SUMSTATS

Description of GWAS summary statistics

The repository shows how to provide input for PW-pipeline, FM-pipeline including GCTA, and possibly others here.

Briefly, the format has the following columns,

Column	Name	Description
1	SNP	RSid
2	A1	Effect allele
3	A2	Other allele
4	freqA1	A1 frequency
5	beta	effect estimate
6	se	standard error of effect
7	P	P-value
8	N	sample size
9*	chr	chromosome
10*	pos	position

* These two columns can be obtained from https://genome.ucsc.edu/ as shown below.

It might be worthwhile to check for options with the sumstats as defined in ldsc, https://github.com/bulik/ldsc, and particularly its munge_sumstats.py utility.

Information from UCSC

The chromosomal positions for the current build can be downloaded from the UCSC website, which should be helpful for GWAS summary statistics either using chromosomal positions from different build or without these at all.

```
wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/snp150.txt.gz
gunzip -c snp150.txt.gz | \
awk '{split($2,a,"_");sub(/chr/,"",a[1]);print a[1],$4,$5}' | \
sort -k3,3 > snp150.txt
```

where it first obtains build 37 positions, sorts them by RSid into the file `snp150.txt`.

Examples

BMI

We take data reprted by Locke, et al. (2015) as example which requires build 37 positions from UCSC described above.

```
# GWAS summary statistics
wget
http://portals.broadinstitute.org/collaboration/giant/images/1/15/SNP_gwas_mc
_merge_nogc.tbl.uniq.gz
gunzip -c SNP_gwas_mc_merge_nogc.tbl.uniq.gz |
awk 'NR>1' | \
sort -k1,1 | \
join -11 -23 - snp150.txt | \
awk '($9!="X" && $9!="Y" && $9!="Un")' > bmi.txt
```

where file containing the GWAS summary statistics is downloaded, its header dropped, sorted and positional information added leading to a file named `bmi.txt`. We also filter out nonautosomal SNPs.

The list of 97 SNPs can be extracted as follows,

```
R --no-save <<END
library(openxlsx)
xlsx <- "https://www.nature.com/nature/journal/v518/n7538/extref/nature14177-
s2.xlsx"
snps <- read.xlsx(xlsx, sheet = 4, colNames=FALSE, skipEmptyRows = FALSE,
cols = 1, rows = 5:101)
snplist <- sort(as.vector(snps[,1]))
write.table(snplist, file="97.snps", row.names=FALSE, col.names=FALSE,
quote=FALSE)
END
```

T2D

The data was reported by Scott, et al. (2017),

```
R -q --no-save <<END

library(openxlsx)
library(dplyr)

xlsx <-
"http://diabetes.diabetesjournals.org/highwire/filestream/79037/field_highwir
e_adjunct_files/1/DB161253SupplementaryData2.xlsx"

# Supplementary Table 3. Results for established, novel and additional
distinct signals from the main analysis.
ST3 <- read.xlsx(xlsx, sheet = 3, colNames=TRUE, skipEmptyRows = FALSE, cols
= 1:20, rows = 2:130) %>%
       rename(P="p-value.in.stage.1") %>% within(
       {
          beta=log(OR)
          L <- as.numeric(substr(CI,1,4))
          U <- as.numeric(substr(CI,6,9))
          se=abs(log(L)-log(U))/3.92
       }) %>% select(
          SNP=rsid,
```

```
          A1=EA,
          A2=NEA,
          freqA1=EAF,
          beta,
          se,
          P,
          N=Sample.size,
          chr=Chr,
          pos=Position_b37
       )
write.table(ST3, file="ST3", row.names=FALSE, col.names=FALSE, quote=FALSE)

# Supplementary Table 4. BMI-unadjusted association analysis model
ST4 <- read.xlsx(xlsx, sheet = 4, colNames=TRUE, skipEmptyRows = FALSE, cols
= 1:12, rows = 3:132) %>% rename(
          "CI"="CI.95%",
          "P"="P-value") %>% within(
       {
          beta=log(OR)
          L <- as.numeric(substr(CI,1,4))
          U <- as.numeric(substr(CI,6,9))
          se=abs(log(L)-log(U))/3.92
          P=2*(1-pnorm(abs(beta/se)))
       }) %>% select(
          SNP=rsid,
          A1=allele1,
          A2=allele2,
          freqA1=freq1,
          beta,
          se,
          P,
          N,
          chr,
          pos=position_b37
       )
write.table(ST4, file="ST4", row.names=FALSE, col.names=FALSE, quote=FALSE)

END
```

where we generate data based on the paper's supplementary tables ST3 and ST4; the former is in line with the paper (by specifying `_db`=depict and `p_threshold`=0.00001 when calling PW-pipeline).

References

GIANT (Genetic Investigation of ANthropometric Traits) data

Locke AE, et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518(7538):197-206. doi: 10.1038/nature14177

DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) data

Scott R, et al. (2017) An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes 66:2888–2902.