A Brief Primer on Biobank Data Analysis

Kisung Nam

Graduate School of Data Science

Seoul National University

2024. 1. 18.

Contents

What is Biobank?

Some Definitions of "Biobank":

- **A biobank** is a type of biorepository that stores biological samples (usually human) for use in research *(Wikipedia)*
- **"인체유래물은행"**이란 인체유래물 또는 유전정보와 그에 관련된 역학정보, 임상정보 등을 수집, 보존하여 이를 직접 이용하거나 타인에게 제공하는 기관을 말한다 *(생명윤리법 제2조)*

UK Biobank and Korean Biobank

UK Biobank

- **The UK Biobank project** is a prospective cohort study with deep genetic and phenotypic data collected on approximately 500,000 individuals from across the UK, aged between 40 and 69

Korean Biobank (KoGES data)

- The Korean National Institute of Health recruited more than 210,000 participants aged 40-69 via the Korean Genome and Epidemiology Study (KoGES)

Data included in UK Biobank

Genotype Data

- Genome-wide genetic data of ∼500k participants
- ∼800k called genotypes and ∼96 million imputed genotypes

Phenotype Data

- Continuous phenotypes: Height, BMI, Blood pressure, ...
- Binary phenotypes: Disease (Diabetes, Cancer, ...)
- Questionnaires: Diet, Mental health, ...
- Images: Brain MRI, Heart MRI, ...

Data included in KoGES data

Genotype Data

- Genome-wide genetic data of ∼72k participants
- ∼467k called genotypes and ∼8 million imputed genotypes

Phenotype Data

- Continuous phenotypes: Height, BMI, Blood pressure, ...
- Binary phenotypes: Disease (Diabetes, Cancer, ...)

Using Biobank Data

What Can We Do?

- **Genome-wide association studies (GWAS)**
- Identify ancestral diversity and cryptic relatedness
- Haplotype estimation and genotype imputation
- and more...

Overview of UK Biobank Data

- Can be found at `DATA/UKBB/` in storage node

Data	Folder	Format	Readable
Imputed Genotypes	imp	bgen	No
Called Genotypes	cal	bed, bim, fam	No, Yes, Yes
Whole Exome Sequencing	WES	bed, bim, fam	No, Yes, Yes
Phenotypes	Pheno	tab	Yes
ICD - Phecode Mapping	PheCode	txt	Yes
Bulk Phenotypes	bulk	-	-
Sample Relatedness	info	dat	Yes
PC scores	PC	txt	Yes

Genotype File Formats

Various File Formats of Genotype Data

- Variant call format (.vcf)
- PLINK text (.ped, .map)
- PLINK binary (.bed, .bim, .fam)
- Oxford (.bgen)

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/KBN/NIH_DATA_KCHIP$ head 210428_1_KCHIP_72296.ped | cut -f 1-40 -d$' '
NIH2008809721 NIH2008809721 0 0 2 1 C C A C A A A A T T A A G T C T T T T G C C A G C C A G T C C C C C
NIH2008245858 NIH2008245858 0 0 1 1 C C C C A A A A T T A A T T T T T T G G C C A A C C A A T T C C C C
NIH2008881655 NIH2008881655 0 0 1 1 C C A A A A A A T T A A G T T T T T G G C C A G C C A G T C C C C C
NIH2008275629 NIH2008275629 0 0 1 1 C C C C A A A A T T A A T T T T T T G G T C G G C C G G C C C C C C
NIH2008681639 NIH2008681639 0 0 2 1 C C C C A A A A T T A A T T T T T T G G C C A G C C A G T C C C C C
NIH2008829796 NIH2008829796 0 0 2 1 C C A C A A A G A T T G A G T C T T T T G G T C G G C C A G C C C C C C
NIH2008779874 NIH2008779874 0 0 2 1 C C C C A A A A A T T A A G T T T T T G G C C A A T C A G C C G C C C
NIH2008481652 NIH2008481652 0 0 1 1 C C 0 0 A A A A A T T G A G T C T T T G G T C A G C C A A T C C C C C
NIH2008863857 NIH2008863857 0 0 1 1 C C C C A A A G A T T A A T T T T T T G G C C A G C C A G T C C C C C
NIH2008433200 NIH2008433200 0 0 2 1 C C C C A A A A A T T A A T T T T T T G G C C A G C C A G T C C C C C
```

Genotype File Formats (ctd.)

Comparison Between File Formats

Format	Readable	Dosage	File Size
VCF (.vcf)	O	O	Large
PLINK text (.ped, .map)	O	X	Large
PLINK binary (.bed, .bim, .fam)	X	X	Small
Oxford (.bgen)	X	O	Medium

Variant Call Format (.vcf)

- Each row: variants (SNPs)

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/KBN/유전정보/KCHIP_72298$ zcat CHR22_annoINFO_filINFO0.8_72K.vcf.gz | head -27 | sed 1,16d | cut -f 1-10 -d$'\t'
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NIH2007486348
22      16439593        22:16439593_G/A G       A       .       PASS    IQS=0.855729;AC=24775;AN=144596 GT:DS:GP        0/0:0.033:0.967,0.033,0
22      16440500        22:16440500_T/C T       C       .       PASS    IQS=0.844602;AC=24352;AN=144596 GT:DS:GP        0/0:0.033:0.967,0.033,0
22      16441330        22:16441330_G/C G       C       .       PASS    IQS=0.861663;AC=24957;AN=144596 GT:DS:GP        0/0:0.033:0.967,0.033,0
22      16441441        22:16441441_A/G A       G       .       PASS    IQS=0.861663;AC=24957;AN=144596 GT:DS:GP        0/0:0.033:0.967,0.033,0
22      16442005        22:16442005_A/G A       G       .       PASS    IQS=0.861663;AC=24957;AN=144596 GT:DS:GP        0/0:0.033:0.967,0.033,0
22      16444235        22:16444235_T/C T       C       .       PASS    IQS=0.863802;AC=24955;AN=144596 GT:DS:GP        0/0:0.033:0.967,0.033,0
22      16458883        22:16458883_A/G A       G       .       PASS    IQS=0.879597;AC=24938;AN=144596 GT:DS:GP        0/0:0.032:0.968,0.032,0
22      16459520        22:16459520_G/T G       T       .       PASS    IQS=0.882013;AC=24843;AN=144596 GT:DS:GP        0/0:0.028:0.972,0.028,0
22      16460679        22:16460679_C/G C       G       .       PASS    IQS=0.835569;AC=23377;AN=144596 GT:DS:GP        0/0:0.028:0.972,0.028,0
22      16462088        22:16462088_C/T C       T       .       PASS    IQS=0.843069;AC=23000;AN=144596 GT:DS:GP        0/0:0.027:0.972,0.027,0
```

PLINK text (.ped, .map)

- Each row in .ped file: samples (individuals)
- Columns in .ped file = Rows in .map file

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/KBN/NIH_DATA_KCHIP$ head 210428_1_KCHIP_72296.ped | cut -f 1-40 -d$' '
NIH2008809721 NIH2008809721 0 0 2 1 C C A C A A A A T T A A G T C T T T T G C C A G C C A G T C C C C C
NIH2008245858 NIH2008245858 0 0 1 1 C C C C A A A A T T A A T T T T T T G G C C A A C C A A T T C C C C
NIH2008881655 NIH2008881655 0 0 1 1 C C A A A A A A T T A A G T T T T T G G C C A G C C A G T C C C C C
NIH2008275629 NIH2008275629 0 0 1 1 C C C C A A A A T T A A T T T T T T G G T C G G C C G G C C C C C C
NIH2008681639 NIH2008681639 0 0 2 1 C C C C A A A A T T A A T T T T T T G G C C A G C C A G T C C C C C
NIH2008829796 NIH2008829796 0 0 2 1 C C A C A A A G A T T G A G T C T T T T G G T C G G C C A G C C C C C C
NIH2008779874 NIH2008779874 0 0 2 1 C C C C C A A A A T T A A G T T T T T T G G C C A A T C A G C C G C C C
NIH2008481652 NIH2008481652 0 0 1 1 C C 0 0 A A A A T T G A G T C T T T T G G T C A G C C A A T C C C C C
NIH2008863857 NIH2008863857 0 0 1 1 C C C C A A A G A T T A A T T T T T T T G G C C A G C C A G T C C C C C
NIH2008433200 NIH2008433200 0 0 2 1 C C C C A A A A T T A A T T T T T T T G G C C A G C C A G T C C C C C
```

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/KBN/유전정보/Genotype$ head 201215_1_1KG_8840.map
1       rs143225517     0       751756
1       rs3094315       0       752566
1       rs3131972       0       752721
1       rs3131971       0       752894
1       rs61770173      0       753405
1       rs2073814       0       753474
1       rs2073813       0       753541
1       rs3131969       0       754182
1       rs3131968       0       754192
1       rs3131967       0       754334
```

PLINK Binary File (.bed, .bim, .fam)

.fam file

- Sample information file accompanying a .bed file
- Columns: Fam ID, Ind ID, Pat ID, Mat ID, Sex, Phenotype

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/KBN/NIH_DATA_KCHIP/plink$ head 210428_1_KCHIP_72296.fam
NIH2008809721 NIH2008809721 0 0 2 1
NIH2008245858 NIH2008245858 0 0 1 1
NIH2008881655 NIH2008881655 0 0 1 1
NIH2008275629 NIH2008275629 0 0 1 1
NIH2008681639 NIH2008681639 0 0 2 1
NIH2008829796 NIH2008829796 0 0 2 1
NIH2008779874 NIH2008779874 0 0 2 1
NIH2008481652 NIH2008481652 0 0 1 1
NIH2008863857 NIH2008863857 0 0 1 1
NIH2008433200 NIH2008433200 0 0 2 1
```

PLINK Binary File (.bed, .bim, .fam) (ctd.)

.bim file

- Variant information file accompanying a .bed file
- Columns: Chr, rsid, Position in morgans, Position, Allele 1, Allele 2

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/KBN/NIH_DATA_KCHIP/plink$ head 210428_1_KCHIP_72296.bim
1       AX-37361829     0       761252  T       C
1       AX-83277131     0       762485  A       C
1       AX-64175821     0       792643  G       A
1       AX-32416607     0       835499  G       A
1       AX-32429231     0       839461  C       T
1       AX-32455597     0       848456  G       A
1       AX-32460873     0       850371  G       T
1       AX-32461903     0       850780  C       T
1       AX-113228954    0       852820  C       T
1       AX-32469373     0       852964  T       G
```

PLINK Binary File (.bed, .bim, .fam) (ctd.)

.bed file

- Representation of genotype calls at biallelic variants
- Must be accompanied by .bim and .fam files.

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/KBN/NIH_DATA_KCHIP/plink$ xxd -b -l 100 210428_1_KCHIP_72296.bed
00000000: 01101100 00011011 00000001 11111111 11111111 11111111  l.....
00000006: 11111111 11111111 11111110 11111111 11111110 10111111  ......
0000000c: 11111111 11111111 11111111 11111110 11111111 11111111  ......
00000012: 11111111 11111011 11111111 11111111 11111111 11111111  ......
00000018: 11111011 11111111 11111111 11111111 11111111 11111111  ......
0000001e: 11111111 11111111 11111111 11111111 11111111 11111111  ......
00000024: 11111111 11111111 11111111 11111111 11111111 11111111  ......
0000002a: 11111111 11111111 11111111 11111111 11111111 11111111  ......
00000030: 11111111 11111111 10111111 11111110 11111111 11111111  ......
00000036: 11111111 11101111 11111111 11111111 11111111 11001111  ......
0000003c: 11111111 11111111 11111111 11111111 10111011 11111111  ......
00000042: 11111110 11111111 11111011 11101111 11111111 11111111  ......
00000048: 11111111 11111111 11111111 11111111 11101111 11111111  ......
0000004e: 11111111 11111111 11111111 11111111 11111111 11111111  ......
00000054: 11111111 11111111 11111111 11110111 11111111 11111111  ......
0000005a: 11111110 11111111 11111111 11111111 11111111 11111111  ......
00000060: 11111111 11111101 11110111 11110111                    ....
```

Oxford File (.bgen)

.bgen file

- A compressed binary format for typed and *imputed* genotype data
- Compatible for command-line, R and Python

Reading Imputed Genotype Data (.bgen) File

- Using `seqminer` package in R
- Using `bgen_reader` package in Python

Oxford File (.bgen) (ctd.)

Reading a .bgen file (Python)

- Variants

```
>>> bgen = read_bgen("/Users/nam/ukb_imp_chrXY_v3.bgen", verbose=False)
>>> variants = bgen["variants"]
>>> v = variants.loc[0:10].compute()
>>> print(v)
             id         rsid chrom    pos  nalleles allele_ids   vaddr
0   X:60014_T_C  rs370048753  PAR1  60014         2        T,C      72
1   X:60014_T_G  rs370048753    XY  60014         2        T,G   10411
2   X:60017_C_T  X:60017_C_T    XY  60017         2        C,T   31991
3   X:60060_G_C  rs148832940    XY  60060         2        G,C   37474
4   X:60072_G_C  rs116895855    XY  60072         2        G,C   67459
5   X:60072_G_T  rs116895855    XY  60072         2        G,T  170647
6   X:60112_G_C  rs111065979    XY  60112         2        G,C  271868
7   X:60146_G_C  rs138058540    XY  60146         2        G,C  402640
8   X:60153_C_G  rs111264342    XY  60153         2        C,G  429576
9   X:60181_G_C  rs189980076    XY  60181         2        G,C  551856
10  X:60184_A_G  rs183336868    XY  60184         2        A,G  564127
```

- (Imputed) Genotypes of sample for id $= 0$

```
>>> genotype = bgen["genotype"]
>>> g = genotype[0].compute()
>>> print(g)
{'probs': array([[1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       ...,
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.]]), 'phased': False, 'ploidy': array([2, 2, 2, ..., 2, 2, 2]),
'missing': array([False, False, False, ..., False, False, False])}
>>> print(len(genotype))
45906
>>> print(len(g['probs']))
486443
>>> m = np.mean(g['probs'], axis=0)
>>> print(m)
[9.99214176e-01 7.80028114e-04 5.79637870e-06]
```

Phenotypes Data

- Phenotype information (∼7,500 columns) for ∼500k samples
- Can find code in UK Biobank showcase

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/DATA/UKBB/Pheno$ head ukb42597.tab | cut -f 1-10 -d$'\t'
f.eid   f.46.0.0        f.46.1.0        f.46.2.0        f.46.3.0        f.47.0.0        f.47.1.0        f.47.2.0        f.47.3.0        f.48.0.0
1000019 16      NA      NA      NA      20      NA      NA      NA      80
1000022 30      NA      NA      NA      40      NA      NA      NA      97
1000035 48      NA      NA      NA      48      NA      NA      NA      85
1000046 46      NA      NA      NA      46      NA      NA      NA      95
1000054 52      NA      NA      NA      46      NA      NA      NA      98
1000063 52      NA      NA      NA      46      NA      NA      NA      99
1000078 24      NA      NA      NA      32      NA      NA      NA      100
1000081 28      NA      NA      NA      28      NA      NA      NA      128
1000090 30      NA      NA      NA      33      NA      NA      NA      98
```

Phenotypes Data

- f.46.0.0 means "Hand grip strength (left) Instance 0"

Data-Field 46

Description: Hand grip strength (left)

Category: Hand grip strength - Physical measures - UK Biobank Assessment Centre

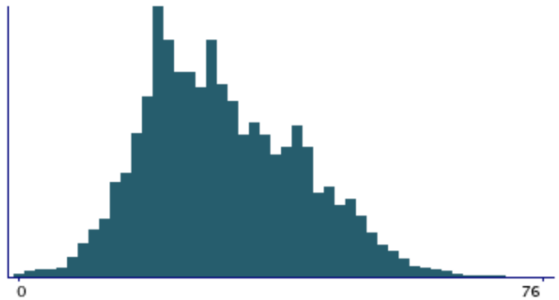
Participants	499,267	Value Type	Integer, Kg	Sexed	Both sexes	Debut	Jan 2012
Item count	569,378	Item Type	Data	Instances	Defined (4)	Version	Mar 2020
Stability	Complete	Strata	Primary	Array	No		

Data | 4 Instances | Notes | 5 Categories | 2 Related Data-Fields | 0 Tabulations | 2 Resources

Instance 0 : Initial assessment visit (2006-2010) at which participants were recruited and consent given

499,095 participants, 499,095 items

Maximum	89
Decile 9	46
Decile 8	40
Decile 7	36
Decile 6	32
Median	28
Decile 4	25
Decile 3	22
Decile 2	20
Decile 1	16
Minimum	0



- There are 86 distinct values.
- Mean = 29.5462
- Std.dev = 11.3273
- 36 items above graph maximum of 76

Principal Component (PC) Data

PC Score of Each Individual

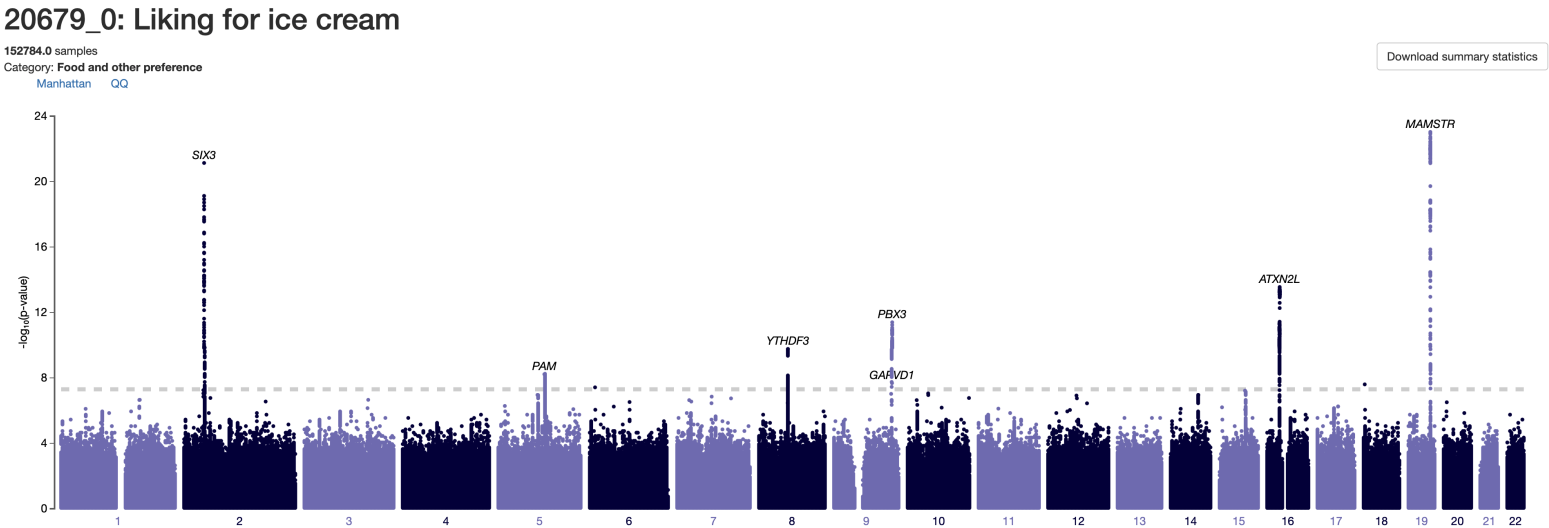
- PC scores are required for GWAS

```
(base) kisungnam@login0:/media/leelabsg-storage0/DATA/UKBB/PC$ head PEDMASTER_UNRELATED_WhiteBritish_20180612_v2_MAPPED.txt
FID     IID     Sex     birthYear       PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10    X153    X
1265060 1265060 1       1956    -16.8861        3.38776 -0.496867       0.617104        -3.90062        -2.19504        1.94369 -
4279687 4279687 1       1938    -14.1724        4.57744 -4.61973        2.05025 2.0004  -2.85618        2.4178  -2.58268        2
4608614 4608614 1       1944    -10.5828        1.87179 -2.91263        -1.04566        -5.07404        0.0109573       -0.205352
2087882 2087882 1       1949    -12.5769        4.26968 -3.1915 2.89076 -7.3564 2.36535 -1.49095        2.14493 3.6423  3.03614 0
1526004 1526004 1       1941    -10.8651        2.52371 -2.79921        -2.98971        -5.64579        2.53417 0.56558 0.422705
2530063 2530063 1       1943    -15.3836        4.45268 -2.205  1.19426 -6.6835 0.30293 0.225121        -0.69315        3.78101 1
1347805 1347805 2       1960    -12.9173        3.34111 -1.74563        -0.10012        -8.09168        1.13279 -2.03264        -
3278616 3278616 2       1956    -11.9514        4.58447 0.618089        4.84009 8.15947 0.594526        -1.3289 0.75907 -3.17168
2438170 2438170 2       1942    -13.2883        3.21761 -1.30171        1.36956 4.88423 0.542946        0.730243        -2.53709
```

What is GWAS?

GWAS

- An observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait
- Typically focus on associations between SNPs and traits like major human diseases, but can be applied to any other genetic variants



Quantitative Traits vs. Binary Traits

Quantitative Traits

- Phenotypes that depends on the cumulative actions of many genes and the environment which produce a continuous distribution
- Example: height, weight, blood pressure, etc.
- Use linear regression - for each individual i and genotype p

$$y_i = \alpha_0 + \sum_j \alpha_j X_{ij} + G_{ip}\beta_p + \epsilon$$
$$= \alpha_0 + \alpha' X_i + G_{ip}\beta_p + \epsilon$$

y: outcome (phenotype)
X: covariates (age, gender, PC)
G: genotype values (minor allele count)

Quantitative Traits vs. Binary Traits (ctd.)

Binary Traits

- Phenotypes that there are only two possible values
- Example: affected/unaffected
- Use logistic regression - for each individual i and genotype p

$$\text{logit}(\mathbb{P}(y_i = 1)) = \alpha_0 + \alpha' X_i + G_{ip}\beta_p$$

y: outcome (phenotype)
X: covariates (age, gender, PC)
G: genotype values (minor allele count)

Goal of GWAS

Goal of GWAS

- To find the associations between SNPs and traits
- For which genotype p, $\beta_p \neq 0$?

Meaning of β_p

- For linear model, β_p means the effect size of G_p
- For logistic model, β_p means the log odds ratio between $G_i = 0$ and $G_i = 1$

What is PLINK?

PLINK

- An open-source whole-genome analysis toolset
- Can handle .bim/.bed/.fam files and .bgen files
- More information
  - PLINK 1.07: https://zzz.bwh.harvard.edu/plink/index.shtml
  - PLINK 1.9: https://www.cog-genomics.org/plink2

GWAS for Standing Height - Overview

Model

- For each individual i and genotype p,

$$y_i = \alpha_0 + \sum_{j=1}^{10} \alpha_j \text{PC}_j + \alpha_{11}\text{Age}_i + \alpha_{12}\text{Gender}_i + G_{ip}\beta_p + \epsilon$$

  $\Rightarrow$ For each p, calculate α_j's and β_p

- Repeat this process for all available p's
- Parameter of interest: β_p

  $\Rightarrow$ Find p such that p-value for $\beta_p < 5 \times 10^{-8}$ (with correction)

GWAS for Standing Height - Data Preprocessing

Step 1. Build a Phenotype (.fam) File

- Using the given .fam file, build a phenotype file
- Be careful not to change the order of individuals

```
kisungnam@login:/home/lee7801/DATA/UKBB/cal$ head ukb45227_cal_chr20_v2_s488264.fam
3267751 3267751 0 0 1 Batch_b001
4085874 4085874 0 0 2 Batch_b001
1488844 1488844 0 0 2 Batch_b001
2299675 2299675 0 0 2 Batch_b001
4189123 4189123 0 0 2 Batch_b001
3600970 3600970 0 0 1 Batch_b001
1111417 1111417 0 0 2 Batch_b001
1414057 1414057 0 0 2 Batch_b001
2503961 2503961 0 0 1 Batch_b001
1802736 1802736 0 0 2 Batch_b001
```

```
kisungnam@login:~/plink$ head ukb_cal_chr22_v2.fam
3267751 3267751 0 0 1 177
4085874 4085874 0 0 2 166
1488844 1488844 0 0 2 158
2299675 2299675 0 0 2 158
4189123 4189123 0 0 2 165.5
3600970 3600970 0 0 1 186
1111417 1111417 0 0 2 162.5
1414057 1414057 0 0 2 166
2503961 2503961 0 0 1 167
1802736 1802736 0 0 2 156.5
```

GWAS for Standing Height - Data Preprocessing (ctd.)

Step 1. Build a Phenotype (.fam) File

- Phenotypes can be found in the .tab file
  - Find the phenotype code in UK Biobank showcase (height = 50)

```
(base) leelabsg@cpu03:/media/leelabsg_storage01/DATA/UKBB/Pheno$ head ukb42597.tab | cut -f 1-10 -d$'\t'
f.eid   f.46.0.0        f.46.1.0        f.46.2.0        f.46.3.0        f.47.0.0        f.47.1.0        f.47.2.0        f.47.3.0        f.48.0.0
1000019 16      NA      NA      NA      20      NA      NA      NA      80
1000022 30      NA      NA      NA      40      NA      NA      NA      97
1000035 48      NA      NA      NA      48      NA      NA      NA      85
1000046 46      NA      NA      NA      46      NA      NA      NA      95
1000054 52      NA      NA      NA      46      NA      NA      NA      98
1000063 52      NA      NA      NA      46      NA      NA      NA      99
1000078 24      NA      NA      NA      32      NA      NA      NA      100
1000081 28      NA      NA      NA      28      NA      NA      NA      128
1000090 30      NA      NA      NA      33      NA      NA      NA      98
```

- Some useful functions and commands
  - `left_join` or `merge` functions in R (or `vlookup` in Excel)
  - `cat` *infile* | `awk` '{`print` \$1, \$3, \$4}' > *outfile*

GWAS for Standing Height - Data Preprocessing (ctd.)

Step 2. Build a Covariate File

- Use "Unrelated, White British" sample to get PCs of unrelated individuals

```
(base) kisungnam@login0:/media/leelabsg-storage0/DATA/UKBB/PC$ ll
total 816600
drwxrwx---  2 root      leelabsg      4096 Jul 15 13:01 ./
drwxrwx--- 16 root      leelabsg      4096 Jan 18  2022 ../
-rw-r--r--  1 kisungnam user     182542844 Apr 14  2022 PEDMASTER_ALL_20180514_v1_MAPPED.txt
-rwxrwx---  1 root      leelabsg 181009974 Sep  8  2021 PEDMASTER_ALL_20180514_v1.txt*
-rwxrwx---  1 root      leelabsg 154154268 Sep  8  2021 PEDMASTER_UNRELATED_ALL_20180514_v1.txt*
-rwxrwx---  1 root      leelabsg 155095466 Sep  8  2021 PEDMASTER_UNRELATED_ALL_20180514_v2.txt*
-rwxrwx---  1 root      leelabsg  37911987 Sep  8  2021 PEDMASTER_UNRELATED_WhiteBritish_20180612_v1.txt*
-rw-r--r--  1 leelabsg  user      41620035 Jul 15 13:01 PEDMASTER_UNRELATED_WhiteBritish_20180612_v2_MAPPED.txt
-rwxrwx---  1 root      leelabsg  38871039 Sep  8  2021 PEDMASTER_UNRELATED_WhiteBritish_20180612_v2.txt*
-rwxrwx---  1 root      leelabsg  44965164 Sep  8  2021 PEDMASTER_WhiteBritish_20180612_v1.txt*
-rwxrwx---  1 root      leelabsg      2142 Sep  8  2021 .Rhistory*
```

GWAS for Standing Height - Data Preprocessing (ctd.)

Step 2. Build a Covariate File

- Useful if the file is ordered differently with the .fam file

```
→  cal plink --bfile ukb_cal_chr22_v2 --covar ourdata.txt --write-covar
PLINK v1.90p 64-bit (16 Jun 2020)             www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plink.log.
Options in effect:
  --bfile ukb_cal_chr22_v2
  --covar ourdata.txt
  --write-covar

16384 MB RAM detected; reserving 8192 MB for main workspace.
12968 variants loaded from .bim file.
488377 people (223467 males, 264797 females, 113 ambiguous) loaded from .fam.
Ambiguous sex IDs written to plink.nosex .
486817 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
--covar: 17 covariates loaded.
24647 people had missing value(s), and 144703 people were not seen in the covariate file.
Before main variant filters, 488377 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.972879.
12968 variants and 488377 people pass filters and QC.
Phenotype data is quantitative.
Covariates written to plink.cov .
→  cal head plink.cov
FID IID Sex birthYear PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 X153 X193 X365 X411 Age
3267751 3267751 1 1967 -12.1725 5.39163 -1.28103 0.841765 -5.26521 -1.78657 3.10992 -2.63085 2.39288 0.307537 0 0 0 0 51
4085874 4085874 2 1962 -13.0245 6.41514 -0.183365 2.92761 -5.88964 0.940534 1.14106 -1.98213 -2.70226 2.50775 0 0 0 0 56
1488844 1488844 2 1957 -11.4712 3.48383 -1.15458 3.08383 7.6516 -0.913399 -1.54879 1.47893 -1.20895 1.0649 0 0 0 0 61
2299675 2299675 2 1943 -12.1327 4.02976 -0.98808 0.750294 -2.36431 0.431658 -0.534071 -0.654367 -6.59593 -1.53356 0 0 0 0 75
4189123 4189123 2 1952 -12.2171 3.50821 -1.62599 -1.2268 -5.3458 3.81679 0.579155 -1.16958 1.01855 -0.795227 0 0 -9 0 66
3600970 3600970 1 1959 -11.5847 5.41131 0.785815 -0.940348 -6.08543 1.46541 2.6986 0.0793596 -2.85344 3.51401 0 0 0 0 59
1111417 1111417 2 1967 -11.2134 3.8501 -2.73488 0.286941 1.47179 1.1056 2.08845 0.930548 -15.8034 -0.36375 0 0 0 0 51
1414057 1414057 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9
2503961 2503961 1 1956 -12.8808 3.36515 0.233168 -3.95787 1.98784 -0.909523 -2.50145 -1.04002 -8.80647 -1.624 0 0 0 0 62
```

GWAS for Standing Height - Running PLINK

Running PLINK

- Enter a command by specifying each file and parameter
- Example command

```
./plink
--bed /home/lee7801/DATA/UKBB/cal/ukb_cal_chr2_v2.bed
--bim /home/lee7801/DATA/UKBB/cal/ukb_snp_chr2_v2.bim
--fam height_pheno.fam
--linear
--covar covar.cov
--out chr2
```

GWAS for Standing Height - Running PLINK (ctd.)

Outputs

- .assoc.linear: main result (p-values of each SNP)
- .log: log
- .nosex: IDs with ambiguous sex

GWAS for Standing Height - Running PLINK (ctd.)

.assoc.linear

```
kisungnam@login:~/plink$ head -27 chr2.assoc.linear
 CHR             SNP         BP   A1       TEST    NMISS       BETA         STAT            P
   2      rs10172629      11944    T        ADD   342681   -0.07482       -1.071       0.2844
   2      rs10172629      11944    T        Sex   342681      -13.3       -613.6            0
   2      rs10172629      11944    T        PC1   342681    0.02394        3.389    0.0007006
   2      rs10172629      11944    T        PC2   342681   -0.00733      -0.9983       0.3181
   2      rs10172629      11944    T        PC3   342681    0.01767        2.493      0.01267
   2      rs10172629      11944    T        PC4   342681   -0.02114       -3.999    6.368e-05
   2      rs10172629      11944    T        PC5   342681   -0.08749       -37.81            0
   2      rs10172629      11944    T        PC6   342681    0.01756        2.598     0.009376
   2      rs10172629      11944    T        PC7   342681   -0.04646       -7.671    1.718e-14
   2      rs10172629      11944    T        PC8   342681    0.06326        10.53    6.559e-26
   2      rs10172629      11944    T        PC9   342681  -0.000604      -0.2491       0.8033
   2      rs10172629      11944    T       PC10   342681   -0.05583       -10.62    2.546e-26
   2      rs10172629      11944    T        Age   342681    -0.1596         -118            0
   2       rs7595668      16937    A        ADD   335808   -0.02245      -0.4165       0.6771
   2       rs7595668      16937    A        Sex   335808      -13.3       -607.6            0
   2       rs7595668      16937    A        PC1   335808     0.0239        3.349    0.0008097
   2       rs7595668      16937    A        PC2   335808  -0.006317      -0.8519       0.3943
   2       rs7595668      16937    A        PC3   335808     0.0173        2.416      0.01568
   2       rs7595668      16937    A        PC4   335808   -0.01938       -3.63    0.0002833
   2       rs7595668      16937    A        PC5   335808   -0.08857       -37.89            0
   2       rs7595668      16937    A        PC6   335808    0.01747        2.558      0.01052
   2       rs7595668      16937    A        PC7   335808   -0.04662       -7.624    2.474e-14
   2       rs7595668      16937    A        PC8   335808    0.06389        10.53    6.503e-26
   2       rs7595668      16937    A        PC9   335808 -0.0004022      -0.1643       0.8695
   2       rs7595668      16937    A       PC10   335808   -0.05544       -10.44    1.718e-25
   2       rs7595668      16937    A        Age   335808    -0.1594       -116.8            0
```

GWAS for Standing Height - Running PLINK (ctd.)

.log and .nosex

```
kisungnam@login:~/plink$ cat chr2.log
PLINK v1.90b6.18 64-bit (16 Jun 2020)
Options in effect:
  --bed /home/lee7801/DATA/UKBB/cal/ukb_cal_chr2_v2.bed
  --bim /home/lee7801/DATA/UKBB/cal/ukb_snp_chr2_v2.bim
  --covar covar.cov
  --fam ukb_cal_chr22_v2.fam
  --linear
  --out chr2

Hostname: login
Working directory: /home/kisungnam/plink
Start time: Mon Aug  3 19:06:21 2020

Random number seed: 1596449181
514643 MB RAM detected; reserving 257321 MB for main workspace.
61966 variants loaded from .bim file.
488377 people (223467 males, 264797 females, 113 ambiguous) loaded from .fam.
Ambiguous sex IDs written to chr2.nosex .
486817 phenotype values loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
--covar: 12 covariates loaded.
144703 people had missing value(s).
Before main variant filters, 488377 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.967432.
61966 variants and 488377 people pass filters and QC.
Phenotype data is quantitative.
Writing linear model association results to chr2.assoc.linear ... done.

End time: Tue Aug  4 10:03:30 2020
```

```
kisungnam@login:~/plink$ head chr2.nosex
-1      -1
-2      -2
-3      -3
-4      -4
-5      -5
-6      -6
-7      -7
-8      -8
-9      -9
-10     -10
```

GWAS for Standing Height - Results

Organization of Results

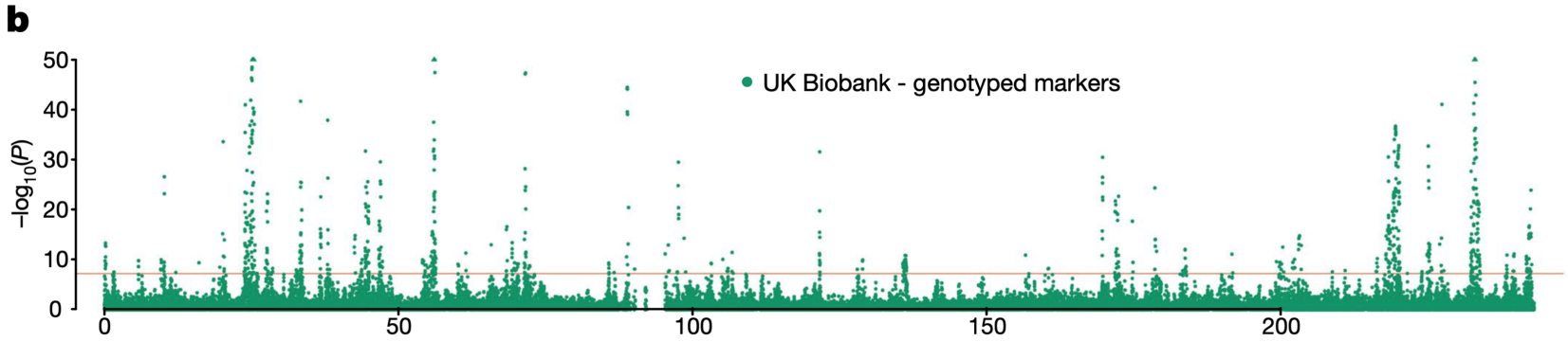
- Read .assoc.linear file in R
- Select rows where the value of TEST column is 'ADD'
- Remove rows where the p-value is 'NA'
- Draw Manhattan Plot and QQ Plot using qqman library in R

```
> library(qqman)
> chr2 <- read.csv("./Downloads/cal/plink/chr2.assoc.linear", header=T, sep="")
> chr2_1 <- chr2[chr2$TEST == 'ADD',]
> chr2_2 <- chr2_1[is.na(chr2_1$P) == FALSE,]
> manhattan(chr2_2, main = "Manhattan Plot", ylim = c(0, 40), cex = 0.8, cex.axis = 0.9, col = "skyblue")
> qq(chr2_2$P, ylim = c(0, 50), cex = 0.8, col = "skyblue")
```

GWAS for Standing Height - Results (ctd.)

Manhattan Plot

- We want to replicate the results of previous study



Results (p-values) of association tests between height and genotypes for chromosome 2

Bycroft et al. The UK Biobank resource with deep phenotyping and genomic data, *Nature* (2018)

GWAS for Standing Height - Results (ctd.)

Our Results - Manhattan Plot

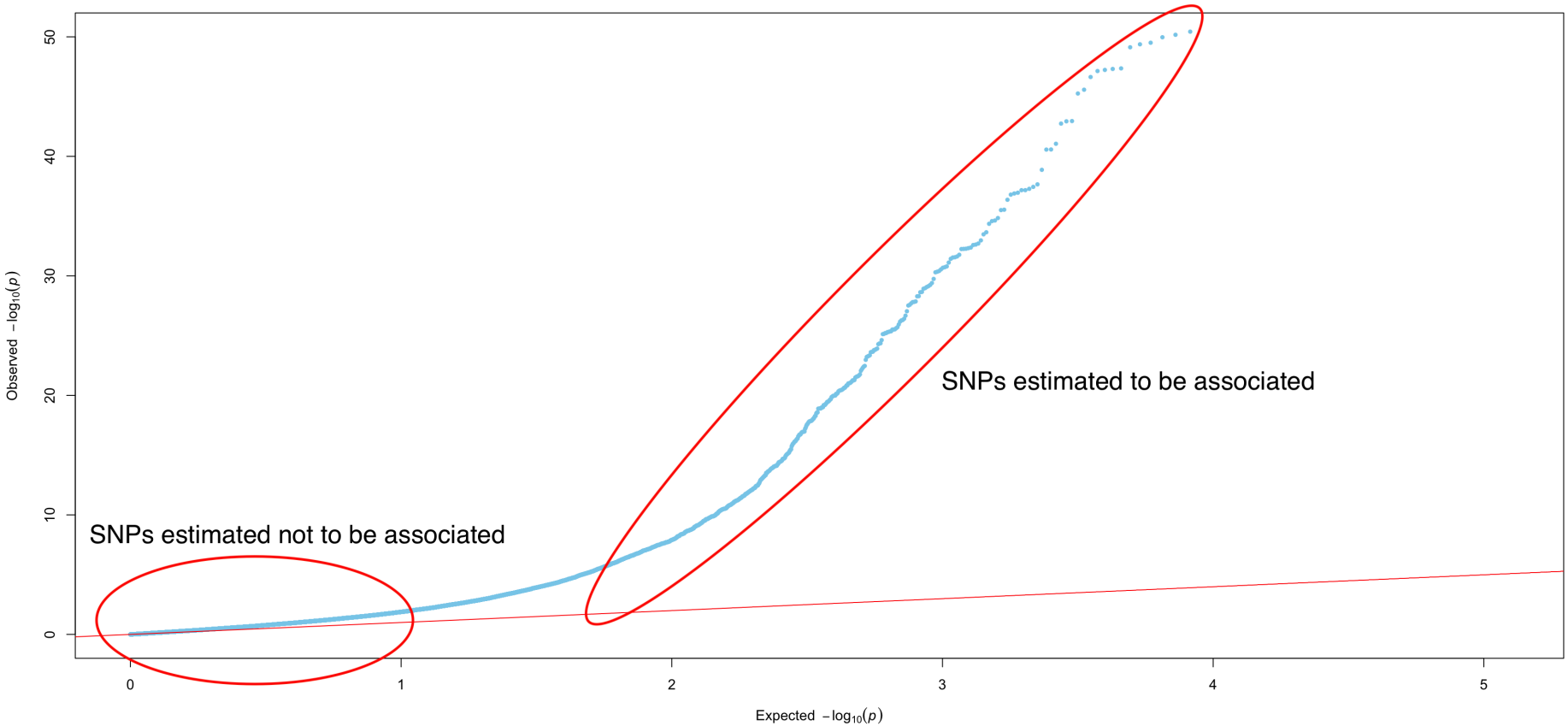
Results (p-values) of association tests between height and genotypes for chromosome 2

(Top: *Bycroft et al*, Bottom: *Replication*)

GWAS for Standing Height - Results (ctd.)

Our Results - QQ Plot

- Under the null, p-value of each SNP $\sim$ Unif(0, 1)



GWAS for Type 2 Diabetes

Model

- We can assume that

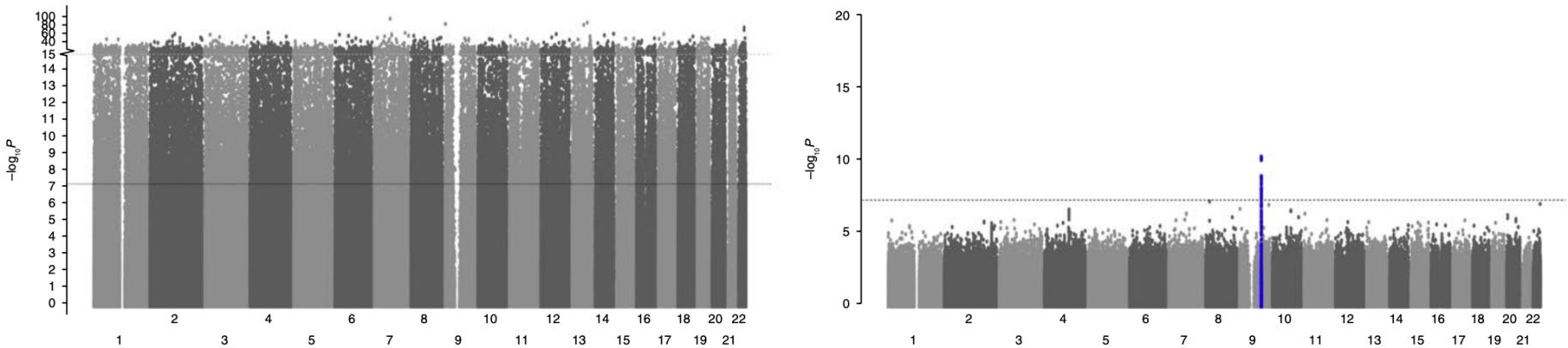
$$\text{logit}(\mathbb{P}(y_i = 1)) = \alpha_0 + \alpha' X_i + G_{ip}\beta_p$$

  where y is affected/unaffected by type 2 diabetes, G is genotype

- We assume the same covariates as in the previous example

Limitations of GWAS Using PLINK

- PLINK uses the Wald test which is computationally expensive
- Assumes asymptotic normality of test statistics which can be violated when case-control ratio is extremely skewed
- Cannot account for sample relatedness



Zhou et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies, *Nature Genetics* (2018)