Genome-wide Association Studies

Practical 2: Do the GWAS

Data Use Agreement

- To maximize your learning experience, we will be working with genuine human genetic data
- Access to this data requires agreement to the following in to comply with human genetic data ethics regulations
- Please email pctgadmin@imb.uq.edu.au to confirm that you agree with the following:
  - “I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

Data

- Data for this practical is found in the directory:
  - `/data/module1/6_unrelGWAS/`
- Three files:
  - `gwasQC.bed` → binary file containing all genotypes
  - `gwasQC.bim` → information about SNP markers
  - `gwasQC.fam` → information about individuals
- Heaps of phenotypes → Choose your own adventure!
  - Fasting glucose, fasting insulin, ferritin, height, neuroticism, sleep duration, smoking (pack years), systolic blood pressure, waist-to-hip ratio
- Covariate file: age, sex, PC 1-5 (`covariates.cov`)
- Pre-adjusted phenotypes from this morning

GWAS

- Command: `--assoc`

```
plink --bfile /data/module1/gwas/part2/gwas --assoc --pheno <file>
```

```
[allan@analysis1 ~]$ plink --bfile /data/module1/gwas/part2/gwas --assoc --pheno
 /data/module1/gwas/part2/Fasting_Insulin_QC.phen
PLINK v1.90b6.26 64-bit (2 Apr 2022)           www.cog-genomics.org/plink/1.9/
(C) 2005-2022 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plink.log.
Options in effect:
  --assoc
  --bfile /data/module1/gwas/part2/gwas
  --pheno /data/module1/gwas/part2/Fasting_Insulin_QC.phen

64141 MB RAM detected; reserving 32070 MB for main workspace.
277719 variants loaded from .bim file.
11780 people (5346 males, 6434 females) loaded from .fam.
11770 phenotype values present after --pheno.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 11780 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.995966.
277719 variants and 11780 people pass filters and QC.
Phenotype data is quantitative.
Writing QT --assoc report to plink.qassoc ... done.
```

Manhattan & QQ plots

- Use R

```
library(qqman)
d = read.table("plink.qassoc", head=T)
manhattan(d)
qq(d$P)
```

- Do your plots look good?
- Any evidence for inflation? Calculate the genomic inflation factor:

```
qchisq(1-median(d$P),1)/qchisq(0.5,1)
```

Generate PCs

- **Takes a long time to run! Use the pre-generated PCs in the covariate file**
- Command: `--pca <n>` Calculate the first n PCs

Add covariates

- Command: `--linear --covar <file>`
- **SLOWER!**
- **A LOT SLOWER IF YOU INCLUDE PCS TOO!**

- Alternative: regress the phenotype against the covariates in R and create a new phenotype file with the residuals
- Results in some power loss
- Use your pre-adjust phenotype file from this morning

Summary – Quantitative trait

- Run two or three GWAS:
  - One without adjusting for covariates
  - One adjusting for covariates and PCs
  - One using pre-adjusted phenotypes from this morning

- Compare the GWAS results – Manhattan plots, qq-plots, genomic inflation factors, time taken to run

Binary Phenotype

- Command: --assoc
- Command: --logistic --covar <file>

What is logistic regression? I have no idea where to start…

youTube: Stat Quest logistic regression in R

Set up a small example in R

1. Extract SNP from .bed file using e.g.

```
plink --bfile gwasQC --recode A --snps rs12562034 --out SNP1
```

2. Read into R, use logistic regression & compare to plink results, e.g.

```
glm(bmiBinary ~ snp,family=binomial(link='logit'),data=data)
```

Summary – binary trait

- Run two or three GWAS:
  - One using `--assoc` without adjusting for covariates
  - One using `--logistic` without adjusting for covariates
  - One using `--logistic` fitting covariates

- Extract a single SNP from the `.bed` file and read into R. Use `glm()` to understand the PLINK output.
- Compare the GWAS results – Manhattan plots, qq-plots, genomic inflation factors, time taken to run

STOP

Comparison, quantitative trait

```
(1) plink --bfile gwasQC --assoc --pheno BMI.phen --out raw
```

seconds to run

```
(2) plink --bfile gwasQC --linear --covar covariateFiltered.cov --
pheno BMI.phen –out covariates
```

Many minutes to run... I gave up.

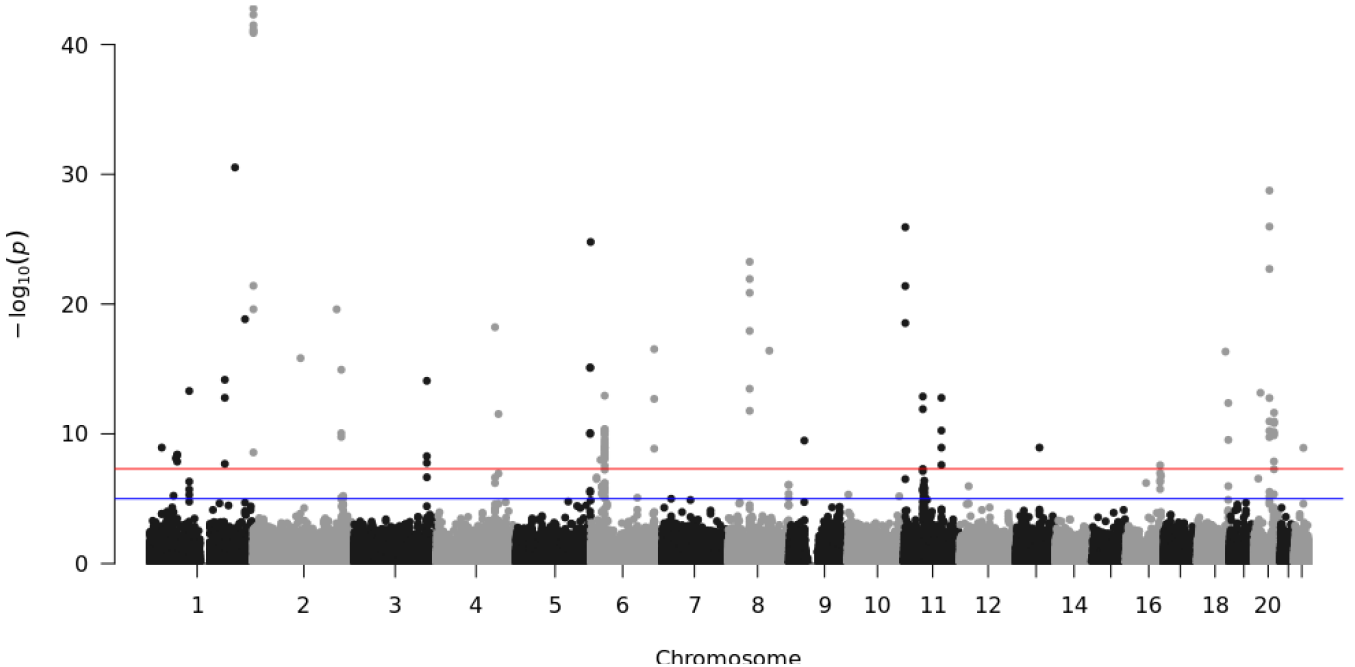
```
(3) plink --bfile gwasQC --assoc --pheno bmiStd.phen --out raw
```

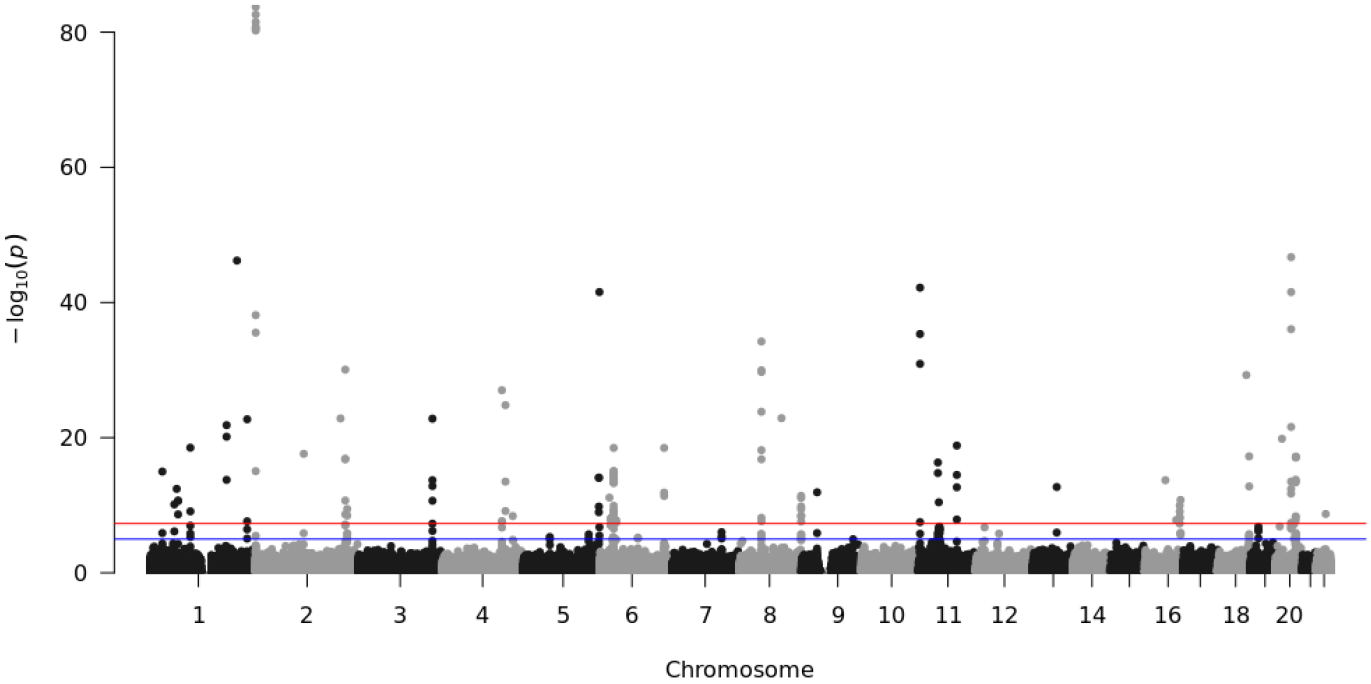
seconds to run

Output, quantitative trait

```
delta2:~/60days/UQWS_2023/5_unrelGWAS$ head raw.qassoc
 CHR          SNP         BP    NMISS       BETA         SE         R2         T            P
   1   rs12562034     768448    11683     0.2037     0.1314  0.0002056      1.55       0.1212
   1    rs4040617     779322    11667    0.02397     0.1193  3.463e-06     0.201       0.8407
   1    rs4970383     838555    11687    0.03148    0.09247  9.915e-06    0.3404       0.7336
   1     rs950122     846864    11564    0.04572     0.1012  1.767e-05     0.452       0.6513
   1    rs6657440     850780    11687   -0.06427     0.0819  5.271e-05   -0.7848       0.4326
   1   rs13303101     862124    11689    0.09545     0.2875  9.434e-06    0.3321       0.7399
   1    rs1110052     873558    11654   -0.01181    0.09043  1.464e-06   -0.1306       0.8961
   1    rs3748592     880238    11697    -0.1481     0.1775  5.951e-05   -0.8343       0.4041
   1    rs3748593     880390    11696    -0.5318     0.2519   0.000381    -2.111      0.03478
delta2:~/60days/UQWS_2023/5_unrelGWAS$
```

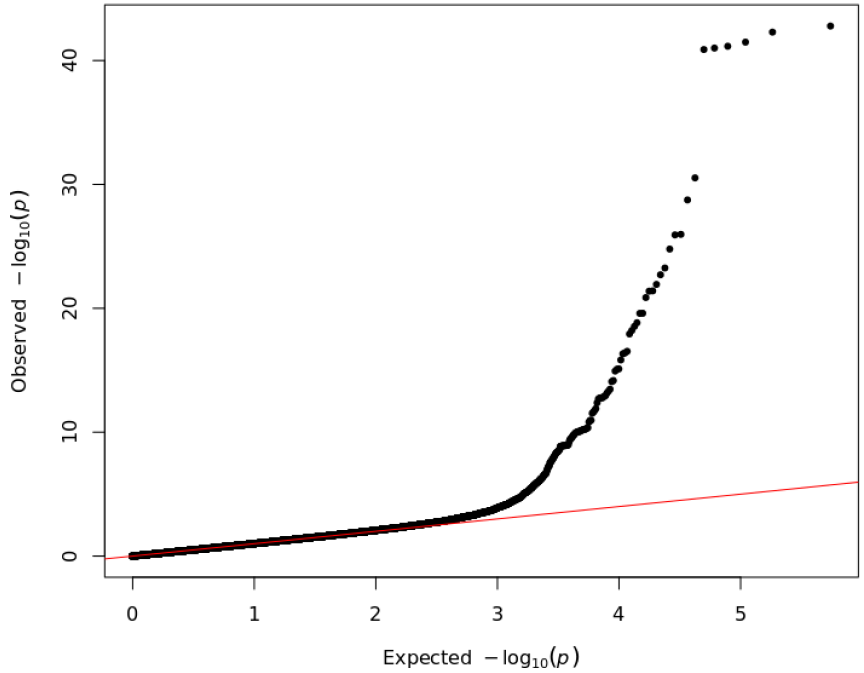
Comparison, raw vs. pre-adjusted

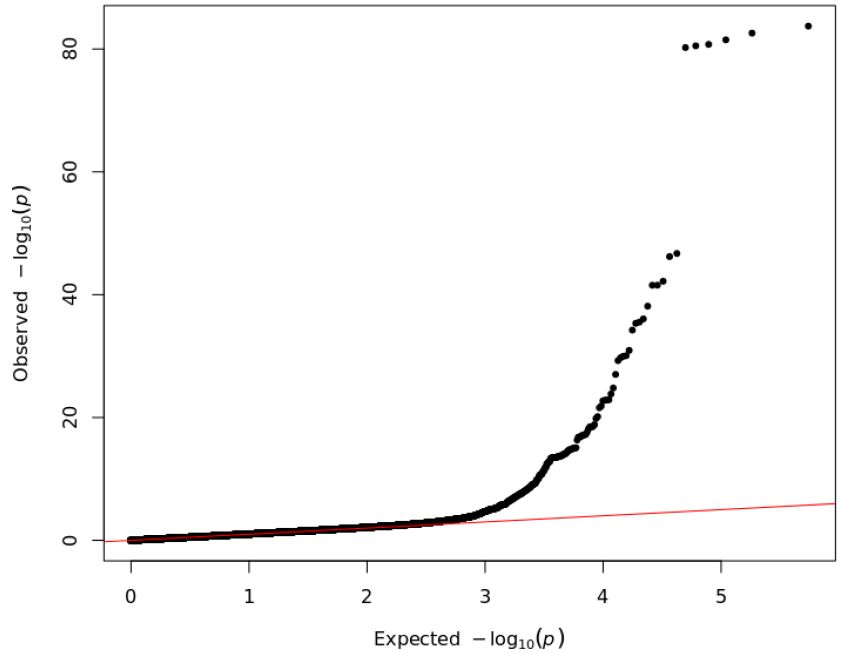




Comparison, raw vs. pre-adjusted

	GIF	P<5x10^{-8}
raw	0.99	102
pre-adjusted	1.01	145





Comparison, binary trait

```
(1) plink --bfile gwasQC --assoc --pheno BMI_binary1.phen --out binary
```

seconds to run

```
(2) plink --bfile gwasQC --logistic --pheno BMI_binary1.phen –out
logistic
```

seconds to run

```
(3) plink --bfile gwasQC --logistic --pheno BMI_binary1.phen –covar
covariates.txt --out logisticCovar
```

took long time to run... I gave up

Output, binary trait

```
delta2:~/60days/UQWS_2023/5_unrelGWAS$ head binary.assoc
 CHR          SNP         BP   A1      F_A      F_U   A2        CHISQ            P           OR
   1   rs12562034     768448    1   0.1102   0.1004    2        3.914      0.04788         1.11
   1    rs4040617     779322    2   0.1286   0.1275    1      0.03569       0.8502        1.009
   1    rs4970383     838555    1   0.2473   0.2474    2    0.0002086       0.9885       0.9995
   1     rs950122     846864    1   0.1982   0.1975    2      0.01094       0.9167        1.004
   1    rs6657440     850780    2    0.385   0.3945    1        1.401       0.2366       0.9611
   1   rs13303101     862124    1  0.01898  0.01958    2      0.07267       0.7875       0.9683
   1    rs1110052     873558    2    0.268   0.2751    1       0.9624       0.3266       0.9645
   1    rs3748592     880238    1  0.05288  0.05384    2      0.06811       0.7941       0.9812
   1    rs3748593     880390    1  0.02259  0.02711    2        3.004      0.08307       0.8294
delta2:~/60days/UQWS_2023/5_unrelGWAS$ head logistic.assoc.logistic
 CHR          SNP         BP   A1       TEST    NMISS         OR         STAT            P
   1   rs12562034     768448    1        ADD    11683       1.11        1.976      0.04816
   1    rs4040617     779322    2        ADD    11667      1.009       0.1885       0.8505
   1    rs4970383     838555    1        ADD    11687     0.9995     -0.01447       0.9885
   1     rs950122     846864    1        ADD    11564      1.004        0.105       0.9164
   1    rs6657440     850780    2        ADD    11687     0.9608       -1.188       0.2349
   1   rs13303101     862124    1        ADD    11689     0.9686      -0.2686       0.7882
   1    rs1110052     873558    2        ADD    11654     0.9638      -0.9908       0.3218
   1    rs3748592     880238    1        ADD    11697      0.981       -0.262       0.7933
   1    rs3748593     880390    1        ADD    11696     0.8264       -1.748      0.08038
```

Output, logistic regression

```
delta2:~/60days/UQWS_2023/5_unrelGWAS$ head logistic.assoc.logistic
 CHR          SNP         BP    A1       TEST    NMISS          OR          STAT             P
   1   rs12562034     768448     1        ADD    11683        1.11         1.976       0.04816
```

```
> model = glm(bmiBinary ~ snp, family=binomial(link='logit'))
> summary(model)

Call:
glm(formula = bmiBinary ~ snp, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7265  -0.6624  -0.6624  -0.6624   1.8026

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.40528    0.02576 -54.556   <2e-16 ***
snp          0.10395    0.05261   1.976   0.0482 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11707  on 11682  degrees of freedom
Residual deviance: 11704  on 11681  degrees of freedom
  (110 observations deleted due to missingness)
AIC: 11708

Number of Fisher Scoring iterations: 4

> exp(0.10395)
[1] 1.109545
```

Binary trait, raw vs. logistic

	GIF	P<5x10^{-8}
raw	0.99	102
pre-adjusted	1.01	145
binary	1.006	64
logistic	1.008	64