Brief Introduction to Genomic informed Drug Target database

BALDER team

2022-11-09

Contents

We have developed scripts and workflows that efficiently generate functional marker data sets from publicly available resources. Functional marker information has been downloaded and processed from:

- Ensembl (link SNPs to genes and proteins)
- GO (gene ontology)
- STRING (protein-protein)
- STITCH (protein-chemical)
- Reactome (biological pathways)
- more ressources will be added …..

Processing included quality control, mapping to LD reference panel and creation of marker sets (e.g., markers linked to genes, proteins, pathways) used in marker set analyses and subsequently be used to help the biological interpretation of genome-wide association studies.

This includes screening functional marker sets (e.g. biological pathways, protein complexes, gene ontology terms) for association with complex diseases.

It is also possible to test specific biological hypothesis such as:

Genes, proteins, metabolites, pathways underlying T2DM are enriched for association signal with T2DM

Drugs used for treatment of T2DM are linked to genes, proteins, metabolites, pathways enriched for association signal with T2DM

Our workflow allows us to quickly process new functional marker sets and we will therefore continue to identify and process functional marker data relevant for T2DM and other complex disease.

The practical is based on the R package `gact` (Rohde et al.2022)). This package provides an infrastructure for working with large-scale genomic association data linked to different types of genomic features.

Load packages used

The most recent version of `gact` can be obtained from github:

```r
library(devtools)
devtools::install_github("psoerensen/gact")
```

```r
library(gact)
library(qgg)
library(corrplot)
library(data.table)
```

Download and install GDT database

The function `gact()` dowload and install the GDT database:

```
# Set working for database
dbdir <- "C:/Users/au223366/Dropbox/Projects/balder/gdtdb"

# Download data bases from repository
GAlist <- gact(version = "t2dm-gact-0.0.1", dbdir = dbdir, task = "download")
GAlist$features
```

```
## [1] "Markers"            "Genes"              "Proteins"
## [4] "GO"                 "Pathways"           "ProteinComplexes"
## [7] "ChemicalComplexes"
```

```
GAlist$studies
```

```
## NULL
```

```
# Information about features in GDT database
GAlist$features
```

```
## [1] "Markers"            "Genes"              "Proteins"
## [4] "GO"                 "Pathways"           "ProteinComplexes"
## [7] "ChemicalComplexes"
```

Add new summary statisticsto GDT database

The function `getStat()` extract data from the database:

```
# CARDIoGRAMplusC4D.txt.gz
fname_stat <- "C:/Users/au223366/Dropbox/Projects/balder/data/CARDIoGRAMplusC4D.txt.gz"
stat <- fread(fname_stat, data.table = FALSE)
head(stat)
stat <- stat[, c(1:6, 9:11)]
colnames(stat) <- c("marker", "chr", "pos", "ea", "nea", "eaf", "b", "seb",
    "p")

GAlist <- updateStatDB(GAlist = GAlist, stat = stat, source = "CARDIoGRAMplusC4D.txt.gz",
    trait = "CAD", type = "binary", gender = "both", reference = "PMID:26343387",
    n = 184305, ncase = 60801, ncontrol = 123504)
```

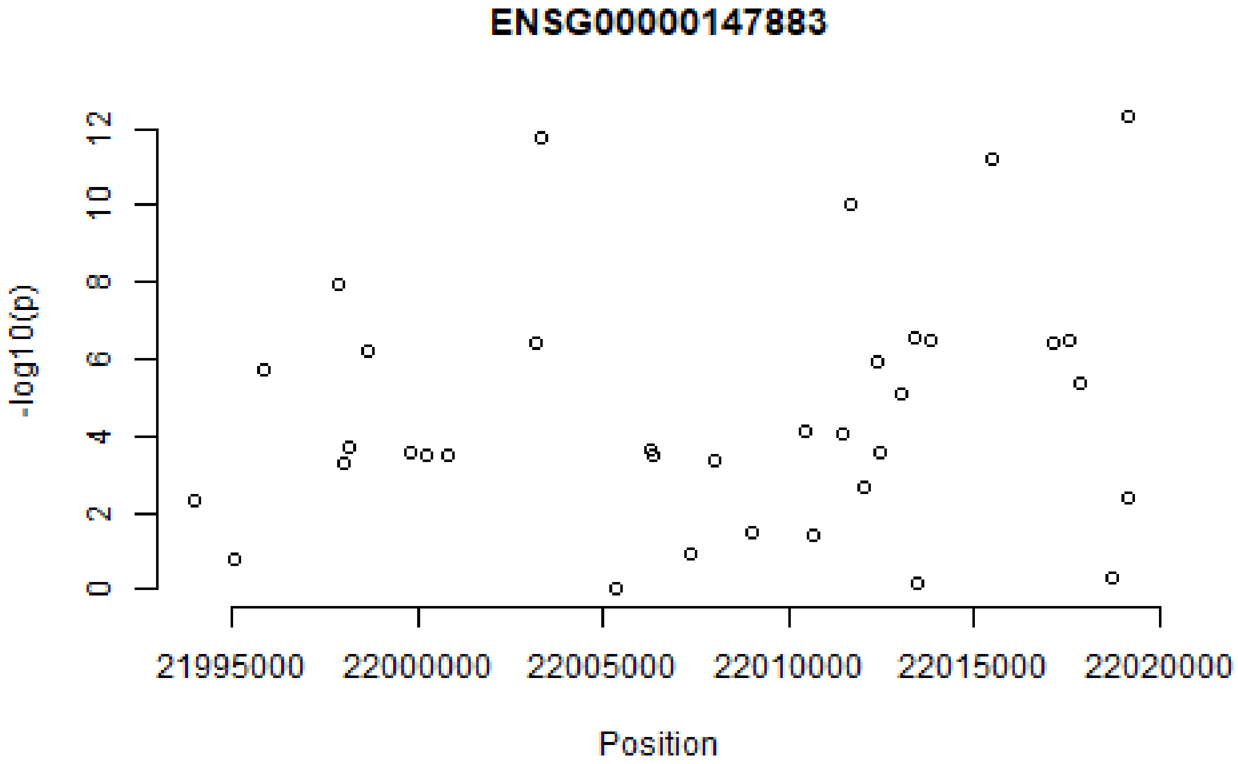
Extract data from GDT database

The function `getStat()` extract data from the database:

```
# Extract data from T2D for genomic feature Markers
stat <- getStat(GAlist = GAlist, trait = "t2d", feature = "Markers")
head(stat)
```

```
##                  rsids chr    pos a1 a2    af       b   seb    p     n
## rs2000096    rs2000096   1 567867  G  A 0.000 -0.5200 0.630 0.41 28130
## rs12238997 rs12238997   1 693731  G  A 0.130 -0.0088 0.017 0.60 28130
## rs72631875 rs72631875   1 705882  A  G 0.063  0.0110 0.037 0.76 28130
## rs55727773 rs55727773   1 706368  A  G 0.500  0.0140 0.015 0.37 28130
## rs12184267 rs12184267   1 715265  T  C 0.041 -0.0610 0.061 0.32 28130
## rs12184277 rs12184277   1 715367  G  A 0.040 -0.0580 0.061 0.34 28130
```

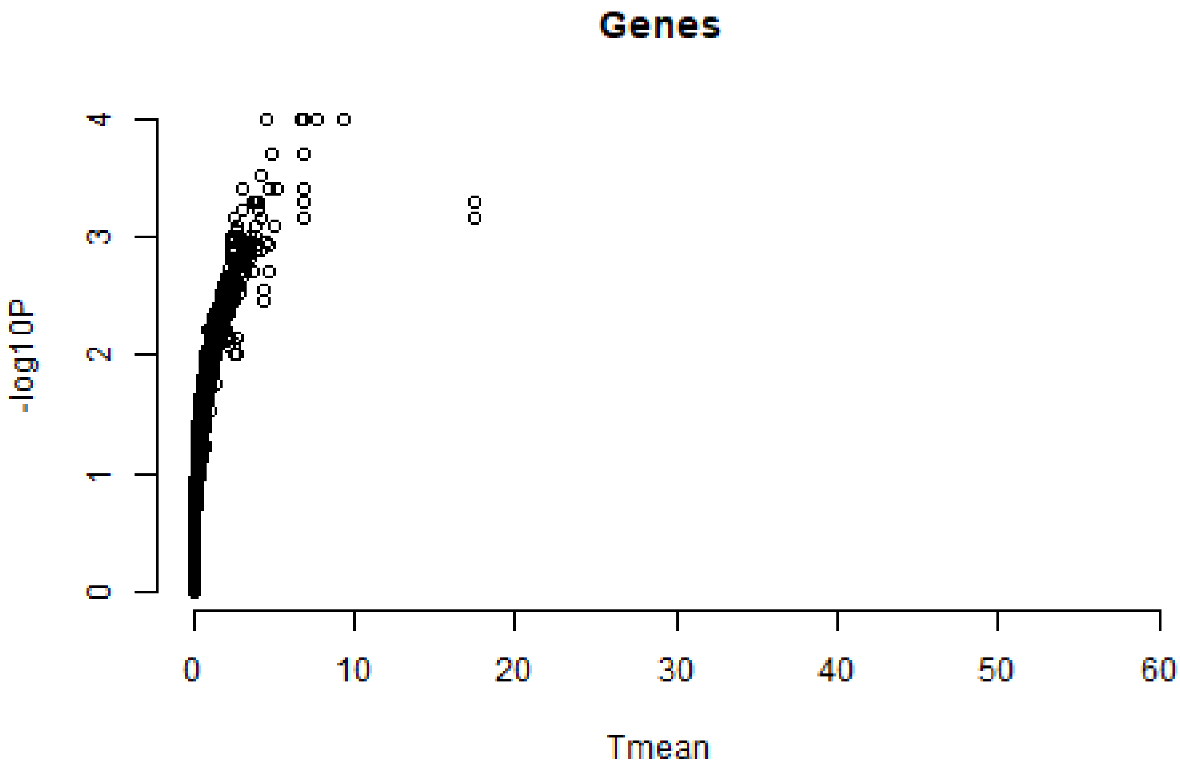
```
# Extract marker sets for ENSG00000147883 and plot
rsids <- getSets(GAlist = GAlist, feature = "Genes", featureID = "ENSG00000147883")
plot(y = -log10(stat[rsids, ]$p), x = stat[rsids, ]$pos, ylab = "-log10(p)",
    xlab = "Position", frame.plot = FALSE, main = "ENSG00000147883")
```



```
# Extract data from T2D for genomic feature Genes
stat <- getStat(GAlist = GAlist, trait = "t2d", feature = "Genes")
head(stat)
```

```
##                 Ensembl Gene ID   Symbol   m     stat      p
## ENSG00000121410 ENSG00000121410     A1BG  58  0.00000 1.0000
## ENSG00000175899 ENSG00000175899      A2M 153 20.78100 0.1477
## ENSG00000256069 ENSG00000256069    A2MP1  64  0.00000 1.0000
## ENSG00000171428 ENSG00000171428     NAT1 135  0.00000 1.0000
## ENSG00000156006 ENSG00000156006     NAT2 102 20.90052 0.0951
## ENSG00000196136 ENSG00000196136 SERPINA3  64  0.00000 1.0000
```

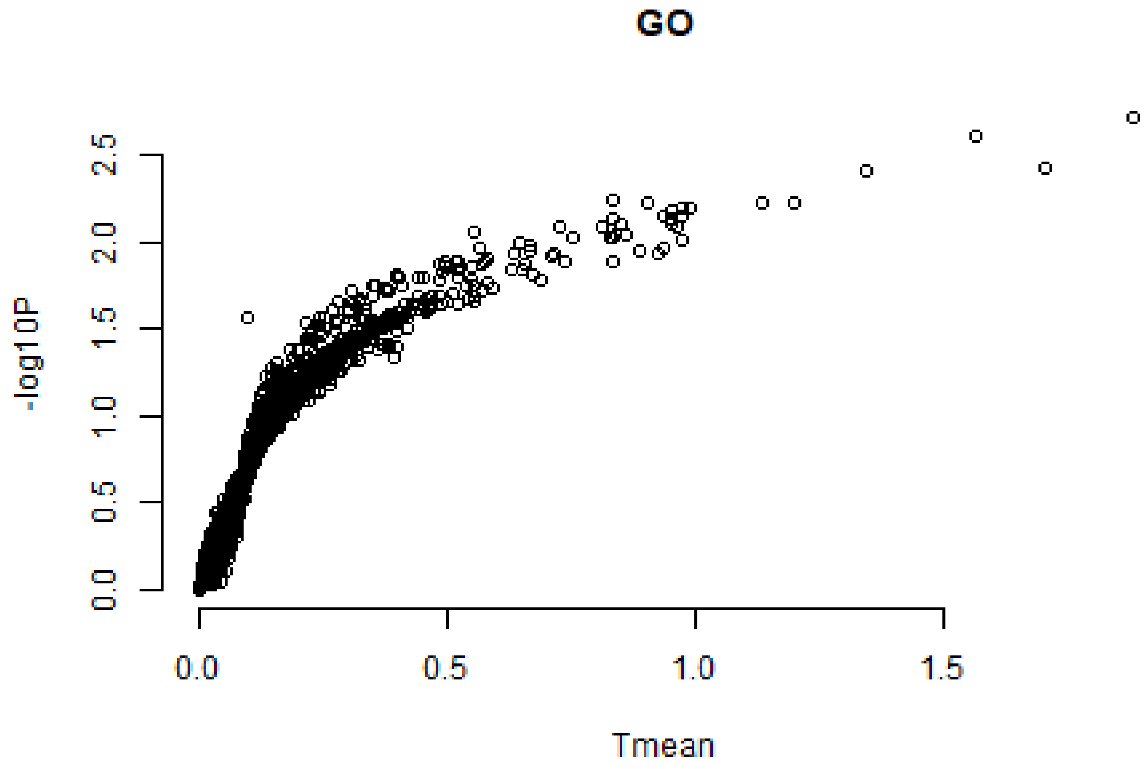
```
plot(x = stat$stat/stat$m, y = -log10(stat$p), ylab = "-log10P", xlab = "Tmean",
    frame.plot = FALSE, main = "Genes")
```



```
# Extract data from T2D for genomic feature Gene Ontology (GO) where
# output format is a data frame
stat <- getStat(GAlist = GAlist, trait = "t2d", feature = "GO")
head(stat)
```

```
##                   GO ID    m      stat      p
## GO:0000002 GO:0000002 2165 143.36901 0.3208
## GO:0000012 GO:0000012 1885 107.38110 0.3577
## GO:0000027 GO:0000027 1867  72.77531 0.5031
## GO:0000028 GO:0000028  861  71.43454 0.2242
## GO:0000038 GO:0000038 2397  63.58818 0.6599
## GO:0000045 GO:0000045 8223 749.36205 0.2368
```

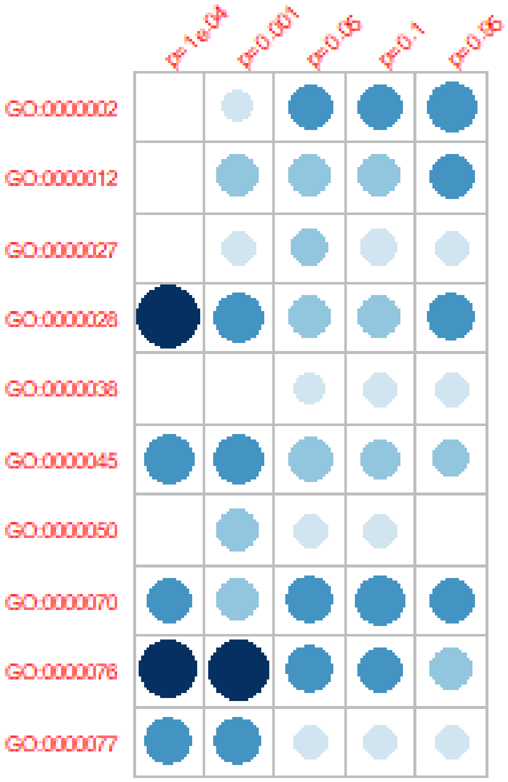
```
plot(x = stat$stat/stat$m, y = -log10(stat$p), ylab = "-log10P", xlab = "Tmean",
    frame.plot = FALSE, main = "GO")
```



```r
# Extract data from T2D for genomic feature Gene Ontology (GO) where
# output format is list
stat <- getStat(GAlist = GAlist, trait = "t2d", feature = "GO", format = "list",
    cls = c("p=1e.04", "p=0.001", "p=0.05", "p=0.1", "p=0.95"))
str(stat)
```

```
## List of 3
##  $ m   : Named int [1:4547] 2165 1885 1867 861 2397 8223 1050 3519 1077 5192 ...
##   ..- attr(*, "names")= chr [1:4547] "GO:0000002" "GO:0000012" "GO:0000027" "GO:0000028" ...
##  $ stat: num [1:4547, 1:5] 0 0 0 17 0 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4547] "GO:0000002" "GO:0000012" "GO:0000027" "GO:0000028" ...
##   .. ..$ : chr [1:5] "p=1e.04" "p=0.001" "p=0.05" "p=0.1" ...
##  $ p   : num [1:4547, 1:5] 1 1 1 0.0781 1 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4547] "GO:0000002" "GO:0000012" "GO:0000027" "GO:0000028" ...
##   .. ..$ : chr [1:5] "p=1e.04" "p=0.001" "p=0.05" "p=0.1" ...
```

```r
# Plot results for genomic feature Gene Ontology (GO)
colbar <- colorRampPalette(c("#FFFFFF", "#D1E5F0", "#92C5DE", "#4393C3",
    "#2166AC", "#053061"))
corrplot(-log10(stat$p[1:10, ]), is.corr = FALSE, tl.cex = 0.7, tl.srt = 45,
    col = colbar(6), cl.pos = "n", mar = c(1, 1, 1, 1))
```



Extract and write data from GDT database

The function `writeStat()` extract and write data from the database:

```
writeStat(GAlist = GAlist, feature = "GO", trait = "t2d", file.csv = "go_t2dm_gcta.csv")
writeStat(GAlist = GAlist, feature = "Pathways", trait = "t2d", file.csv = "pathways_t2dm_gcta.csv")
writeStat(GAlist = GAlist, feature = "ProteinComplexes", trait = "t2d",
    file.csv = "proteincomplexes_t2dm_gcta.csv")
writeStat(GAlist = GAlist, feature = "ChemicalComplexes", trait = "t2d",
    file.csv = "chemicalcomplexes_t2dm_gcta.csv")
writeStat(GAlist = GAlist, feature = "Genes", trait = "t2d", file.csv = "genes_t2dm_gcta.csv")
```

Extract marker set data from GDT database

The marker sets in the database can be extracted using:

```
geneSets <- getSets(GAlist = GAlist, feature = "Genes")
chemSets <- getSets(GAlist = GAlist, feature = "ChemicalComplexes2Genes")
```

Extract data for chemical "CIDm00004091" from GDT database

The marker sets in the database can be extracted using:

```
chemStat <- getStat(GAlist = GAlist, trait = "t2d", feature = "ChemicalComplexes",
    cls = c("p=1e.04", "p=0.001", "p=0.05", "p=0.1", "p=0.95"))
chemStat["CIDm00004091", ]
```

```
##                Chemical ID     m stat.p.1e.04 stat.p.0.001 stat.p.0.05
## CIDm00004091 CIDm00004091 13369     279.0648     525.4034    2521.156
##              stat.p.0.1 stat.p.0.95 p.p.1e.04 p.p.0.001 p.p.0.05 p.p.0.1
## CIDm00004091   3560.588    6134.161    0.1882    0.2233   0.2662  0.2413
##              p.p.0.95
## CIDm00004091   0.2592
```

```
genesStat <- getStat(GAlist = GAlist, trait = "t2d", feature = "Genes")
ensgIDs <- getSets(GAlist = GAlist, feature = "ChemicalComplexes2Genes",
    featureID = "CIDm00004091")
head(genesStat[ensgIDs, ])
```

```
##                   Ensembl Gene ID Symbol   m     stat      p
## ENSG00000050344 ENSG00000050344 NFE2L3 109 6.698962 0.3215
## ENSG00000065970 ENSG00000065970  FOXJ2 127 0.000000 1.0000
## ENSG00000100448 ENSG00000100448   CTSG  26 0.000000 1.0000
## ENSG00000103121 ENSG00000103121   CMC2 274 6.950413 0.5150
## ENSG00000104899 ENSG00000104899    AMH  72 0.000000 1.0000
## ENSG00000104918 ENSG00000104918   RETN  40 0.000000 1.0000
```