib

path\_annovar=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW//EXTERNALTOOLS/ANNOVAR/annovar

**CONFIG FILE EQTL RUN**

1. **RUNINFO CONFIG FILE**

tped=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/hapmap3\_final\_truncated.tped

tfam=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/hapmap3\_final.tfam

output.dir=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/OUTPUT

temp.folder=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/TEMP

expr.data=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/Final\_expression.txt

expr.na.symbol=NaN

expr.loc=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/Final\_geneloc.txt

only.summary=NO

cluster.queue=4-days

cluster.max.mem=20g

email=prodduturi.naresh@mayo.edu

num.permutation=10

cis.pval=1e-4

trans.pval=1e-6

sge=YES

run.dir=INNER\_TEST2

filtering=YES

fil.min.num.gnt=5

fil.max.na.gnt=0.5

fil.maf.cutoff.gnt=0.05

fil.max.mean.geneexp=10000

fil.min.mean.geneexp=-10000

fil.min.std.geneexp=0

fil.max.std.geneexp=10000

fil.max.na.geneexp=0.1

covariate.data=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/Final\_covariate.txt

cisDist=100000

geneannot.attach.genesumm=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/gene\_annotation.txt

**CONFIG FILE ANNOTATION**

1. **RUNINFO CONFIG FILE**

tped=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/hapmap3\_final\_truncated.tped

tfam=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/hapmap3\_final.tfam

snp.list=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/OUTPUT/INNER\_TEST2/snp.list

temp.folder=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/ANNOVAR\_TEMP2

annot=geneanno\_\_gene,regionanno\_\_mce46way,regionanno\_\_tfbs,regionanno\_\_band,regionanno\_\_segdup,regionanno\_\_dgvMerged,regionanno\_\_gwascatalog,filter\_\_1000g2012apr,filter\_\_snp135\_\_webfrom,filter\_\_avsift\_\_webfrom,filter\_\_ljb2\_sift\_\_webfrom,filter\_\_esp6500\_all\_\_webfrom,filter\_\_gerp++gt2\_\_webfrom

db=/data5/bsi/RandD/Workflow/temp/ANNOVAR\_DB/

build=19

**CONFIG FILE LD**

snp.list=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/snplist\_ld

temp.folder=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/TEMP/

run.dir=INNER\_TEST2

ld.window.size=1000000

ld.window.r2=0.3

output.dir=/data5/bsi/bioinf\_ext1/s113625.eQTL/EQTL\_WORKFLOW/sampledata/OUTPUT

outputfile=ld\_out