**Bivariate GWAS steps:**

1. Data Preparation:

* **Alzheimer disease data (AD)**

Ancestry of the study sample: European

Data source: Bellenguez, C. et al. Nat Genet. 2022; 54:412–36.

Genome build: GRCh38

MAF of SNP: 0.01

LD: 0.2

* **Bone mineral density (femoral neck BMD, forearm BMD, lumber spine BMD)**

Ancestry of the study sample: European

Data source: Zheng et. al. *Nature* 526, 112–117 (2015)

Genome build:  build hg19.

MAF of SNP: 0.5

LD: 0.2

**\*\*\*\*\*\*\*** So far largest eBMD metaanalysis was done by Morris et al. *Nat Genet* 51, 258–266 (2019), however did not use the summary statistic data because of too small P value

**Step 1:** **AD data pre-processing**

Converting AD summary statistics genome build to build hg19 using liftover

**Number of SNPs**: **21101115**

MAF filtering: As the MAF for BMD data is >0.5 should not I filter out more SNPs from AD data?

**Step 2:** **BMD data pre-processing**

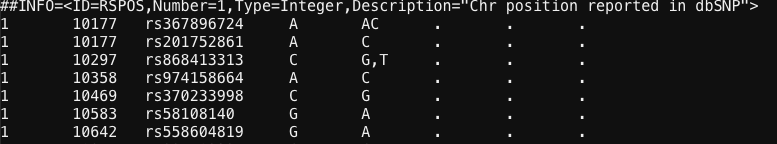
The BMD (fnBMD, faBMD, lsBMD) data does not have any RSID info, has only CHR and POS info, as I am using ELX software for bivariate GWAS and I requires RSID, I got the RSID from dbSNP (00-All37.vcf.gz)

**dbSNP source:**

**Total SNP: 30880449**

https://ftp.ncbi.nih.gov/snp/organisms/[human\_9606\_b151\_GRCh37p13/](https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/) 2018-04-27 12:09

**number of SNPs in fnBMD data**: **10586900**

* I tried to find the matched number of lines in dbSNP with fnBMD data position and the number of matched lines are **27991724,** indicating duplicates, in several cases there are more than one SNPs example below:
* When I merged fnBMD data with DBSNP data based on matched CHR, POS and ALT allele, I lost several SNPs (~30%), **6901000**

**How to prevent SNP losing?**

--- Extracting the duplicated rows from matched lines

---Manually check those lines with my fnBMD data and decide which line to keep based on my reference allele and alternative allele from fnBMD data

Or,

--- Removing the lines from matched data that has more than two nucleotide bases in a position and checking the number, in such case also have to carefully check the bases from fnBMD data

(I checked fnBMD data REF and ALT allele and there is not a single line with more than two base pairs so I followed this process for preventing SNP lose)

---- Extract duplicate lines

---- Remove SNPs with INDEL

--- rejoin with merged file and check

**Bivariate GWAS:**

1. Remove SNPs from AD that has MAF less than 0.5
2. Some papers before doing bivariate or LDSC (Linkage Disequilibrium Score Regression) is a statistical method used to estimate heritability and genetic correlations, excluded high LD SNPs from APOE and HLA region.

Paper 1: The apolipoprotein E (APOE) region has a large effect on the risk of AD; hence, we excluded APOE and the 500 kilobase (kb) flanking region (hg19, 19:44,909,039–45,912,650) from the AD GWAS. We also excluded SNPs in the 26 to 36 megabase region of chromosome six from the data given the complex LD structure in the human major histocompatibility complex (MHC). Notably, in analyses both with and without the APOE region, LDSC reveals a significant genetic correlation between AD and GIT traits (https://doi.org/10.1038/s42003-022-03607-2)

* I can perform with or without these region as most of the papers got both results

1. As AD data and BMD data do not have same number of SNPs included and our objective is to get the pleiotropic effect of SNPs, at first we extracted the common SNPs from both of the data
2. We merged the data.
3. Check whether the bases are harmonized.
4. **P value threshold for SNPs to include for bivariate GWAS (different papers used different thresholds --- I have not decided yet)**

**Paper 1:** Pac Symp Biocomput. 2019 ; 24: 272–283.

Title: **Detecting potential pleiotropy across cardiovascular and neurological diseases using univariate, bivariate, and multivariate methods on 43,870 individuals from the eMERGE network**

we applied a p-value threshold of 0.005 on the univariate GWAS results and filtered out any SNPs that did not meet this threshold. We also filtered out SNPs with MAF = 0.5 to remove ambiguity about which allele was chosen as the referent allele in univariate analyses. Finally, we identified a list of common SNPs and estimated a p-value for each of 2,080 “pairs” of phenotypes using a chi-squared test with two degrees of freedom. Although we conducted a reduced number of tests, it should be noted that we corrected for multiple comparisons using the original “unfiltered” SNP set to control our type I error rate well.

**Paper 2:**

**Title: Bivariate Genome-Wide Association Scan Identifies 6 Novel Loci Associated with Lipid Levels and Coronary Artery Disease**

we narrowed our set of variants to those that were nominally associated with both traits from single-trait association data (***P*<5×10−3** for the lipid trait and CAD). To focus on sites where we could hypothesize that the same causal variant contributes to both the lipid and CAD signals, we filtered out variants where the patterns of association for both traits did not statistically overlap. To determine this, we used the COLOC software,

**Paper 3:**

**Title: Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation**

The results were considered genome-wide significant when (1) the bivariate p-values were < 5 × 10−8 and (2) the bivariate p-value was at least one order of magnitude lower than both individual trait p-values and (3) when the **individual trait p-values were at least nominally significant (p-value<0.05**). When multiple SNPs were significant in a locus, the SNP with the lowest p-value was chosen for replication. The eLC method is imple- mented in eLX package using C++

**Paper 4: they pruned based on z score**

**Multivariate Analysis of Anthropometric Traits Using Summary Statistics of Genome-Wide Association Studies from GIANTConsortium**

SNPs with large effect sizes may represent true association, and consequently may inflate correlation among summary statistics. Therefore, we removed SNPs whose summary statistics Z scores were greater than 1.96 or less than -1.96. The final SNP sets for correlation estimation include 81,322 SNPs for height, 82,012 SNPs for BMI, and 81,130 SNPs for WHRadjBMI.

**Paper 5: Has a very beautifully organized method section in their additional file that I can follow**

**Multi-trait association studies discover pleiotropic loci between Alzheimer’s disease and cardiometabolic traits**

Given that Z-scores should be standard 8 normal distributed, we can estimate the bivariate normal distribution of the SNP Z-scores. We 9 can then perform an association test for each SNP using a chi-squared test with two degrees of 10 freedom to calculate a bivariate P-value(2–4). 11 The first step to perform a bivariate scan is to harmonize the alleles of the summary 12 statistics data from the two GWAS studies using the harmonise\_data function in MRbase(5). We 13 next use PLINK to obtain independent SNPs using the command “--indep-pairwise 1000 5 0.2” 14 with the 1000 Genomes phase 3 European ancestry (1kG EUR) as the linkage reference 15 panel(6). We estimate the mean and covariance matrix of the bivariate normal distribution of 16 SNP Z-scores using the independent SNPs. As described above, we then used these 17 parameters to calculate a bivariate P-value for each harmonized SNP(2–4). Independent 18 associated loci were produced using the “--clump-r2 0.2” command in PLINK(1). We further 19 filtered independent associated loci differently for the “AD-centric” and the “locus discovery” 20 experiment to detect loci of interest (see Materials and Methods: AD-centric Analysis and Locus 21 Discovery Analysis).

\*\*\*\* **All these papers and this analytic plan are saved in lary lab in team Bivariate GWAS/ Paper to follow new folder**

**4/3/2024**

**BMD- extracted dbsnp data from build 37 -------30,880,449**

**No of SNP in fnbmd sumstat. ------ 10,586,900**

**Extract matched lines between these two based on chr and pos ----- 10,765,637**

**Checked the duplicates and most of the cases there were INDELS (more than two bases)**

**Removed the indels ------- fnbmd matched lines no indel ----- 10,535,335**

**Fnbmd data sample after merging now ----- 10,535,335**

**AD Number of SNPs**: **21,101,114**

**AD common SNPS with fnbmd and AD and removing NA rows to get the complete data ------------------------ 9,295,515**

**Unique SNPs after harmonization ----** 9,286,380

**SNP with P-values <0.05 for both traits ----- 42,852**

**After analysis number of total SNP with P-value <5x10-8 if consider dLC.pval --- 800**

**After analysis number of total SNP with P-value <5x10-8 if consider OB.pval ---320**

**Based on dLC.pval we found total in chr1 ----- 2 hits**

**Chr2----2 hits**

**Chr4 ---1 hit**

**Chr6---2 hits**

**Chr7---1 hit**

**Chr8---1 hit**

**Chr10---2 hits**

**Chr11---3 hits**

**Chr 12—1 hit**

**Chr 13-1 hit**

**Chr 15-1 hit**

**Chr 16—5 hits**

**Chr 17-3 hits /probably 2 hits as 1 big ld region**

**Chr19-3 hits**

**##### Date: 27/02/2024**

**#### Author: Nayema**

**#### Merging file containing rsid from dbsnp with bmd summary statistic files that does not have rsid**

**##################################################**

**#1. Set environment**

**###############################**

**getwd()**

**setwd("/work/larylab/NAYEMA/BIVARIATE\_GWAS/")**

**######################################################################################**

**#2. Load Packages**

**######################################################################################**

**#First make sure you have the biomaRt and devtools packages installed:**

**#Load packages**

**#install.packages("devtools")**

**library(devtools)**

**#install\_github("MRCIEU/TwoSampleMR")**

**library(TwoSampleMR)**

**#install.packages("ggplot2")**

**library(ggplot2)**

**#install.packages("knitr")**

**library(knitr)**

**#install\_github("MRCIEU/MRInstruments")**

**#The current results were achieved with using version TwoSampleMR\_0.4.11**

**library(MRInstruments)**

**#install.packages("reshape")**

**library("reshape")**

**library(dplyr)**

**#####################**

**#3. read the summary statistics and the matched file from dbsnp for getting RSID**

**#####################**

**#### femoral neck -fnbmd summary statistics data for european people downloaded from musculoskeleta knowledge portal**

**#### link http://www.gefos.org/?q=content/data-release-2015**

**## from dbsnp data where the values in the CHROM and POS columns match with the corresponding values in fn2stu.MAF0\_.005.pos\_.out\_, fa2stu.MAF0\_.005.pos\_.out\_, ls2stu.MAF0\_.005.pos.out\_**

**############ fnbmd**

**fn2stu <- read.table("fn2stu.MAF0\_.005.pos.out\_", header = TRUE)**

**matched\_lines <- read.table("matched\_lines\_fnbmd\_no\_indel.txt", header=FALSE) # lines of fnbmd based on chr and pos with dbsnp build 37 data**

**#check the data**

**head(fn2stu)**

**head(matched\_lines)**

**####### give them column names**

**colnames(matched\_lines) <- c("chromosome","position","snp","other\_allele", "reference\_allele",".", "..1", "..2")**

**matched\_lines <- matched\_lines[, 1:5]# i need only these 5 columns**

**# Merge matched line with fnbmd based on matching values in specified columns to get rsid**

**merged\_datafn <- merge(matched\_lines, fn2stu, by = c("chromosome", "position"))**

**head(merged\_datafn)**

**##### extract only required columns**

**merged\_datafn2 <- merged\_datafn[, c(1:3, 6:11, 15)]**

**head(merged\_datafn2)**

**#change column names**

**colnames(merged\_datafn2) <- c("chromosome","position","variant\_id", "rs\_number","reference\_allele.fnbmd","other\_allele.fnbmd","eaf.fnbmd","beta.fnbmd","se.fnbmd","p.value.fnbmd")**

**######### save the fnbmd data**

**# Write the data frame to a file without row names**

**write.table(merged\_datafn2, file = "/work/larylab/NAYEMA/BIVARIATE\_GWAS/Bivariate\_gwas/data/fnbmd\_data\_with\_rsid.txt", sep = "\t", quote = FALSE, row.names = FALSE)**

**#############################**

**############ Alzheimers data , it was converted from build 38 to 37 using liftover**

**data\_ad<-read.table("AD.txt",fill = TRUE, header = TRUE)**

**head(data\_ad)**

**save.image(file='myEnvironment.RData')**

**####################################################**

**# 4. merge fnbmd with AD**

**####################################################**

**merged\_datafnbmd\_ad <- merge( merged\_datafn2, data\_ad, by = "variant\_id", all.x = TRUE)**

**head(merged\_datafnbmd\_ad)**

**# Remove rows with NA values which means this variant\_id is not present in AD\_data**

**subset\_merged\_datafnbmd\_ad <- merged\_datafnbmd\_ad[complete.cases(merged\_datafnbmd\_ad), ]**

**head(subset\_merged\_datafnbmd\_ad)**

**############## save merged data**

**write.table(subset\_merged\_datafnbmd\_ad, file = "/work/larylab/NAYEMA/BIVARIATE\_GWAS/Bivariate\_gwas/data/merged\_fnbmd\_ad\_common\_variant\_id.txt", sep = "\t", quote = FALSE, row.names = FALSE)**

**###############################**

**## Extract only columns required**

**subset\_merged\_datafnbmd\_ad2 <- subset\_merged\_datafnbmd\_ad[, c(1:16, 20:23)]**

**head(subset\_merged\_datafnbmd\_ad2)**

**### Rename the columns for harmonization**

**colnames(subset\_merged\_datafnbmd\_ad2) <- c("SNP", "chromosome.outcome", "position.outcome", "rs\_number", "effect\_allele.outcome",**

**"other\_allele.outcome", "eaf.outcome", "beta.outcome", "se.outcome", "pval.outcome","pval.exposure",**

**"chromosome.exposure", "position.exposure", "effect\_allele.exposure", "other\_allele.exposure", "eaf.exposure","beta.exposure",**

**"se.exposure","n.cases","n.control")**

**############################**

**### The alleles for AD and fnbmd are not harmonized therefore need to harmonize them**

**################################**

**#LOAD IN AD DATA WITH FNBMD FOR COMMON SNPS**

**AD\_data <- subset\_merged\_datafnbmd\_ad2[, c("SNP", "chromosome.exposure", "position.exposure", "effect\_allele.exposure", "other\_allele.exposure","beta.exposure",**

**"se.exposure","eaf.exposure","pval.exposure")]**

**head(AD\_data)**

**str(AD\_data)**

**AD\_data1$exposure<-"AD"**

**AD\_data1$id.exposure<-"AD.1"**

**#LOAD IN FNBMD DATA WITH AD FOR COMMON SNPS**

**AD\_data <- subset\_merged\_datafnbmd\_ad2[, c("SNP", "chromosome.exposure", "position.exposure", "effect\_allele.exposure", "other\_allele.exposure","beta.exposure",**

**"se.exposure","eaf.exposure","pval.exposure")]**

**head(AD\_data)**

**str(AD\_data)**

**AD\_data1$exposure<-"AD"**

**AD\_data1$id.exposure<-"AD.1"**

**pQTL**

1. GWAS of genetic variants associated with bmd
2. GWAS of genetic variants associated with AD
3. **PLEOTROPIC VARIANTS BETWEEN THESE TWO TRAITS**
4. **DIRECTION OF EFFECT should be opposite (ex. Having negative effect on Alzheimer but positive effect on bmd)\*\*\*\*\*\*\*\*\***
5. **DO I NEED SEPARATE OR SAME DIRCTION?\*\*\*\*\*\*\*\***
6. integrated pQTL variants with genetic variants from the MUSCULOSKELETAL KNOWLEDGE PORTAL  and employed Mendelian randomization (MR)[16](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6093935/#CR16) to reveal proteins with potentially causal effects on AD.  OR VICE VERSA

awk -F'\t' '$2 == $4 && $3 == $5' harmonised\_fnbmd\_ad\_common\_variant\_id.txt > unique\_harmonised\_fnbmd\_ad\_common\_variant\_id.txt