UKB_relatedness

Hakhamanesh Mostafavi

5/14/2020

```
library(data.table)
library(tidyverse)
library(igraph)

home_dir="~/sherlock/ukbiobank/data/"
qc_file=paste0(home_dir,"sample_qc/sample_qc_all.fam") # set of individuals passe
d QC
rel_file=paste0(home_dir,"sample_qc/ukb2498_rel_s488374.dat") # relatedness table

d_qc=fread(qc_file,header=F); colnames(d_qc)=c("FAMID","IID")
d_rel=fread(rel_file)
```

```
print(paste("number of related pairs before QC = ", as.character(nrow(d_rel))))
d_rel_filter=d_rel[ ((d_rel$ID1 %in% d_qc$IID) & (d_rel$ID2 %in% d_qc$IID)),]
print(paste("number of related pairs after QC = ", as.character(nrow(d_rel_filte
r))))
```

```
calc_related_samples <- function(df,b1,b2,degree){
  d=df[(df$Kinship<b1) & (df$Kinship>b2),]
  N_pairs=nrow(d)
  N_samples=length(unique(c(d$ID1,d$ID2)))
  print(paste0(degree,": #pairs, #samples = ", as.character(N_pairs), ", ", as.ch
aracter(N_samples)))
}

## KING table: {mono twins,1st,2nd,3rd}=1/{2^1.5,2^2.5,2^3.5,2^4.5}
calc_related_samples(d_rel_filter,0.5,(1/2^1.5),"Monozygotic twins")
calc_related_samples(d_rel_filter,(1/2^1.5),(1/2^2.5),"1st degree")
calc_related_samples(d_rel_filter,(1/2^2.5),(1/2^3.5),"2nd degree")
calc_related_samples(d_rel_filter,(1/2^3.5),(1/2^4.5),"3rd degree")
calc_related_samples(d_rel_filter,0.5,(1/2^4.5),"3rd degree or closer")
```

```
df <- select(d_rel_filter,ID1,ID2)
df$ID1=as.character(df$ID1); df$ID2=as.character(df$ID2)
G=graph_from_data_frame(df, directed = FALSE, vertices = NULL)

clst=clusters(G) #identify clusters
dj=as.data.frame(clst$membership) #membership info per individual
dj$ID=as.character(rownames(dj)); colnames(dj) <- c("membership", "ID")
dk=merge(df,dj, by.x="ID1", by.y="ID") #membership info per pair

# manually chose two clusters (3 and 5) to plot
G3=graph_from_data_frame(dk[dk$membership==3,], directed = FALSE, vertices = NUL
L)
G5=graph_from_data_frame(dk[dk$membership==5,], directed = FALSE, vertices = NUL
L)

plot(G3)
plot(G5)
```

```
library(igraph)

max_unrelated_sample <- function(d_qc,dg,thresh,degree){

  df <- select(dg[(dg$Kinship>thresh),],ID1,ID2)
  df$ID1=as.character(df$ID1); df$ID2=as.character(df$ID2); row.names(df)=1:nrow
(df)
  G=graph_from_data_frame(df, directed = FALSE, vertices = NULL)

  #### FAILED attempt, sample size too large for this approach

  # s=largest_ivs(G) #returns all possible sets (nodes to keep)
  # mat=as.matrix(sapply(s[1], as_ids)) #select the first set
  # list_ids=c(mat) #final output of unrelated sample IDs

  #### in-house implementation (analyze one cluster at a time)

  clst=clusters(G) #identify clusters

  dj=as.data.frame(clst$membership) #membership info per individual
  dj$ID=as.character(rownames(dj)); colnames(dj) <- c("membership", "ID")
  dk=merge(df,dj, by.x="ID1", by.y="ID") #membership info per pair

  #loop over clusters
  list_ids=vector() # individuals to keep from d_rel_filter data
  for(k in seq(1,max(dk$membership))) {

    dt=dk[dk$membership==k,]
    A_temp=dt[,1:2]
    G_temp=graph_from_data_frame(A_temp, directed = FALSE, vertices = NULL)
    s_temp=largest_ivs(G_temp)
    mat=as.matrix(sapply(s_temp[1], as_ids))
    list_ids=c(list_ids,mat)
  }

  all_indivs=unique(c(df$ID1,df$ID2)) # all QCed indivs
  list_related_ids=all_indivs[!(all_indivs %in% list_ids)] # related individuals
to remove

  N_unrelated=nrow(d_qc[!(d_qc$IID %in% list_related_ids),])
  print(paste0(degree," or closer: #unrelated indivs = ", as.character(N_unrelate
d)))

}

## KING table: {mono twins,1st,2nd,3rd}=1/{2^1.5,2^2.5,2^3.5,2^4.5}
max_unrelated_sample(d_qc,d_rel_filter,(1/2^1.5),"Monozygotic twins")
max_unrelated_sample(d_qc,d_rel_filter,(1/2^2.5),"1st degree")
```

```
max_unrelated_sample(d_qc,d_rel_filter,(1/2^3.5),"2nd degree")
max_unrelated_sample(d_qc,d_rel_filter,(1/2^4.5),"3rd degree")
```