GTEx summary statistics database

Gao Wang

Last updated: August 14, 2015

1 Databases

1.1 Multi-tissue summary stats

Summary statistics from the GTEx v6 analysis are stored in `MatrixEQTLSumStats.h5` (130GB, in HDF5 format). It contains 38,933 genes (less than previous version! See the list `/project/mstephens/datab` each gene has several summary statistics matrix of dimension $N_{cisSNPs} \times N_{tissues}$. For this release $N_{tissues} = 44$. For `MatrixEQTL` the summary statistics are $\hat{\beta}$, T-stat and P-value.

MatrixEQTLSumStats.h5

```
/ENSG00000008735.10 'MatrixEQTL summary statistics of GTEx Release 2015.04.15'
/ENSG00000008735.10/beta (5636, 44)
/ENSG00000008735.10/p-value (5636, 44)
/ENSG00000008735.10/t-stat (5636, 44)
/ENSG00000008735.10/rownames (5636,)
/ENSG00000008735.10/colnames (44,)
...
```

There are two special tables in `MatrixEQTLSumStats.h5`. One called `max`, which is the data for the "best" gene-snp pair identified per gene as the snp pair having maximum abs(t-stat) across all snps and all tissues and there is no missing data for any tissue. It contains 16,069 gene-snp pairs. The other called `null`, which organize gene-snp pair data in the same format as `max` only that the gene-snp pairs in this table are NOT the best gene-snp pair, but just random samples from all non-best gene-snp pairs. For portability these two tables are also stored separately in `MatrixEQTLSumStats.Portable.h5` (58MB, in HDF5 format).

MatrixEQTLSumStats.Portable.h5

```
/max (Group) 'max'
/max/beta (16069, 44)
/max/p-value (16069, 44)
/max/t-stat (16069, 44)
/max/rownames (16069,)
/max/colnames (44,)
```

1.2 Meta-database

Unfortunately GTEx names *cis*SNP by their genomic coordinate in Human Genome Assembly hg19 (b37), not by rsID. To make it possible to search with rsID I created another database `snp-gene.db` (1.6GB, in SQLite format).

snp-gene.db

```
3_52771872_C_T_b37      rs59638016      ENSG00000016864.12,ENSG00000055955.11,ENSG00000055957.6 ...
3_52774530_C_T_b37      rs2159644       ENSG00000016864.12,ENSG00000055955.11,ENSG00000055957.6 ...
3_52774835_AT_A_b37     rs35270761,rs397785457  ENSG00000016864.12,ENSG00000055955.11,ENSG00000055957.6 ...
...
```

The 3 columns are

- SNP name in GTEx convention
- rsID
- cis-genes, *i.e.*, genes whose TSS is within 100,000bp up/down-stream of the SNP.

This database will be used prior to searching the summary statistics.

Note

- There may be multiple rsID associated with the same coordinate. Indeed such records exists in dbSNP 144, and eventually these rsIDs should merge into one name (as always have happened in history).
- There are 10,297,646 SNPs in total in this dataset. 9,794,339 of them (95.1%) has rsID in dbSNP 144.

2 Queries

I provide a simple R script `SumstatQuery.R` to make queries on this dataset. It has:

- A function that lists for given rsID the GTEx SNP ID and all its cis-genes.
- A function that loads sumstats data given gene name and GTEx SNP ID.

To use the script, `RSQLite` and `rhdf5` should be installed, *e.g.*,

R

```
install.packages("RSQLite")
source("http://bioconductor.org/biocLite.R")
biocLite("rhdf5")
```

2.1 Example query

The demo script below shows an example to extract information for gene ENSG00000171960.6 and SNP rs6600419, and examples to extract the best/null gene-snp pairs.

R

```
source("/project/mstephens/gtex/scripts/SumstatQuery.R")
##
# Load data for given gene
##
dat <- GetSS("ENSG00000171960.6", "/project/mstephens/gtex/analysis/april2015/query/MatrixEQTLSumStats.h5")
print(names(dat))
## [1] "beta"    "p-value" "t-stat"  "z-score"
dim(dat$"beta")
## [1] 6344   44
dat$"beta"[1:4,1:4]
##                     Adipose_Subcutaneous Adipose_Visceral_Omentum Adrenal_Gland Artery_Aorta
## 1_42124161_C_T_b37            0.04608095                      NaN           NaN   0.06785597
## 1_42124246_A_AC_b37          -0.03087348              -0.03222961    -0.1286249  -0.13490575
## 1_42124311_G_A_b37            0.12400229               0.08011222           NaN  -0.21858459
## 1_42124614_C_T_b37            0.04630340                      NaN           NaN   0.06785597
##
# Output information for the SNP of interest
##
dat$"p-value"["1_43124701_A_G_b37", ]
 ##                 Adipose_Subcutaneous              Adipose_Visceral_Omentum
 ##                         1.632499e-01                          3.462833e-02
 ##                        Adrenal_Gland                          Artery_Aorta
 ##                         8.168959e-01                          5.600338e-01
 ##                      Artery_Coronary                         Artery_Tibial
 ##                         7.659426e-01                          7.397466e-01
 ## Brain_Anterior_cingulate_cortex_BA24           Brain_Caudate_basal_ganglia
 ## ...
dat$"z-score"["1_43124701_A_G_b37", ]
 ##                 Adipose_Subcutaneous              Adipose_Visceral_Omentum
 ##                          -1.39422418                           -2.11267803
 ##                        Adrenal_Gland                          Artery_Aorta
 ##                          -0.23153929                            0.58279135
 ##                      Artery_Coronary                         Artery_Tibial
 ##                           0.29768626                            0.33218889
 ## Brain_Anterior_cingulate_cortex_BA24           Brain_Caudate_basal_ganglia
 ## ...
###
# Load the best gene-snp data
###
mdat <- GetSS("max", "/project/mstephens/gtex/analysis/april2015/query/MatrixEQTLSumStats.h5")
dim(mdat$"t-stat")
mdat$"p-value"[1:4,1:4]
mdat$"t-stat"["ENSG00000000419.8_20_49461813_G_C_b37",]
## Is this gene-snp pair most significant in spleen?
dat <- GetSS("ENSG00000000419.8", "/project/mstephens/gtex/analysis/april2015/query/MatrixEQTLSumStats.h5")
idx.to.show <- matxMax(abs(dat$"t-stat"))
rownames(dat$"t-stat")[idx.to.show[1]]
colnames(dat$"t-stat")[idx.to.show[2]]
###
# Load the "null" gene-snp data
###
ndat <- GetSS("null", "/project/mstephens/gtex/analysis/april2015/query/MatrixEQTLSumStats.h5")
dim(ndat$"t-stat")
##
# Look up GTEx SNP ID
##
ShowSNP("rs6600419", "/project/mstephens/gtex/analysis/april2015/query/snp-gene.db")
## GTEx SNP ID: 1_43124701_A_G_b37
## cisGenes: ENSG00000065978.13,ENSG00000164007.6,ENSG00000171960.6,ENSG00000200254.1,ENSG00000234917.1,ENSG00000236180.2
##
# Look up rs ID given GTEx SNP ID
##
ShowSNP("1_43124701_A_G_b37", "/project/mstephens/gtex/analysis/april2015/query/snp-gene.db")
## rsID(s): rs6600419
## cisGenes: ENSG00000065978.13,ENSG00000164007.6,ENSG00000171960.6,ENSG00000200254.1,ENSG00000234917.1,ENSG00000236180.2
##
# Create matched training/testing sets
##
```

```
N1 <- 8000
N2 <- 16069
strong.train <- SubsetMatLists(mdat, seq(1, N1))
strong.test <- SubsetMatLists(mdat, seq(N1 + 1, N2))
strong.train.genes <- as.character(lapply(strsplit(rownames(strong.train$beta), "_"), function(x) x[1]))
strong.test.genes <- as.character(lapply(strsplit(rownames(strong.test$beta), "_"), function(x) x[1]))
null.genes <- as.character(lapply(strsplit(rownames(ndat$beta), "_"), function(x) x[1]))
null.train <- SubsetMatLists(ndat, which(null.genes %in% strong.train.genes))
null.test <- SubsetMatLists(ndat, which(null.genes %in% strong.test.genes))
```

For `MatrixEQTLSumStats.h5`, for given rsID you first search the GTEx SNP ID via `ShowSNP`, then load the cisGene of interest via `GetSS` and extract information using the GTEx SNP ID as the row name key.

> **Note**
>
> If R complains the given row name does not exist, try to flip the DNA strand in the GTEx SNP ID and search again. For example, instead of searching for `1_43124701_A_G_b37` you search for `1_43124701_T_C_b37`.

The best gene-snp data can be loaded to memory, via `GetSS("max", "MatrixEQTLSumStats.h5")`, or the portable version `GetSS("max", "MatrixEQTLSumStats.Portable.h5")`

Same for the “null” gene-snp data, via `GetSS("null", "MatrixEQTLSumStats.h5")`, or the portable version `GetSS("null", "MatrixEQTLSumStats.Portable.h5")`