Ldpred2

Bayesian approach to computing polygenic risk scores

Computing Sources: Google Cloud Platform(GCP), Nipa Ubuntu Server(고성능 컴퓨팅 서비스), Texas Advanced Computing Center(TACC)

Google Cloud Platform

- Session logged out too quickly

Nipa Ubuntu Server

- Extremely slow for finding correlation on one chromosome on R
- Downloaded R studio but could not access it through https://localhost:8787 (Still do not know why)

Texas Advanced Computing Center

- A very nice stampede guideline: https://portal.tacc.utexas.edu/user-guides/stampede2

Accessing TACC

- Create an account on Xsede site and connect it with duo mobile application
- Through terminal, log into xsede site by typing ssh -l your_*xsede_id* login.xsede.org and then your password
- Then, type gsissh stampede2 to go into TACC

File location

Your directory

```
ABCD.bed                chr12.OMNI.interpolated_genetic_map
ABCD.bim                chr13.OMNI.interpolated_genetic_map
ABCD.bk                 chr14.OMNI.interpolated_genetic_map
ABCD.fam                chr15.OMNI.interpolated_genetic_map
ABCD.rds                chr16.OMNI.interpolated_genetic_map
ABCD.valid.sample       chr17.OMNI.interpolated_genetic_map
ABCD_QCed.bed           chr18.OMNI.interpolated_genetic_map
ABCD_QCed.bim           chr19.OMNI.interpolated_genetic_map
ABCD_QCed.bk            chr2.OMNI.interpolated_genetic_map
ABCD_QCed.fam           chr20.OMNI.interpolated_genetic_map
ABCD_QCed.het           chr21.OMNI.interpolated_genetic_map
ABCD_QCed.log           chr22.OMNI.interpolated_genetic_map
ABCD_QCed.nosex         chr3.OMNI.interpolated_genetic_map
ABCD_QCed.prune.in                   chr4.OMNI.interpolated_genetic_map
ABCD_QCed.prune.out                  chr5.OMNI.interpolated_genetic_map
ABCD_QCed.rel.id                   chr6.OMNI.interpolated_genetic_map
ABCD_QCed.snplist                   chr7.OMNI.interpolated_genetic_map
chr1.OMNI.interpolated_genetic_map  chr8.OMNI.interpolated_genetic_map
chr10.OMNI.interpolated_genetic_map  chr9.OMNI.interpolated_genetic_map
chr11.OMNI.interpolated_genetic_map  pca
```

- GWAS summary stat file: directory private!
- > It should look like this…

```
CNCR_AD                PGC_EATING.txt
CNCR_ANTISOCIAL.txt      PGC_OCD
CNCR_DEPRESSION.txt      PGC_SCZ
CNCR_DEPRESSION_SUB.txt  PGC_UKB_MDD
CNCR_IQ.txt             SSAGC_ASP.txt
CNCR_NEUROTICISM.txt     SSAGC_DRINK.txt
CNCR_WORRY_SUB.txt       SSAGC_RISK4PC.txt
ETC_INSOMNIA            SSAGC_RISKTOL_MA.txt
ETC_PTSD_EA             SSAGC_SMOKER_MA.txt
ETC_SNORING.txt         UKB_AUDIT.txt
GWAS_CP_all.txt         UKB_BMI.txt
GWAS_CP_all_ldpred.txt   UKB_CANNABIS.txt
GWAS_CP_all_ldpred2.txt  UKB_GENERALHAPPINESS.txt
GWAS_EA_excl23andMe.txt  UKB_GENERALHAPPINESS_HEALTH.txt
PGC_ADHD_EA             UKB_GENERALHAPPINESS_MEANINGFUL.txt
PGC_ASD                UKB_HAPPINESS.txt
PGC_BIP_2018           adas
PGC_CROSS.txt
```

R script for PRS computation (my case: bipolar disorder)

- Saved in directory private
- The code looks like this.. (Blue comments are for additional info)

```
#install.packages("dplyr")
library(bigsnpr)
options(bigstatsr.check.parallel.blas = FALSE) # For multi-thread


obj.bigSNP <- snp_attach("privaate_____________ABCD_QCed.rds")
str(obj.bigSNP, max.level = 2, strict.width = "cut")

G   <- obj.bigSNP$genotypes
CHR <- obj.bigSNP$map$chromosome
POS <- obj.bigSNP$map$physical.pos
y   <- obj.bigSNP$fam$affection - 1

sumstats <- bigreadr::fread2("/private directory…../ABCD_summarystats/PGC_BIP_2018")
#If you wish to find PRS on other GWAS summarystats change PGC_BIP_2018 part
str(sumstats)
```

```
set.seed(1)
ind.val <- sample(nrow(G), 400)
ind.test <- setdiff(rows_along(G), ind.val)

sumstats$beta <- log(sumstats$OR)
sumstats$n_eff <- 4 / (1 / sumstats$Nca + 1 / sumstats$Nco)
sumstats$Nca <- sumstats$Nco <- NULL
sumstats$HetPVa <- sumstats$HetDf <- sumstats$Direction <-
sumstats$HetISqt<-sumstats$Neff <- sumstats$ngt<- sumstats$INFO <- sumstats$OR <-NULL
sumstats$FRQ_A_20352 <- sumstats$FRQ_U_31358 <- NULL
names(sumstats) <- c("chr", "rsid", "pos", "a0", "a1","beta_se", "p","beta", "n_eff")
# check the format and contents of sumstats by str(sumstats)
map <- obj.bigSNP$map[-(2:3)]
names(map) <- c("chr", "pos", "a0", "a1")
info_snp <- snp_match(sumstats, map)

library(R.utils)
library(data.table)
library(magrittr)

POS2 <- snp_asGeneticPos(CHR, POS, dir = "private___")
# Get maximum amount of cores
NCORES <- nb_cores()
# Start doing analysis on each chromosome
fam.order <- as.data.table(obj.bigSNP$fam)
fam.order[, Inf.est := 0]

# add progress bar
pb = txtProgressBar(min = 0, max = 22, initial = 0)


#inf.model 로 chr 별 결과 각각 print
for (chr in 1:22) {
  setTxtProgressBar(pb, chr)
  # extract current chromosome
  chr.idx <- which(info_snp$chr == chr)
  df_beta <- info_snp[chr.idx,
                      c("beta", "beta_se", "n_eff")]
  ind.chr <- info_snp$`_NUM_ID_`[chr.idx]
  # calculate LD
  corr0 <- snp_cor(
    G,
    ind.col = ind.chr,
    ncores = NCORES,
    infos.pos = POS2[ind.chr],
    size = 3 / 1000
  )
  corr <- bigsparser::as_SFBM(as(corr0, "dgCMatrix"))
  # Perform LDSC analysis to get h2 estimate
  ldsc <- snp_ldsc2(corr0, df_beta)
  h2_est <- ldsc[["h2"]]
  # Get adjusted beta from infinitesimal model
  beta_inf <- snp_ldpred2_inf(corr, df_beta, h2 = h2_est)
  # Get infinitesimal PRS
  pred_inf <- big_prodVec(G,
                          beta_inf,
                          ind.row = ind.test,
                          ind.col = ind.chr)
  # add up the calculated PRS
  fam.order[, Inf.est := Inf.est + pred_inf] # I thought I could add prs scores to
fam.order data by adding another column.
}
print("Completed")
write.csv(fam.order, "private________/bipolarld2.csv")
```

Job submission Script

- This script is saved in private directory…. as bipld2script.sh on TACC

```
#!/bin/bash
#----------------------------------------------------
# Sample Slurm job script
#   for TACC Stampede2 SKX nodes
#----------------------------------------------------

#SBATCH -J bipld2           # Job name
#SBATCH -o bipld2.o%j       # Name of stdout output file
#SBATCH -e bipld2.e%j       # Name of stderr error file
#SBATCH -p skx-dev     # Queue (partition) name / dev도 있고 skx-normal, long 같은 경우는
knl
#SBATCH -N 4               # Total # of nodes (must be 1 for serial) > qlimits 라고
커맨드에 치면 limits 볼 수 있음
#SBATCH -n 32               # Total # of mpi tasks (should be 1 for serial)
#SBATCH -t 2:00:00          # Run time (hh:mm:ss)
#SBATCH --mail-user=private___________
#SBATCH --mail-type=all     # Send email at begin and end of job


# Other commands must follśw all #SBATCH directives...
private________/ld2 #stdout/ err output file goes in! WORKSPACE
module load Rstats
pwd
date

# Launch serial code...

ibrun Rscript bipld2.R # dev 경우 ibrun 했고 아닌경우는 Rscript file명.R

# ----------------------------------------------------
```

- I wanted to run the R script with skx-large, but I couldn't for whatever reason. So I tried it with long. It uses KNL if queue is long so probably inaccurate partition name. This script is saved in ------------------- on TACC

```
login3.stampede2(678)$ cat bipld2large.sh
#!/bin/bash
```

```
#----------------------------------------------------
# Sample Slurm job script
#   for TACC Stampede2 SKX nodes
#----------------------------------------------------

#SBATCH -J bipld2           # Job name
#SBATCH -o bipld2.o%j       # Name of stdout output file
#SBATCH -e bipld2.e%j       # Name of stderr error file
#SBATCH -p long     # Queue (partition) name
#SBATCH -N 1               # Total # of nodes (must be 1 for serial)
#SBATCH -n 1               # Total # of mpi tasks (should be 1 for serial)
#SBATCH -t 24:00:00        # Run time (hh:mm:ss)
#SBATCH --mail-user=blahblah (private)
#SBATCH --mail-type=all    # Send email at begin and end of job

# Other commands must follśw all #SBATCH directives...
cd -------------------------- private
module load Rstats
pwd
date

# Launch serial code...

Rscript bipld2.R
```

How to run r script on TACC by job shell file

1. Send necessary files for job submission- my case: shell file(.sh) and your rscript

   ```
   (base)Desktop % scp shell_file_name tacc_id@stampede2.tacc.utexas.edu:any_directory_you_wish_to_send_file
   ```

   - Then you will need to enter your tacc account password and tacc return token code(6 digits) from duo mobile

   My case)

   ```
   (base) heewon@Heewonui-MacBookPro Desktop % scp bipld2script.sh _______private_____@stampede2.tacc.utexas.edu:-------------private
   ```

2. Login xsede through terminal

   ```
   ssh -l xsede_id login.xsede.org
   ```

   My case:

   ```
   ssh -l private_ login.xsede.org
   ```

   - And then, enter your password

3. Access to tcaa, stampede2

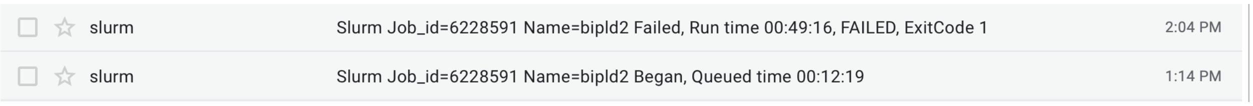
   ```
   gissh stampede2
   ```

4. Go to directory or folder that you sent files to and check if they are correctly sent

5. Submit job using shell file
   ```
   sbatch shell_file_name
   ```
   My case: `sbatch bipld2script.sh`

6. Monitor your job schedule
   ```
   login1$ squeue -u your_tacc_id
   ```
   My id: private______ so `squeue -u private____`

7. If you chose email option in .sh file, you will get an email at the beginning and the end of the job as below. Also, in the same directory/folder in stempede2, output and error files will be created as *project_name.ejob_number,* and *project_name.ojob_number*



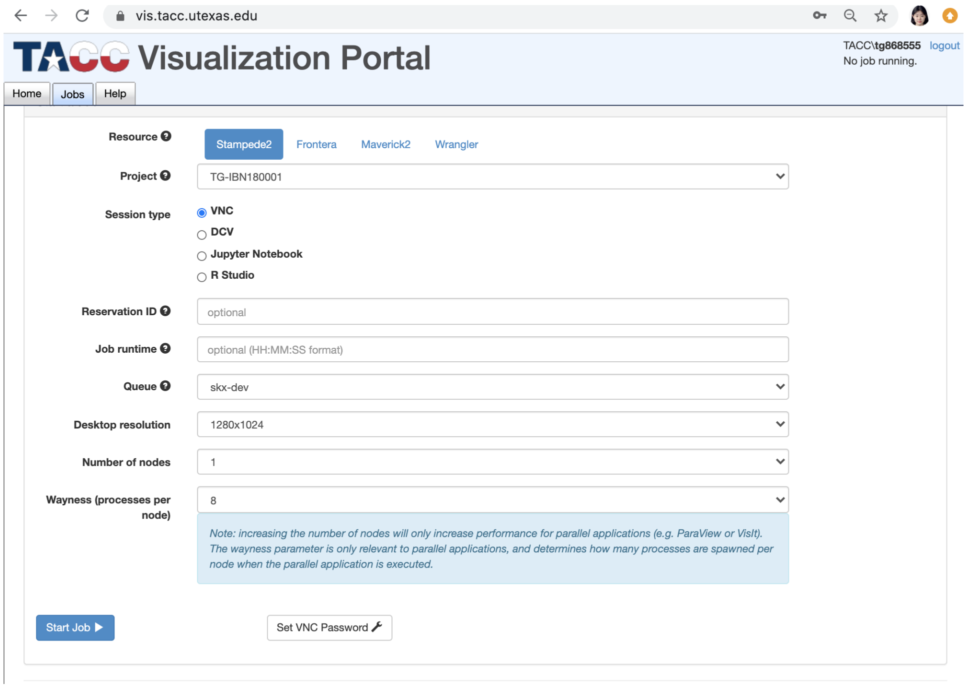
```
☐ ☆ slurm    Slurm Job_id=6228591 Name=bipld2 Failed, Run time 00:49:16, FAILED, ExitCode 1    2:04 PM
☐ ☆ slurm    Slurm Job_id=6228591 Name=bipld2 Began, Queued time 00:12:19                      1:14 PM
```

8. After the job is finished, open output file and error files to see how it went
   ```
   cat bipld2.o6233457 # Output file for job number 6233457
   cat bipld2.e6228979 # Error file for job number 6228979
   ```

Using R studio on TACC Visualization Portal

- I wasn't a hundred percent sure if my Rscipt is working in adding pred_info for all chromosomes at the end in getting the final PRS score
- So I used interactive web based Rstudio on TACC Visualization portal, which was really helpful in checking what values are in data sets, objects, and data frames
- I tried to find snp correlation for chromosome 22

  1) Go to this site https://vis.tacc.utexas.edu/
  2) Sign in to your TACC account(not the same with xsede account be connected) at the top right corner
  3) Go to job section and start interactive R studio which looks like this



vis.tacc.utexas.edu

TACC Visualization Portal

TACC\tg868555 logout
No job running.

Resource: Stampede2 | Frontera | Maverick2 | Wrangler

Project: TG-IBN180001

Session type:
● VNC
○ DCV
○ Jupyter Notebook
○ R Studio

Reservation ID: optional

Job runtime: optional (HH:MM:SS format)

Queue: skx-dev

Desktop resolution: 1280x1024

Number of nodes: 1

Wayness (processes per node): 8

Note: increasing the number of nodes will only increase performance for parallel applications (e.g. ParaView or VisIt). The wayness parameter is only relevant to parallel applications, and determines how many processes are spawned per node when the parallel application is executed.

- the screen freezes often / only 2 hours

Errors and why we failed to get PRS using Ldpred2

- Whether the job was queued in skx-dev or skx-normal, there were error messages saying…

```
에러: 크기가 18.2 Gb인 벡터를 할당할 수 없습니다
실행이 정지되었습니다
경고메시지(들):
시스템 호출에 실패했습니다: 메모리를 할당할 수 없습니다
slurmstepd: error: *** JOB 6229759 ON c458-064 CANCELLED AT 2020-08-13T09:37:21 DUE TO TIME LIMIT ***

Error in validityMethod(as(object, superClass)) :
  아직까지는 지원되지 않는 긴 벡터들입니다: ../../src/include/Rinlinedfuns.h:519
Calls: snp_cor ... validObject -> anyStrings -> isTRUE -> validityMethod
실행이 정지되었습니다
경고메시지(들):
시스템 호출에 실패했습니다: 메모리를 할당할 수 없습니다
```

- In Rstudio TACC visualization portal, it got stuck in the line where the code finds snp_correlation no matter what chromosome number was. (Probably the smallest 22)

  This line: `corr0 <-snp_cor(G,ind.col=ind.chr2,ncores=NCORES,infos.pos=POS2[ind.chr2],size=3/1000)`