국립보건연구원
National Institute of Health

# 한국인칩 유전체정보 분석(2)
# 실습

국립보건연구원 미래의료연구부 유전체연구기술개발과

# 목차



- Data Description
- Single variant test
- Gene-based test
- Meta analysis
- PRS analysis

# Data Description



- **Subject**
  - **Artificial Participants from KoGES** (Korean Genome and Epidemiology Study)
- **Genotype data**
  - **Imputed genotype data, Chromosome 16**
  - **platform : Korea Biobank Array (KBA)**
- **Phenotype data**
  - **high density lipoprotein cholesterol (HDL)**
- **Public data**
  - **BBJ HDL GWAS summary statistic (META, PRS)**
  - **1KP EAS 504 sample (PRS – ref.panel)**

# A Summary of Genomic Data Analysis

국립보건연구원 National Institute of Health

| Analysis | Tool | Input Data |
|---|---|---|
| **Single variant test** | EPACTS | VCF file for Genotypes<br>PED file for Phenotypes and Covariates |
| **Gene-based test** | EPACTS | Annotated VCF file<br>PED file for Phenotypes and Covariates |
| **Meta analysis** | METAL | TXT file with association result |
| **PRS analysis** | PRS-CS<br>plink | PLINK compact binary file format (*.bed, *.bim, and *.fam)<br>TXT file with association result |

**Program Download**

- **EPACTS** : http://csg.sph.umich.edu/kang/epacts/download/index.html
- **METAL** : http://csg.sph.umich.edu//abecasis/Metal/download/
- **Plink** : https://www.cog-genomics.org/plink2
- **PRS-CS** : https://github.com/getian107/PRScs

# DATA directory

국립보건연구원
National Institute of Health

- 01_EPACTS : single-variants test, gene-based test
- 02_META : meta analysis
- 03_PRS : PRS analysis
- 04_OUTPUT : analsys results file

# Single variant test

## Statistical tool: EPACTS
## (Effcient and Parallelizable Association Container Toolbox)

Ref. https://genome.sph.umich.edu/wiki/EPACTS

# Single Variant Test



- **Phenotype 데이터에 따른 regression analysis**
  - Quantitative trait(QT)의 경우 Linear regression analysis
  - Case-Control의 경우 Logistic regression analysis
- **EPACTS supported statistical tests**

| Test Name | Phenotypes | Covariates | Computational Time | Description | Implemented by |
|---|---|---|---|---|---|
| b.wald | Binary | YES (Joint) | Slow | Logisitic Wald Test | Hyun Min Kang (simply used glm in R) |
| b.score | Binary | YES (Regressed Out) | Fast | Logistic Score Test (from Lin DY and Tang ZZ, AJHG 2011 89:354-67) | Clement Ma & Hyun Min Kang |
| b.firth | Binary | YES (Joint) | Slow | Firth Bias-Corrected Logistic Likelihood Ratio Test | Clement Ma |
| b.lrt | Binary | YES (Joint) | Slow | Likelihood Ratio Test | Clement Ma |
| b.glrt | Binary | NO | Fast | Genotype Likelihood Ratio Test (use GL or PL field in VCF to perform case-control test) | Hyun Min Kang |
| q.lm | Quantitative | YES (Joint) | Slow | Linear Wald Test | Hyun Min Kang (as implemented in lm in R) |
| q.score | Quantitative | YES (Regressed Out) | Fast | Quantitative Score Test (from Lin DY and Tang ZZ, AJHG 2011 89:354-67) | Clement Ma |
| q.linear | Quantitative | YES (Regressed Out) | Fast | Linear Wald Test | Hyun Min Kang |
| q.reverse | Quantitative | YES (Joint) | Slow | Reverse regression of phenotypes on binary genotypes (dominant model) | Hyun Min Kang |
| q.wilcox | Quantitative | YES (Regressed Out First) | Slow | Nonparametric Reverse regression of phenotypes on binary genotypes (dominant model) | Hyun Min Kang |
| q.emmax | Quantitative | YES (Regressed Out First) | Slow | EMMAX ( Kang et al (2010) Nat Genet 42:348-54 ) | Hyun Min Kang |

# Single Variant Test : Input Data

국립보건연구원
National Institute of Health

- **Genotype data (Variant Call Format , VCF)**

```
$ less -NS KOGO.vcf.gz
```

```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##filedate=2019.7.16
##source=Minimac4.v1.0.1
##contig=<ID=16>
##INFO=<ID=AF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Minor Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy (R-square)">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
##INFO=<ID=IMPUTED,Number=0,Type=Flag,Description="Marker was imputed but NOT genotyped">
##INFO=<ID=TYPED,Number=0,Type=Flag,Description="Marker was genotyped AND imputed">
##INFO=<ID=TYPED_ONLY,Number=0,Type=Flag,Description="Marker was genotyped but NOT imputed">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1">
##minimac4_Command=minimac4 --mapFile genetic_map_chr16_combined_b37.txt --refHaps 16.1000g.Phase3.v5.With.Parameter.Estimates.m3vcf.gz --haps phasing.chr16.v
##bcftools_viewVersion=1.9-207-g2299ab6+htslib-1.9-271-g6738132
##bcftools_viewCommand=view -i 'MAF >= 0.001' KOGO.dose.vcf.gz; Date=Tue Jul 16 17:08:11 2019
##bcftools_viewCommand=view --targets 16:54000001-56000000 KOGO.vcf.gz; Date=Tue Jul 16 18:39:00 2019
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  ID00502 ID00503 ID00504 ID00505 ID00506 ID00507 ID00508 ID00509 ID00510 ID00511 ID0051
16      54000023        16:54000023:G:A G       A       .       PASS    AF=0.00105;MAF=0.00105;R2=0.00781;IMPUTED       GT:DS:GP        0|0:0.005:0.995,0.005,
16      54000190        16:54000190:T:C T       C       .       PASS    AF=0.00118;MAF=0.00118;R2=0.02625;IMPUTED       GT:DS:GP        0|0:0:1,0,0     0|0:0.
16      54000198        16:54000198:C:T C       T       .       PASS    AF=0.01018;MAF=0.01018;R2=0.01336;IMPUTED       GT:DS:GP        0|0:0.017:0.983,0.017,
16      54000220        16:54000220:A:G A       G       .       PASS    AF=0.00112;MAF=0.00112;R2=0.02626;IMPUTED       GT:DS:GP        0|0:0.004:0.996,0.004,
16      54000261        16:54000261:A:T A       T       .       PASS    AF=0.00873;MAF=0.00873;R2=0.01516;IMPUTED       GT:DS:GP        0|0:0.014:0.986,0.014,
```

**## Meta information lines**
**# Header line**
**Each line contains variant informarion such as chromosomal position and individual genotypes**

# Single Variant Test : Input Data

국립보건연구원 National Institute of Health

- **Phenotype (phenotype with pedigree, PED)**

```
$ head KOGO.HDL.ped
```

| #IND_ID | FAM_ID | PAT_ID | MAT_ID | AGE (Covariate) | SEX (Covariate) | HDL (Phenotype) |
|---|---|---|---|---|---|---|
| ID00001 | ID00001 | 0 | 0 | 58 | 1 | 47.746009043413 |
| ID00002 | ID00002 | 0 | 0 | 41 | 2 | 36.6247508915128 |
| ID00003 | ID00003 | 0 | 0 | 61 | 1 | 42.0850347223023 |
| ID00004 | ID00004 | 0 | 0 | 54 | 1 | 50.7957523405631 |
| ID00005 | ID00005 | 0 | 0 | 66 | 1 | 69.977326067975 |
| ID00006 | ID00006 | 0 | 0 | 61 | 1 | 47.4450279047399 |
| ID00007 | ID00007 | 0 | 0 | 41 | 2 | 37.1970715090123 |
| ID00008 | ID00008 | 0 | 0 | 67 | 1 | 30.9517163742377 |
| ID00009 | ID00009 | 0 | 0 | 50 | 1 | 49.0440968362163 |

# Single Variant Test : Analysis Command

국립보건연구원 National Institute of Health

- **Running**

```
$ epacts-single \
     --vcf  KOGO.vcf.gz \
     --ped KOGO.Phenotype.ped --pheno HDL --cov AGE --cov SEX -no-plot \
     --test q.linear --field DS --min-mac 5 --min-maf 0.01 \
     --min-callrate 0.95 --missing NA --region 16:54770000-55070000 \
     --out ../04_OUTPUT/KOGO.HDL.single.q.linear --run 4
```

- **single:** option for single variant test
- **--vcf:** input file in VCF format
- **--ped:** input file in PED format
- **--pheno:** phenotype in PED file
- **--cov:** covariate in PED file
- **--test q.linear:** Linear Wald Test
- **--min-mac:** minimum minor allele count (MAC)
- **--min-maf:** minimum minor allele frequency(MAF)
- **--min-callrate:** minimum call rate
- **--chr:** chromosome #
- **--region :** position begin, end
- **--out:** output file name
- **--run:** algorithm

# Single Variant Test : Analysis Result

국립보건연구원
National Institute of Health

- **Result files**
  - **Output Text of All Test Statistics**
  - **Output Text of Top Associations (top 5,000)**
  - **Q-Q plot of test statistics (stratified by MAF)**
  - **Manhattan Plot of association results**

# Single Variant Test : Analysis Result

국립보건연구원 National Institute of Health

- **Output Text of All Test Statistics**

```
$ less -NS ../04_OUTPUT/KOGO.HDL.single.q.linear.epacts.gz
```

| #CHROM | BEGIN | END | MARKER_ID | NS | AC | CALLRATE | MAF | PVALUE | BETA | SEBETA | TSTAT | R2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 54770105 | | 54770105 | 16:54770105_T/C_16:54770105:T:C | 34497 | 706.82 | 1 | 0.010245 | | 0.59445 | 1.0046 | 1.8869 | 0.53241 8.2 |
| 16 | 54770232 | | 54770232 | 16:54770232_C/T_16:54770232:C:T | 34497 | 25478 | 1 | 0.36929 | 0.64377 | 0.21666 | 0.46852 | 0.46244 | 6.1993e-06 |
| 16 | 54770268 | | 54770268 | 16:54770268_G/A_16:54770268:G:A | 34497 | 7010.3 | 1 | 0.10161 | 0.50391 | 0.50485 | 0.75536 | 0.66836 | 1.295e-05 |
| 16 | 54770445 | | 54770445 | 16:54770445_C/T_16:54770445:C:T | 34497 | 15816 | 1 | 0.22924 | 0.52231 | 0.30832 | 0.48189 | 0.6398 | 1.1867e-05 |
| 16 | 54770476 | | 54770476 | 16:54770476_C/T_16:54770476:C:T | 34497 | 16357 | 1 | 0.23707 | 0.46343 | 0.36211 | 0.49386 | 0.73321 | 1.5585e-05 |
| 16 | 54770477 | | 54770477 | 16:54770477_TC/T_16:54770477:TC:T | | 34497 | 16357 | 1 | 0.23707 | 0.46343 | 0.36211 | 0.49386 | 0.73321 1.5 |
| 16 | 54770478 | | 54770478 | 16:54770478_C/T_16:54770478:C:T | 34497 | 25469 | 1 | 0.36915 | 0.64351 | 0.21686 | 0.46858 | 0.4628 | 6.2092e-06 |

- **NS**: number of phenotyped samples with non-missing genotypes
- **AC**: total non-reference allele count
- **CALLRATE**: fraction of non-missing genotypes
- **MAF**: minor allele frequencies
- **PVALUE**: p-value of single variant test
- **BETA :** effect size of single variant test
- **SEBETA :** standard error of single variant test
- **TSTAT:** score test statistics
- **R2**: varinace explained

# Gene-based test

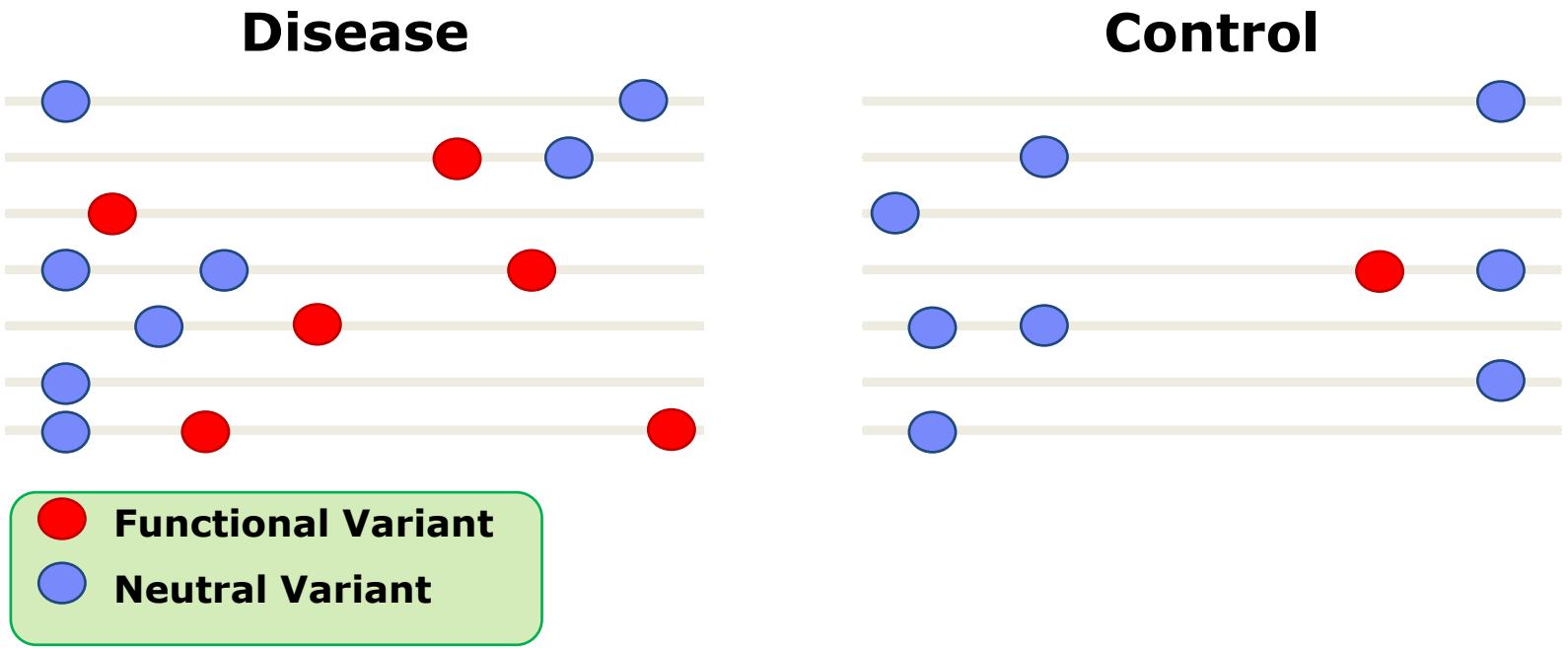
## Statistical tool: EPACTS
## (Effcient and Parallelizable Association Container Toolbox)

Ref. https://genome.sph.umich.edu/wiki/EPACTS

# Gene-based burden test

국립보건연구원
National Institute of Health

- **Rare variants (MAF < 1%)에 대한 효율적인 연관성 분석 방법**
  - rare variants 단일 분석 시 통계 검정력이 낮음 -> **large sample 요구됨**
- **Gene 단위에서 rare allele을 가지고 있는 사람의 비율 비교 분석**



# Gene-based burden test



- **Step 1 : Make input data**
  - annotated file of genotype data (vcf)
  - group definition file based on annotated vcf file
- **Step 2 : Gene-based burden test**

# Gene-based burden test : Input Data

국립보건연구원
National Institute of Health

- **Annotating: VCF file using EPACTS (실습 X)**

```
$ epacts-anno --buildver hg19 --in KOGO.vcf.gz \
     --out ../04_OUTPUT/KOGO.anno.vcf.gz
$ tabix -f -p vcf ../04_OUTPU/KOGO.anno.vcf.gz   #필요 시/
```

- **anno:** annotation
- **--buildver**: reference genome build version
- **--in:** input file name
- **--out:** output file name
- **tabix**: vcf indexing software

# Gene-based burden test : Input Data

국립보건연구원 National Institute of Health

- **Annotated genotype data**

```
$ less -NS KOGO.anno.vcf.gz
```

```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##filedate=2019.7.16
##source=Minimac4.v1.0.1
##contig=<ID=16>
##INFO=<ID=AF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Minor Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy (R-square)">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
##INFO=<ID=IMPUTED,Number=0,Type=Flag,Description="Marker was imputed but NOT genotyped">
##INFO=<ID=TYPED,Number=0,Type=Flag,Description="Marker was genotyped AND imputed">
##INFO=<ID=TYPED_ONLY,Number=0,Type=Flag,Description="Marker was genotyped but NOT imputed">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1">
##minimac4_Command=minimac4 --mapFile genetic_map_chr16_combined_b37.txt --refHaps 16.1000g.Phase3.v5.With.Parameter.Estimates.m3vcf.gz --haps phasing.chr16.vcf --chr 16 --start 54000001 --e
##bcftools_viewVersion=1.9-207-g2299ab6+htslib-1.9-271-g6738132
##bcftools_viewCommand=view -i 'MAF >= 0.001' KOGO.dose.vcf.gz; Date=Tue Jul 16 17:08:11 2019
##bcftools_viewCommand=view --targets 16:54000001-56000000 KOGO.vcf.gz; Date=Tue Jul 16 18:39:00 2019
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  ID00502 ID00503 ID00504 ID00505 ID00506 ID00507 ID00508 ID00509 ID00510 ID00511 ID00512 ID00513 ID00514 ID00515 ID0051
16      54000023        16:54000023:G:A G       A       .       PASS    AF=0.00105;MAF=0.00105;R2=0.00781;IMPUTE;ANNO=Intron:FTO;ANNOFULL=FTO/ENST00000471389.1:+:Intron|FTO/ENST00000394647.
16      54000190        16:54000190:T:C T       C       .       PASS    AF=0.00118;MAF=0.00118;R2=0.02625;IMPUTE;ANNO=Intron:FTO;ANNOFULL=FTO/ENST00000471389.1:+:Intron|FTO/ENST00000394647.
16      54000198        16:54000198:C:T C       T       .       PASS    AF=0.01018;MAF=0.01018;R2=0.01336;IMPUTE;ANNO=Intron:FTO;ANNOFULL=FTO/ENST00000471389.1:+:Intron|FTO/ENST00000394647.
16      54000220        16:54000220:A:G A       G       .       PASS    AF=0.00112;MAF=0.00112;R2=0.02626;IMPUTE;ANNO=Intron:FTO;ANNOFULL=FTO/ENST00000471389.1:+:Intron|FTO/ENST00000394647.
```

```
ANNO=Intron:FTO;ANNOFULL=FTO/ENST00000471389.1:+:Intron|FTO/ENST00000394647.3:+:Intron|FTO/ENST00000431610.2:+
```

# Gene-based burden test : Input Data

국립보건연구원
National Institute of Health

- **Creating marker group file**

```
$ epacts make-group --vcf KOGO.anno.vcf.gz \
     --out ../04_OUTPUT/KOGO.gene.all.grp --format epacts -nonsyn
```

- **make-group:** option to create grp file
- **--vcf:** input file  (VCF format), annotated VCF
- **--out:** output file name
- **--format:** annotation format (epacts, annovar, gatk, snpeff, chaos)
- **--nonsyn:** use nonsynonymous SNPs

# Gene-based burden test : Input Data

국립보건연구원
National Institute of Health

- **Gene group data (13개 gene)**

```
$ less -NS KOGO.gene.all.grp
```

```
CAPNS2   16:55600963_G/C 16:55601192_G/A
CES1     16:55844421_G/A 16:55844927_A/C 16:55850879_G/A 16:55853446_C/A 16:55853481_G/A 16:55853498_G/C 16:55853545_C/A
CES1P1   16:55798603_A/G 16:55806276_T/C 16:55806277_A/G
CES5A    16:55880480_A/C 16:55880681_C/T 16:55880716_T/C 16:55883652_C/T 16:55883674_G/A 16:55893483_G/C 16:55897289_C/T
IRX3     16:54318172_C/A 16:54318528_A/G 16:54319850_G/A 16:54319851_C/A
IRX5     16:54967040_C/T 16:54967096_C/A
IRX6     16:55360297_G/A 16:55361290_G/A 16:55361493_C/T 16:55361523_G/C 16:55361727_A/C 16:55362645_G/A 16:55362716_G/A
LPCAT2   16:55543149_T/G 16:55559513_G/A 16:55559513_G/T 16:55562359_G/A 16:55562466_G/A 16:55579710_G/A 16:55613052_C/T
MMP2     16:55498738_G/C 16:55525759_G/A 16:55527063_T/C 16:55527073_C/T 16:55530864_G/A 16:55536687_A/G 16:55536723_G/T
RP11-21 2I21.2    16:55478526_G/A
RP11-26 L20.3     16:55316584_C/T
RP11-32 4D17.1    16:54279640_G/T 16:54279662_C/T 16:54279712_A/C
SLC6A2   16:55703498_C/T 16:55719143_G/A 16:55725894_C/G 16:55729318_G/A 16:55734106_T/C
```

# Gene-based burden test : Analysis Command

국립보건연구원 National Institute of Health

- **Running**

```
$ epacts-group --vcf KOGO.anno.vcf.gz \
    --ped KOGO.HDL.ped --groupf KOGO.gene.all.grp \
    --min-mac 5 --chr 16 --pheno HDL --cov AGE --cov SEX --test skat \
    --skat-o --unit 1 -min-callrate 0.90 --run 4 --field DS --missing NA\
    --out ../04_OUTPUT/KOGO.HDL.gene.skat
```

- **group:** option for gene-based test
- **--vcf:** input file in VCF format
- **--ped:** ped file name
- **--groupf:** grp file name
- **--out:** output file name
- **--pheno**: column name of phenotype
- **--cov:** column name of covariate in PED file
- **--test:** statistical method
- **--skat-O:** use SKAT-O (SKAT, optimal test) method

# Gene-based burden test : Analysis Result

국립보건연구원 National Institute of Health

```
$ less -NS ../04_OUTPUT/KOGO.HDL.gene.skat.epacts
```

| #CHROM | BEGIN | END | MARKER_ID | NS | FRAC_WITH_RARE | NUM_ALL_VARS | NUM_PASS_VARS | NUM_SING_VARS | PVALUE | STATRHO | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 55600963 | | 55601192 | 16:55600963-55601192_CAPNS2 | | 34497 | 0.0032177 | 2 | 2 | 0 | 0.52552 1 |
| 16 | 55844421 | | 55866957 | 16:55844421-55866957_CES1 | | 34497 | 0.045482 | 21 | 14 | 0 | 0.018027 1 |
| 16 | 55798603 | | 55806277 | 16:55798603-55806277_CES1P1 | | 34497 | 0.0007247 | 3 | 2 | 0 | 0.29357 1 |
| 16 | 55880480 | | 55907857 | 16:55880480-55907857_CES5A | | 34497 | 0.081166 | 20 | 16 | 0 | 0.50162 0 |
| 16 | 54318172 | | 54319851 | 16:54318172-54319851_IRX3 | | 34497 | 5.7976e-05 | 4 | 2 | 0 | 0.96831 NA |
| 16 | 54967040 | | 54967096 | 16:54967040-54967096_IRX5 | | 34497 | 0.00034786 | 2 | 1 | 0 | 0.086471 NA |
| 16 | 55360297 | | 55362842 | 16:55360297-55362842_IRX6 | | 34497 | 0.032756 | 9 | 7 | 0 | 0.041928 0.7 |
| 16 | 55543149 | | 55613052 | 16:55543149-55613052_LPCAT2 | | 34497 | 0.032612 | 7 | 6 | 0 | 0.65149 0 |
| 16 | 55498738 | | 55536782 | 16:55498738-55536782_MMP2 | | 34497 | 0.031829 | 8 | 6 | 0 | 0.22382 1 |
| 16 | 55478526 | | 55478526 | 16:55478526-55478526_RP11-212I21.2 | | NA | NA | 1 | 0 | 0 | NA NA |
| 16 | 55316584 | | 55316584 | 16:55316584-55316584_RP11-26L20.3 | | NA | NA | 1 | 0 | 0 | NA NA |
| 16 | 54279640 | | 54279712 | 16:54279640-54279712_RP11-324D17.1 | | 34497 | 0 | 3 | 1 | 0 | 0.63901 NA |
| 16 | 55703498 | | 55734106 | 16:55703498-55734106_SLC6A2 | | 34497 | 0.002522 | 5 | 5 | 0 | 0.094387 1 |

- **NS**: number of phenotyped samples with non-missing genotypes
- **FRAC_WITH_RARE**: fraction of individual carrying rare variants below --max-maf (default : 0.05) threshold
- **NUM_ALL_VARS**: number of all variants defining the group
- **NUM_PASS_VARS**: number of variants passing the --min-maf, --min-mac, --max-maf, --min-callrate thresholds
- **NUM_SING_VARS**: number of singletons among variants in NUM_PASS_VARS
- **PVALUE**: p-value of burden tests

# Meta analysis

## Statistical tool: METAL

Ref. https://en.wikipedia.org/wiki/Metal; https://genome.sph.umich.edu/wiki/METAL_Quick_Start

# Meta Analysis



- **통계 검정력을 높일 수 있음 (샘플 수 증가)**
- **데이터 결합 분석의 제한점*에 영향을 받지 않음**

  * 각 연구 마다 confounding variable이 다름

  * 수집 데이터 용량 , 포맷, 특성, 동의 등에 따라 공유에 제한

- **접근이 용이한 summary statistic(p-value) 등을 이용하여 분석 효율성을 증가시킴**

# Meta Analysis

국립보건연구원
National Institute of Health

- **각 input file의 필수 정보**
  - 각 연구 간에 동일에 **marker id**
  - **테스트 된 allele (effective allele) / other allele**
- **sample size를 가중치로 분석하는 경우**
  - **p-value**
  - **sample size (표본 크기가 다른 경우)**
- **표준오차(SE)를 가중치로 분석하는 경우**
  - 각 marker에 대한 **estimated effect size**
  - effect size estimate의 **standard error**

# Meta Analysis

국립보건연구원
National Institute of Health

- **METAL command (1)**

```
SCHEME STDERR
MARKER SNP
ALLELE ALT REF
EFFECT BETA
STDERR SE
PVALUE PVALUE
PROCESS DATA1.input
PROCESS DATA2.input
OUTFILE meta.out
ANALYZE HETEROGENEITY
```

- SCHEME: 분석 방법 ** weight 옵션*
- MARKER: marker ID 컬럼명
- ALLELE: effective Allele / Other Allele 컬럼명
- EFFECT: effect size 컬럼명
- STDERR: standard error 컬럼명
- PVALUE: pvalue 컬럼명
- PROCESS: intput 파일명
- OUTFILE: output 파일명
- ANALYZE: meta 분석 실행

** HETEROGENEITY test 추가 옵션*

# Meta Analysis

국립보건연구원
National Institute of Health

- **METAL command (2) –** 각 데이터 마다 변수명이 다른 경우

```
SCHEME STDERR
MARKER SNP
ALLELE ALT REF
EFFECT BETA
STDERR SE
PVALUE PVALUE
PROCESS DATA1.input
MARKER ID
ALLELE A1 A2
EFFECT effect
STDERR StErr
PVALUE P
PROCESS DATA2.input
....

OUTFILE meta.out
ANALYZE HETEROGENEITY
```

# Meta Analysis



- **Step1 : make input data**
  - marker id 통일, 필수 정보 추출
- **Step2 : metal command 작성 및 실행**

# Meta Analysis : Input Data

국립보건연구원
National Institute of Health

```
$ less -NS KOGO.HDL.single.q.linear.epacts.gz
```

| #CHROM | BEGIN | END | MARKER_ID | NS | AC | CALLRATE | MAF | PVALUE | BETA | SEBETA | TSTAT | R2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 54770105 | | 54770105 | 16:54770105_T/C_16:54770105:T:C | 34497 | 706.82 | 1 | 0.010245 | | 0.59445 | 1.0046 | 1.8869 | 0.53241 8.2174e-06 |
| 16 | 54770232 | | 54770232 | 16:54770232_C/T_16:54770232:C:T | 34497 | 25478 | 1 | 0.36929 | 0.64377 | 0.21666 | 0.46852 | 0.46244 | 6.1993e-06 |
| 16 | 54770268 | | 54770268 | 16:54770268_G/A_16:54770268:G:A | 34497 | 7010.3 | 1 | 0.10161 | 0.50391 | 0.50485 | 0.75536 | 0.66836 | 1.295e-05 |
| 16 | 54770445 | | 54770445 | 16:54770445_C/T_16:54770445:C:T | 34497 | 15816 | 1 | 0.22924 | 0.52231 | 0.30832 | 0.48189 | 0.6398 | 1.1867e-05 |
| 16 | 54770476 | | 54770476 | 16:54770476_C/T_16:54770476:C:T | 34497 | 16357 | 1 | 0.23707 | 0.46343 | 0.36211 | 0.49386 | 0.73321 | 1.5585e-05 |
| 16 | 54770477 | | 54770477 | 16:54770477_TC/T_16:54770477:TC:T | | 34497 | 16357 | 1 | 0.23707 | 0.46343 | 0.36211 | 0.49386 | 0.73321 1.5585e-05 |

```
$ less -NS BBJ.HDL-C.autosome.txt
```

| SNP | CHR | POS | REF | ALT | Frq | Rsq | BETA | SE | P | LOG10P | N | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs12922563 | 16 | 53001788 | T | C | 0.8626 | 0.751 | 0.003753 | 0.008767 | 0.6686 | 0.1748 | 70657 | | |
| rs13335763 | 16 | 53002558 | G | A | 0.1283 | 0.763 | -0.002919 | 0.008922 | 0.7436 | 0.1287 | 70657 | | |
| rs7193881 | 16 | 53003344 | C | T | 0.1241 | 0.739 | -0.004507 | 0.009211 | 0.6246 | 0.2044 | 70657 | | |
| rs72812125 | 16 | 53003599 | A | C | 0.1256 | 0.748 | -0.004086 | 0.009103 | 0.6535 | 0.1848 | 70657 | | |
| rs28883922 | 16 | 53003924 | G | A | 0.1241 | 0.738 | -0.004533 | 0.009212 | 0.6226 | 0.2058 | 70657 | | |
| rs74483905 | 16 | 53005526 | G | A | 0.1283 | 0.762 | -0.003684 | 0.008934 | 0.6801 | 0.1674 | 70657 | | |
| rs72812126 | 16 | 53006138 | G | A | 0.1252 | 0.706 | -0.004596 | 0.009381 | 0.6242 | 0.2047 | 70657 | | |

# Meta Analysis : Input Data

국립보건연구원
National Institute of Health

```
## BBJ input data
$ awk '$2 == 16 \
  {print $2":"$3"_"$4"/"$5"_"$2":"$3":"$4":"$5"\t" \
  $2"\t"$3"\t"$4"\t"$5"\t"$8"\t"$9"\t"$10"\t"$12}' \
  BBJ.HDL-C.autosome.txt > ../04_OUTPUT/BBJ.HDL.txt
$ cat TITLE.txt ../04_OUTPUT/BBJ.HDL.txt > ../04_OUTPUT/BBJ_meta_input.txt

## KBA input data
$ gunzip -c KOGO.HDL.single.q.linear.epacts.gz | tail -n+2 | \
  awk '{split($4,arr,":"); print $4"\t"$1"\t"$2"\t" \
  arr[4]"\t"arr[5]"\t"$10"\t"$11"\t"$9"\t"$5}' > ../04_OUTPUT/KBA.HDL.txt
$ cat TITLE.txt ../04_OUTPUT/KBA.HDL.txt > ../04_OUTPUT/KBA_meta_input.txt
```

# Meta Analysis : Input Data

국립보건연구원
National Institute of Health

```
$ less -NS KBA_meta_input.txt
```

| MARKER_ID | CHR | POS | REF | ALT | BETA | SE | PVALUE | | |
|---|---|---|---|---|---|---|---|---|---|
| 16:54770232_C/T_16:54770232:C:T | 16 | 54770232 | | C | T | 0.21666 | 0.46852 | 0.64377 | |
| 16:54770268_G/A_16:54770268:G:A | 16 | 54770268 | | G | A | 0.50485 | 0.75536 | 0.50391 | |
| 16:54770445_C/T_16:54770445:C:T | 16 | 54770445 | | C | T | 0.30832 | 0.48189 | 0.52231 | |
| 16:54770478_C/T_16:54770478:C:T | 16 | 54770478 | | C | T | 0.21686 | 0.46858 | 0.64351 | |
| 16:54770489_T/A_16:54770489:T:A | 16 | 54770489 | | T | A | 0.15652 | 2.2393 | 0.94428 | |
| 16:54770720_C/T_16:54770720:C:T | 16 | 54770720 | | C | T | 0.30862 | 0.93076 | 0.74021 | |
| 16:54770770_T/C_16:54770770:T:C | 16 | 54770770 | | T | C | 0.30862 | 0.93076 | 0.74021 | |
| 16:54770815_A/G_16:54770815:A:G | 16 | 54770815 | | A | G | 0.27046 | 0.926 | 0.77023 | |

```
$ less -NS BBJ_meta_input.txt
```

| MARKER_ID | CHR | POS | REF | ALT | BETA | SE | PVALUE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 16:54770232_C/T_16:54770232:C:T | 16 | 54770232 | | C | T | -0.004667 | | 0.005234 | | 0.3726 |
| 16:54770268_G/A_16:54770268:G:A | 16 | 54770268 | | G | A | -0.01252 | | 0.008248 | | 0.1291 |
| 16:54770445_C/T_16:54770445:C:T | 16 | 54770445 | | C | T | -0.01176 | | 0.005326 | | 0.02722 |
| 16:54770478_C/T_16:54770478:C:T | 16 | 54770478 | | C | T | -0.00476 | | 0.005281 | | 0.3674 |
| 16:54770489_T/A_16:54770489:T:A | 16 | 54770489 | | T | A | 0.012 | 0.01535 | 0.4342 | | |
| 16:54770720_C/T_16:54770720:C:T | 16 | 54770720 | | C | T | -0.0003057 | | 0.01326 | 0.9816 | |
| 16:54770770_T/C_16:54770770:T:C | 16 | 54770770 | | T | C | -0.0003183 | | 0.01326 | 0.9809 | |
| 16:54770815_A/G_16:54770815:A:G | 16 | 54770815 | | A | G | -0.0003277 | | 0.01326 | 0.9803 | |

# Meta Analysis : Analysis Command

국립보건연구원
National Institute of Health

- **Running**

```
$ metal
```

```
#   SEPARATOR          [WHITESPACE|COMMA|BOTH|TAB] (default = WHITESPACE)
#   COLUMNCOUNTING     [STRICT|LENIENT]            (default = 'STRICT')
#   MARKERLABEL        [LABEL]                     (default = 'MARKER')
#   ALLELELABELS       [LABEL1 LABEL2]             (default = 'ALLELE1','ALLELE2')
#   EFFECTLABEL        [LABEL|log(LABEL)]          (default = 'EFFECT')
#   FLIP
#
# Options for filtering input files ...
#   ADDFILTER          [LABEL CONDITION VALUE]     (example = ADDFILTER N > 10)
#                      (available conditions are <, >, <=, >=, =, !=, IN)
#   REMOVEFILTERS
                     =============== 중략 ===============
# Options for general analysis control ...
#   PROCESSFILE        [FILENAME]
#   OUTFILE            [PREFIX SUFFIX]             (default = 'METAANALYSIS','.TBL')
#   MAXWARNINGS        [NUMBER]                    (default = 20)
#   VERBOSE            [ON|OFF]                    (default = 'OFF')
#   LOGPVALUE          [ON|OFF]                    (default = 'OFF')
#   ANALYZE            [HETEROGENEITY]
#   CLEAR

# Options for general run control ...
#   SOURCE             [SCRIPTFILE]
#   RETURN
#   QUIT
```

# Meta Analysis : Analysis Command

국립보건연구원
National Institute of Health

- **Running**

```
SCHEME STDERR
MARKER MARKER_ID
ALLELE ALT REF
EFFECT BETA
STDERR SE
PVALUE PVALUE
PROCESS KBA_meta_input.txt
PROCESS BBJ_meta_input.txt
OUTFILE ../04_OUTPUT/KBA_BBJ_META_HDL .txt   ## .txt 앞에 공백 필수
ANALYZE HETEROGENEITY
QUIT
```

# Meta Analysis : Analysis Result

국립보건연구원
National Institute of Health

```
$ less -NS ../04_OUTPUT/KBA_BBJ_META_HDL1.txt
```

| MarkerName | Allele1 | Allele2 | Effect | StdErr | P-value | Direction | | HetISq | HetChiSq | | HetDf | HetPVal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16:54881826_C/T_16:54881826:C:T | t | c | -0.0003 | 0.0096 | 0.9781 | -- | | 0.0 | 0.233 | | 1 | 0.6296 |
| 16:54813858_A/C_16:54813858:A:C | a | c | -0.0137 | 0.0148 | 0.3545 | -- | | 0.0 | 0.307 | | 1 | 0.5798 |
| 16:54819972_C/T_16:54819972:C:T | t | c | 0.0140 | 0.0148 | 0.345 | ++ | | 0.0 | 0.021 | | 1 | 0.8844 |
| 16:54829523_T/A_16:54829523:T:A | a | t | -0.0028 | 0.0054 | 0.6011 | +- | | 0.0 | 0.097 | | 1 | 0.7552 |
| 16:54859114_T/C_16:54859114:T:C | t | c | -0.0016 | 0.0054 | 0.7647 | +- | | 28.1 | 1.391 | | 1 | 0.2382 |
| 16:54817406_G/A_16:54817406:G:A | a | g | 0.0139 | 0.0148 | 0.3486 | ++ | | 0.0 | 0.306 | | 1 | 0.5799 |
| 16:54870361_G/A_16:54870361:G:A | a | g | -0.0112 | 0.0113 | 0.3234 | +- | | 0.0 | 0.178 | | 1 | 0.6734 |
| 16:55060225_C/T_16:55060225:C:T | t | c | -0.0080 | 0.0078 | 0.3026 | -- | | 0.0 | 0.027 | | 1 | 0.8703 |

- **MarkerName**: name of marker (e.g. rsid)
- **Allele1**: Allele 1 (effective allele)
- **Allele2**: Allele 2 (other allele)
- **Direction**: direction of effect (+,-, or ? denoting positive, negative or missing effect, respectively)
- **HetChiSq:** Heterogeneity statistics
- **HetPVal:** Heterogeneity p-value

# Polygenic Risk Score Analysis

## Statistical tool: PRS-CS, Plink

https://github.com/getian107/PRScs

# PRS Analysis



- **Polygenic Risk Score (PRS)는 개인의 질병 등에 대한 유전적 위험도를 평가하는 점수**
  - $PRS_j = \sum_{i=1}^{n} \beta_i G_{ij}$ ($\beta$: effect size , $i$: SNPs 1, …, n, $j$: individual)
- **PRS는 유전 정보로 예측되는 phenotype 값임**
- **질병 고위험군 선별 및 예측 모형을 구축할 수 있음**

# PRS Analysis

국립보건연구원
National Institute of Health

| Method | # of markers | Software | Limitation |
|---|---|---|---|
| PRS using validated markers | ~ hundreds | - | Limited # of markers |
| Unadjusted PRS | ~ millions | - | No LD info. |
| Clumping(or Pruning) + P-value thresholding | ~ millions | PRSice | Not optimized |
| Bayes based analysis (Individual data) | ~ millions | BayesR | Accurate but slow |
| local LD info (Summary stat.) | ~ millions | LDpred | Less accurate (Fast and efficient) |
| local LD + Shrinkage factor (Summary stat.) | ~ millions | PRS-CS | Less accurate (Fast and efficient) |

# PRS Analysis

국립보건연구원
National Institute of Health

- **Step1 : adjusted effect size calculation (PRS-CS)**
  - Input data
    - GWAS summary statistic (BBJ HDL GWAS Result)
    - Ref panel (1KGP East asian 504 sample, HapMap Phase 3 SNPs)
    - Target genotype data (plink format file)

    *** Only use overlapped SNPs**
- **Step2 : PRS calculation (plink)**

# PRS Analysis : Input Data

국립보건연구원 National Institute of Health

- **GWAS summary statistic (필요변수 추출)**

```
$ head BBJ.HDL-C.autosome.txt
```

| SNP | CHR | POS | REF | ALT | Frq | Rsq | BETA | SE | P | LOG10P | N | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rs12922563 | 16 | 53001788 | T | C | 0.8626 | 0.751 | 0.003753 | 0.008767 | 0.6686 | 0.1748 | 70657 | | | |
| rs13335763 | 16 | 53002558 | G | A | 0.1283 | 0.763 | -0.002919 | 0.008922 | 0.7436 | 0.1287 | 70657 | | | |
| rs7193881 | 16 | 53003344 | C | T | 0.1241 | 0.739 | -0.004507 | 0.009211 | 0.6246 | 0.2044 | 70657 | | | |
| rs72812125 | 16 | 53003599 | A | C | 0.1256 | 0.748 | -0.004086 | 0.009103 | 0.6535 | 0.1848 | 70657 | | | |

```
$ awk '{print $1"\t"$5"\t"$4"\t"$8"\t"$10}' BBJ.HDL-C.autosome.txt > \
    ../04_OUTPUT/BBJ.HDL.SS.txt
$ head ../04_OUTPUT/BBJ.HDL.SS.txt
```

| SNP | ALT | REF | BETA | P | |
|---|---|---|---|---|---|
| rs12922563 | C | T | 0.003753 | 0.6686 | |
| rs13335763 | A | G | -0.002919 | 0.7436 | |
| rs7193881 | T | C | -0.004507 | 0.6246 | |
| rs72812125 | C | A | -0.004086 | 0.6535 | |
| rs28883922 | A | G | -0.004533 | 0.6226 | |

# PRS Analysis : Input Data

국립보건연구원 National Institute of Health

- **Genotype data (vcf 포맷 -> plink 포맷으로 변환)**

```
$ plink --vcf KOGO.vcf.gz --make-bed --out ./04_OUTPUT/KOGO.HDL.PRS
$ ls -ahl KOGO.HDL.PRS*
```

```
ONGui-MacBook-Pro:03_PRS ongs$ ls -ahl KOGO.HDL.PRS.*
-rw-r--r--  1 ongs  staff   196M  6 16 18:17 KOGO.HDL.PRS.bed
-rw-r--r--  1 ongs  staff   806K  6 16 18:17 KOGO.HDL.PRS.bim
-rw-r--r--  1 ongs  staff   842K  6 16 18:17 KOGO.HDL.PRS.fam
```

```
$ head KOGO.HDL.PRS.bim
```

```
16      16:54000023:G:A 0       54000023        A       G
16      16:54000190:T:C 0       54000190        C       T
16      16:54000198:C:T 0       54000198        T       C
16      16:54000220:A:G 0       54000220        G       A
16      16:54000261:A:T 0       54000261        T       A
```

# PRS Analysis : Input Data

국립보건연구원
National Institute of Health

- **Genotype data (SNP ID 변환)**

```
$ awk '{print $2":"$3":"$4":"$5"\t"$1}' BBJ.HDL-C.autosome.txt > rsID_MAT
$ head reID_MAT
```

```
CHR:POS:REF:ALT SNP
16:53001788:T:C rs12922563
16:53002558:G:A rs13335763
16:53003344:C:T rs7193881
16:53003599:A:C rs72812125
16:53003924:G:A rs28883922
```

```
$ plink --bfile KOGO.HDL.PRS --update-name rsID_MAT --make-bed \
      --out ./04_OUTPUT/KOGO.HDL.PRS.rsID
$ ls -ahl KOGO.HDL.PRS.rsID*
```

```
ONGui-MacBook-Pro:03_PRS ongs$ ls -ahl KOGO.HDL.PRS.rsID.*
-rw-r--r--  1 ongs  staff   196M  6 16 18:41 KOGO.HDL.PRS.rsID.bed
-rw-r--r--  1 ongs  staff   775K  6 16 18:41 KOGO.HDL.PRS.rsID.bim
-rw-r--r--  1 ongs  staff   842K  6 16 18:41 KOGO.HDL.PRS.rsID.fam
```

# PRS Analysis : Input Data

국립보건연구원 National Institute of Health

- **Genotype data (overlap data )**

```
$ head KOGO.HDL.PRS.rsID.bim
```

```
16      16:54000588:A:C 0       54000588        C       A
16      16:54000773:G:A 0       54000773        A       G
16      rs1344502       0       54000792        G       A
16      rs1111483       0       54000907        C       A
```

```
$ grep -E "rs" KOGO.HDL.PRS.rsID.bim | awk '{print $2}' > ext_rsID_SNP
$ plink --bfile KOGO.HDL.PRS.rsID --extract ext_rsID_SNP --make-bed \
    --out ./04_OUTPUT/KOGO.HDL.PRS.FINAL
$ head KOGO.HDL.PRS.FINAL.bim
```

```
16      rs1344502       0       54000792        G       A
16      rs1111483       0       54000907        C       A
16      rs7194907       0       54003483        C       T
16      rs8053888       0       54003805        C       T
16      rs7206456       0       54005489        A       G
```

# PRS Analysis : Analysis Command

국립보건연구원 National Institute of Health

- **Running**

```
$ python3 PRScs.py --ref_dir=ldblk_1kg_eas --chrom=16 \
     --bim_prefix=KOGO.HDL.PRS.FINAL --sst_file=BBJ.HDL.SS.txt \
     --n_gwas=136615 --out_dir=../04_OUTPUT/PRS_BETA
$ head ../04_OUTPUT/PRS_BETA_pst_eff_a1_b0.5_phiauto_chr16.txt
```

- **--ref_dir:** LD reference panel (the snpinfo file and hdf5 file)
- **--chrom:** chromosome 옵션
- **--bim_prefix:** bim file for targer genotype dataset
- **--sst_file:** effect size input file name (GWAS summary statistics file)
- **--n_gwas:** sample size of the GWAS
- **--out_dir:** output file name

| | | | | | |
|---|---|---|---|---|---|
| 16 | rs1344502 | 54000792 | G | A | 6.206899e-05 |
| 16 | rs1111483 | 54000907 | C | A | -3.421857e-05 |
| 16 | rs7194907 | 54003483 | C | T | 1.934491e-04 |
| 16 | rs8053888 | 54003805 | C | T | 2.805416e-05 |
| 16 | rs7206456 | 54005489 | A | G | -1.508700e-05 |
| 16 | rs7185783 | 54007822 | T | G | 8.683969e-05 |

# PRS Analysis : Analysis Command

국립보건연구원 National Institute of Health

- **PRS Calculation**

```
$ plink --bfile KOGO.HDL.PRS.FINAL --allow-no-sex \
      --out ../04_OUTPUT/HDL.PRS.OUT \
      --score PRS_BETA_pst_eff_a1_b0.5_phiauto_chr16.txt 2 4 6 header sum
$ head ../04_OUTPUT/HDL.PRS.OUT.profile
```

- **--bfile:** genotype (plink 포맷) input file name
- **--allow-no-sex:** 성별 정보가 없어도 분석이 가능하도록 함
- **--out:** output file name
- **--score:** effect size input file name
- **header sum:** socre file에 header가 있고, scoring을 합산해서 표기하도록 함

| FID | IID | PHENO | CNT | CNT2 | SCORESUM |
|---|---|---|---|---|---|
| ID00502 | ID00502 | -9 | 2164 | 331 | -0.0060191 |
| ID00503 | ID00503 | -9 | 2164 | 311 | -0.00395759 |
| ID00504 | ID00504 | -9 | 2164 | 246 | 0.00456846 |
| ID00505 | ID00505 | -9 | 2164 | 287 | 0.00564633 |
| ID00506 | ID00506 | -9 | 2164 | 298 | -0.0060505 |

# 감사합니다