Day 2
Optional GWAS Practical

Luke Evans & Wei Zhou

Practical originally developed by Katrina Grasby and Lucía Colodro Conde

Questions: Please ask! You can also use the Q&A box under “Common & rare association”

Learning Goals:

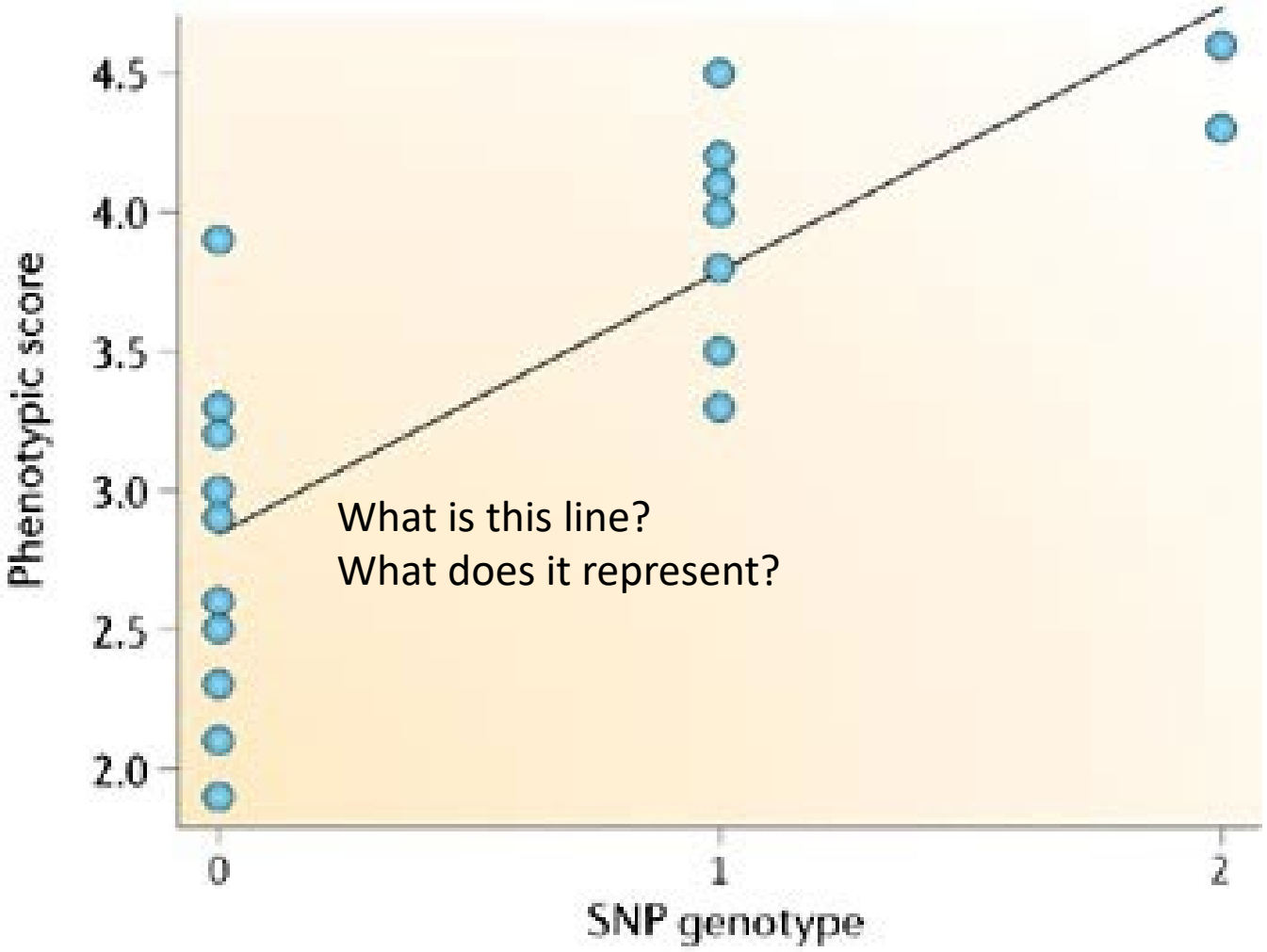
1. Run a GWAS using real data
2. Familiarize yourself with the standard types of output form a GWAS
3. Practice using the command line
4. Improve your familiarity plink2, a common tool in genetics (beyond GWAS, too)
5. Interpret why peaks form in a Manhattan plot, and how to explore the genome around associated loci

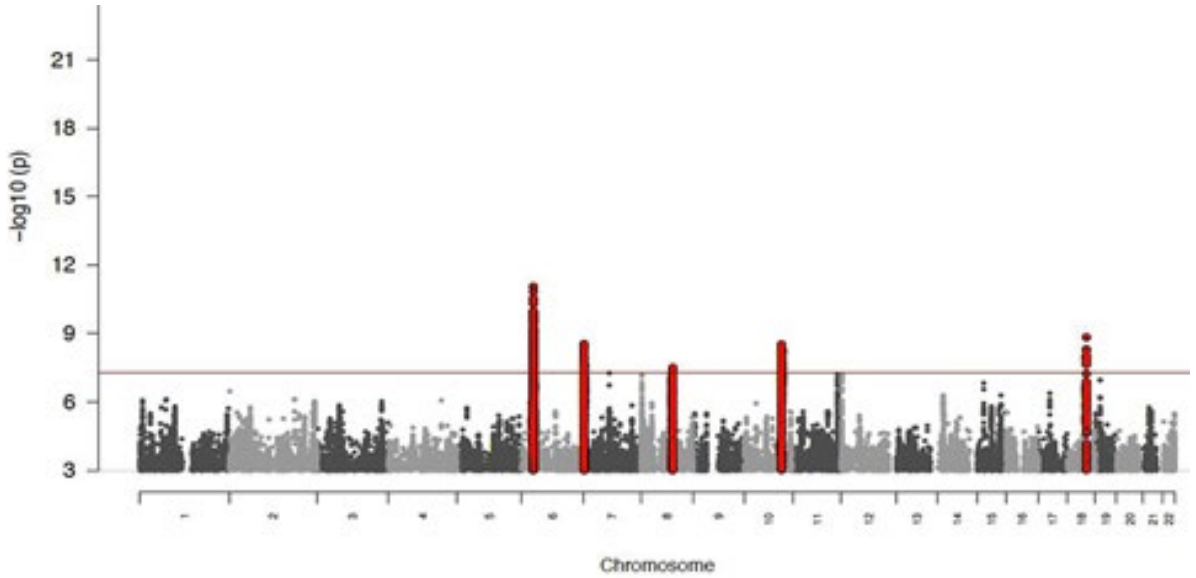
There are 3 exercises in the practical.

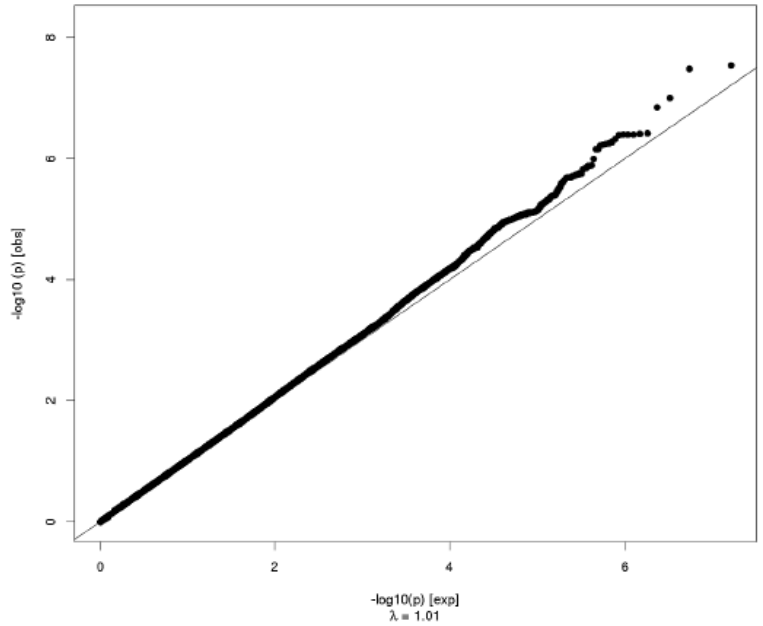
If you don’t finish, that’s OK – these resources are there for you to practice & use as you’re working through your own data.

There are questions throughout the instructions – take some time to either look for the answer (e.g., in the output) or consider what the answers/results are telling you (e.g., why does a q-q plot look a certain way?)

Recap of GWAS from yesterday







Analyzing X Chromosome

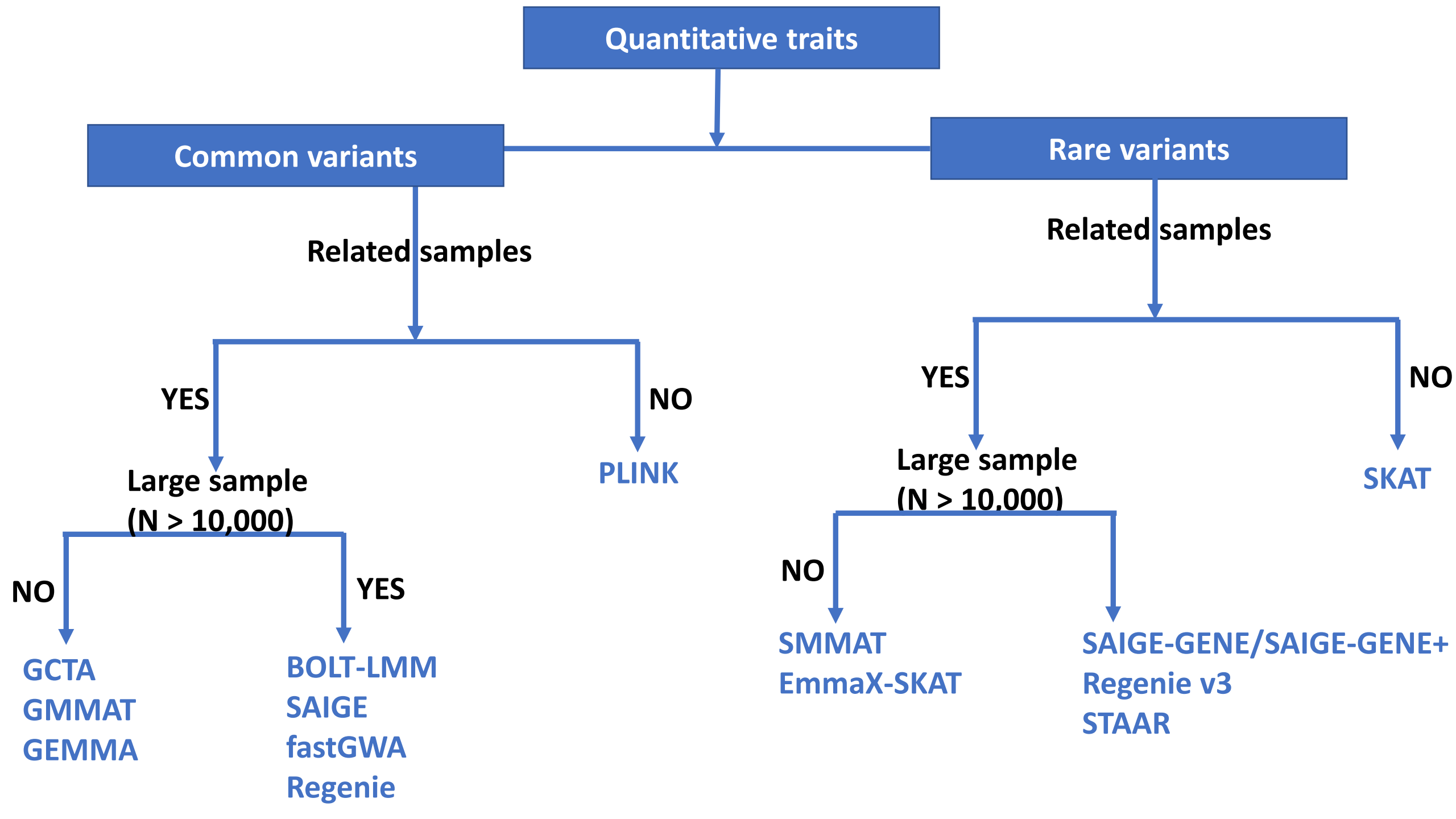
More to come later (Medland)

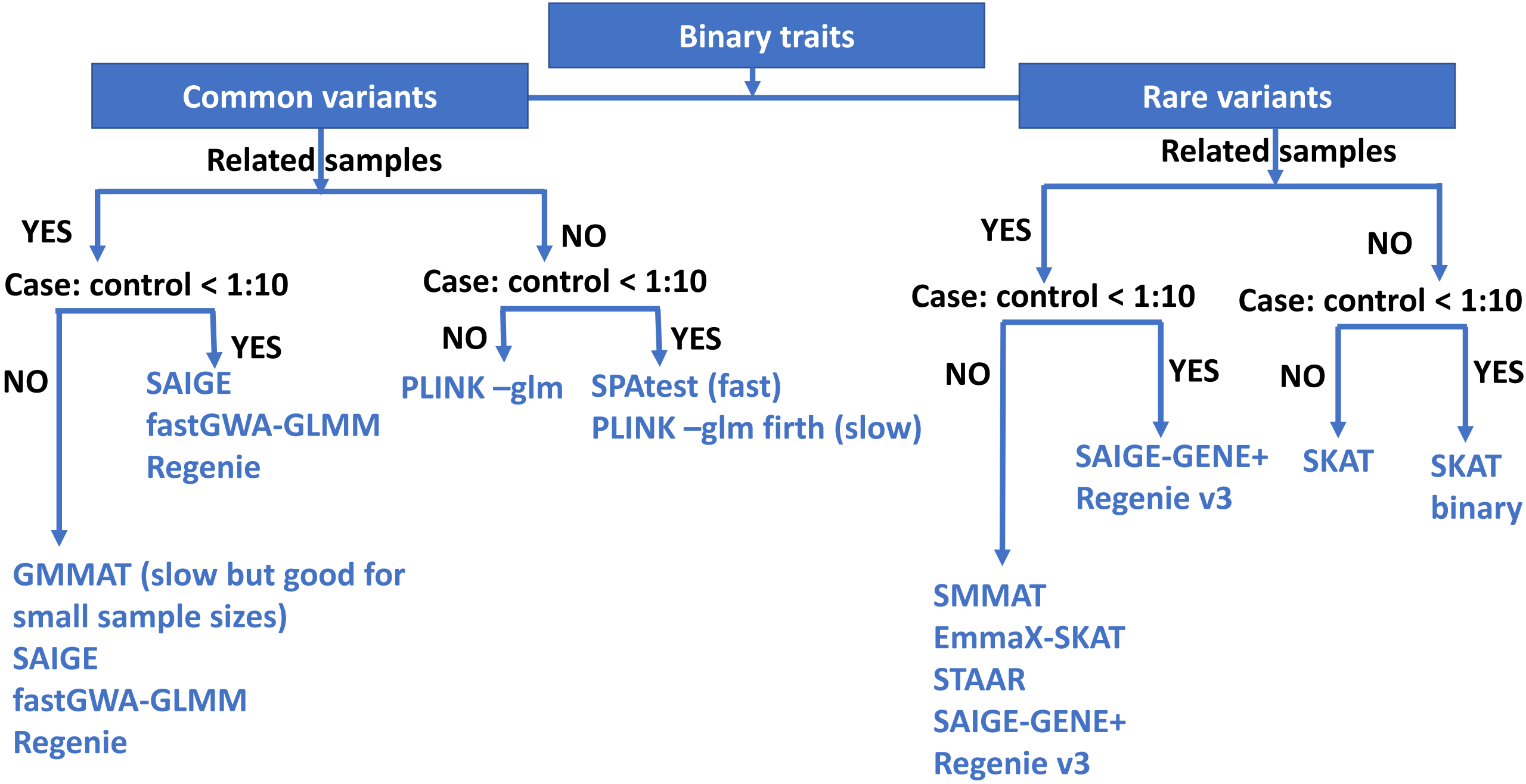
- Often overlooked, but important to analyze, too
- Can analyze it, you just have to do it a little differently than the autosomes
- Imputation can be done and servers understand the different chromosomes and regions on them
- Dosage differences between sexes, dosage compensation and X inactivation are all important features
- X inactivation varies for different tissues

Briefings in Bioinformatics, 2022, **23(5)**, 1–9
https://doi.org/10.1093/bib/bbac287
Review

A systematic review of analytical methods used in genetic association analysis of the X-chromosome

Nick Keur, Isis Ricaño-Ponce, Vinod Kumar and Vasiliki Matzaraki
Corresponding author. Vasiliki Matzaraki, E-mail: Vasiliki.Matzaraki@radboudumc.nl


Quantitative traits
Common variants
Rare variants
Related samples
YES
NO
Large sample
(N > 10,000)
PLINK
NO
YES
GCTA
GMMAT
GEMMA
BOLT-LMM
SAIGE
fastGWA
Regenie
Related samples
YES
NO
Large sample
(N > 10,000)
SKAT
NO
SMMAT
EmmaX-SKAT
SAIGE-GENE/SAIGE-GENE+
Regenie v3
STAAR

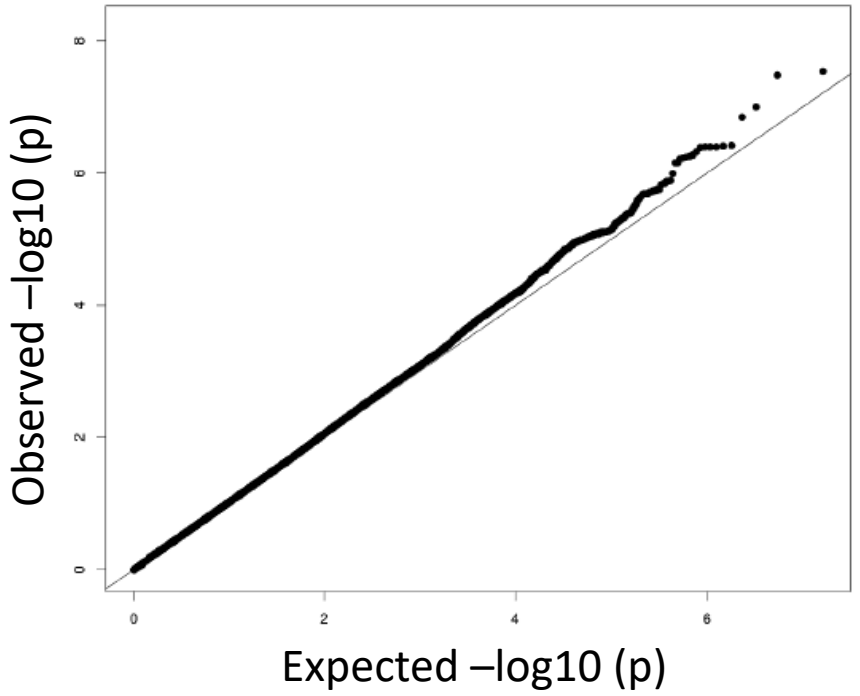

Binary traits
Common variants
Rare variants
Related samples
Related samples
YES
NO
YES
NO
Case: control < 1:10
Case: control < 1:10
Case: control < 1:10
Case: control < 1:10
NO
YES
NO
YES
NO
YES
NO
YES
SAIGE
fastGWA-GLMM
Regenie
PLINK –glm
SPAtest (fast)
PLINK –glm firth (slow)
SAIGE-GENE+
Regenie v3
SKAT
SKAT
binary
GMMAT (slow but good for small sample sizes)
SAIGE
fastGWA-GLMM
Regenie
SMMAT
EmmaX-SKAT
STAAR
SAIGE-GENE+
Regenie v3

There are many tools for GWAS

- The most appropriate method depends on the structure of your data: stratification? Relatedness? Sample size? Etc…
- Mixed linear models: SAIGE, BOLT-LMM, GCTA, FaST-LMM…
- We're going to use plink2 for today's practical: fast, simple command line program that is a general workhorse software for managing data and running analyses.
  - Original plink ped/map format
  - Binary plink format (plink1.9): bed/bim/fam
    - Fast, very useful
  - Plink2 format: pgen/psam/pvar
    - Very fast
    - Saves space compared to others
    - Can include dosage, phase, INFO (similar to VCF format)

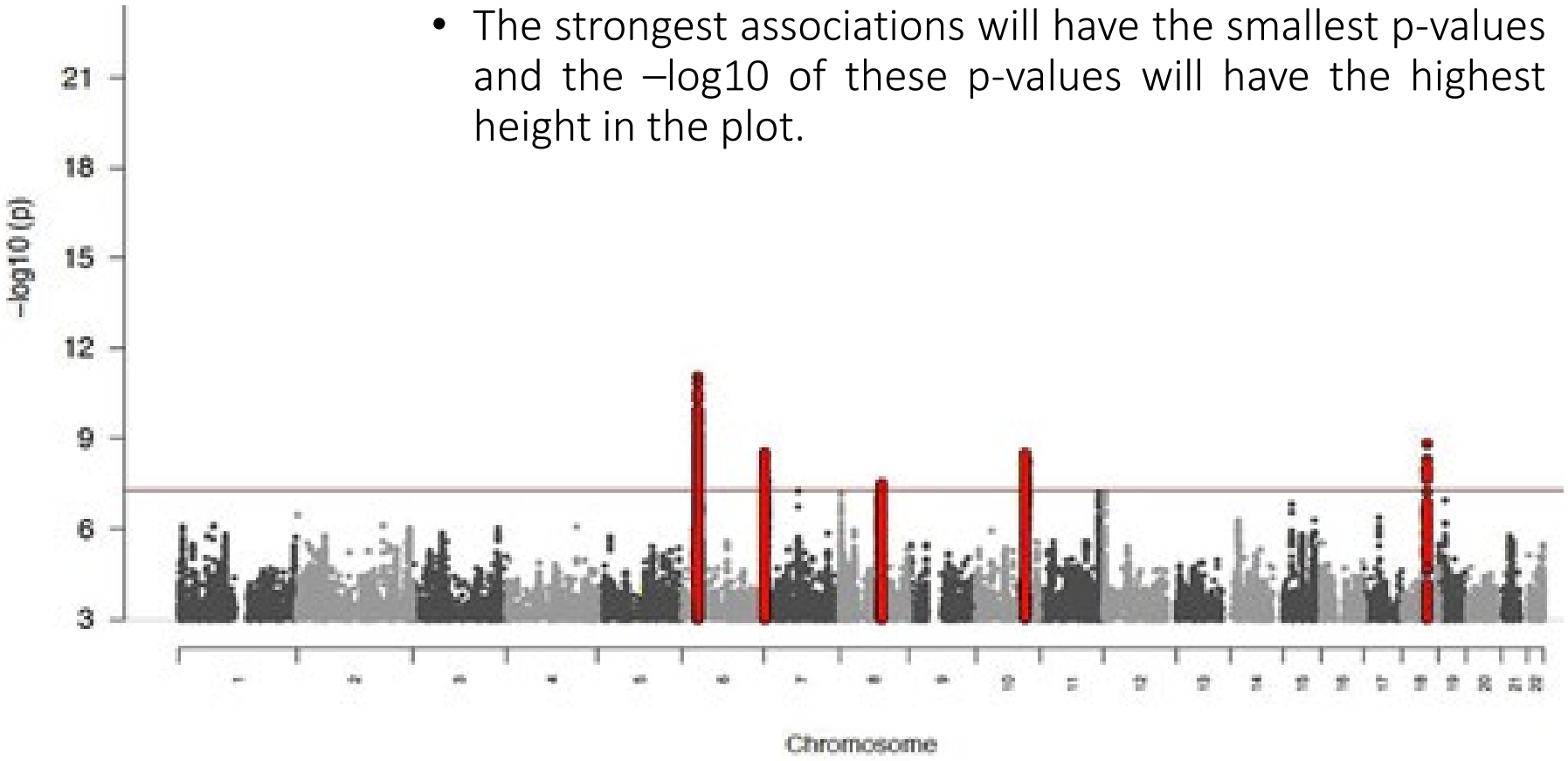
QQ (quantile-quantile) plot

- Checks the overall distribution of test statistics or –log10 p-values of our results with the expectation under the null hypothesis of no association (the diagonal line shows where the points should fall under the null).
- Evaluates systematic bias and inflation (undetected sample duplications, unknown familial relationships, gross population stratification, problems in QC…).

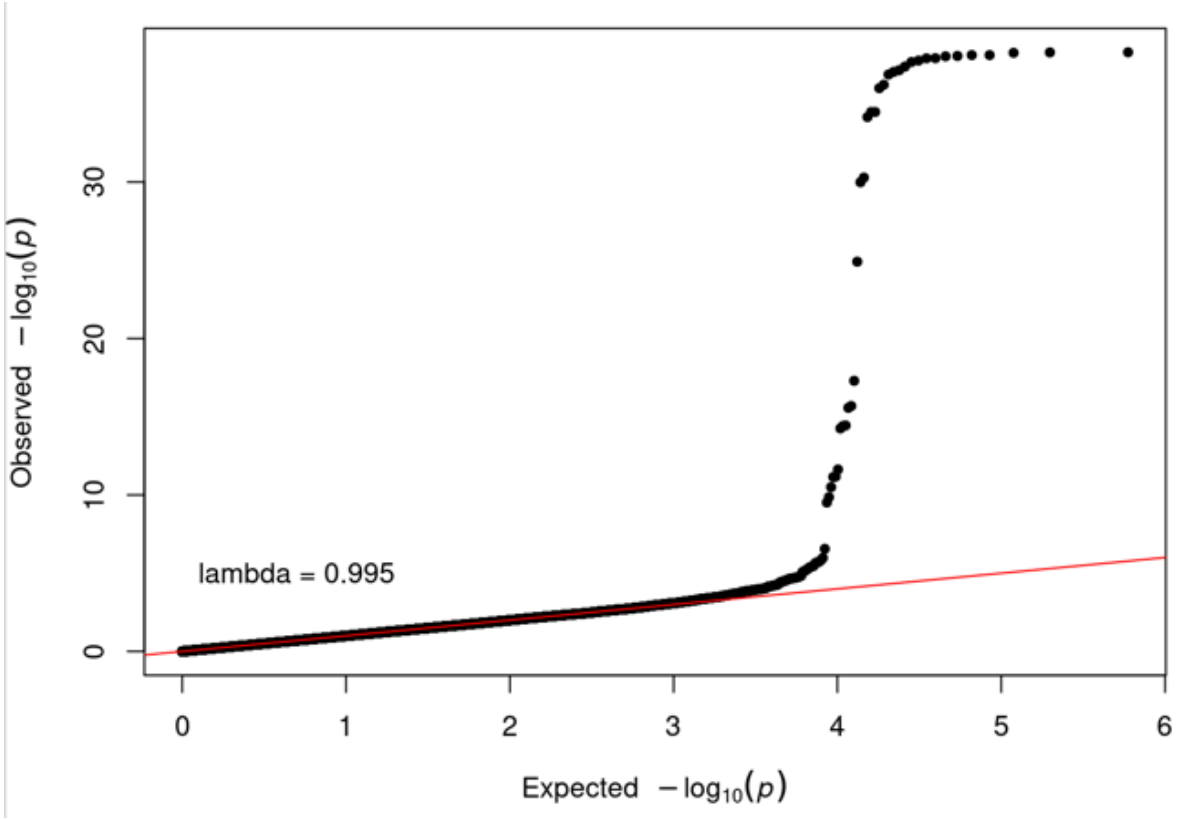


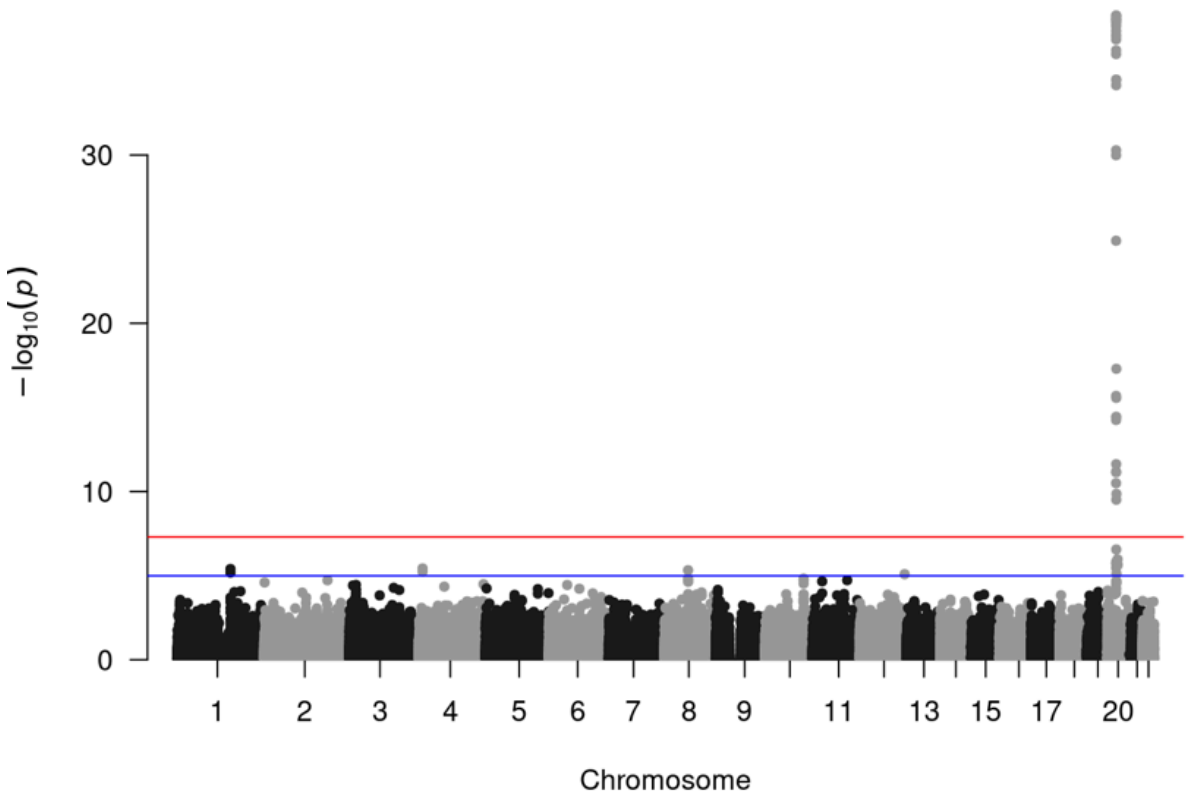
Manhattan plot

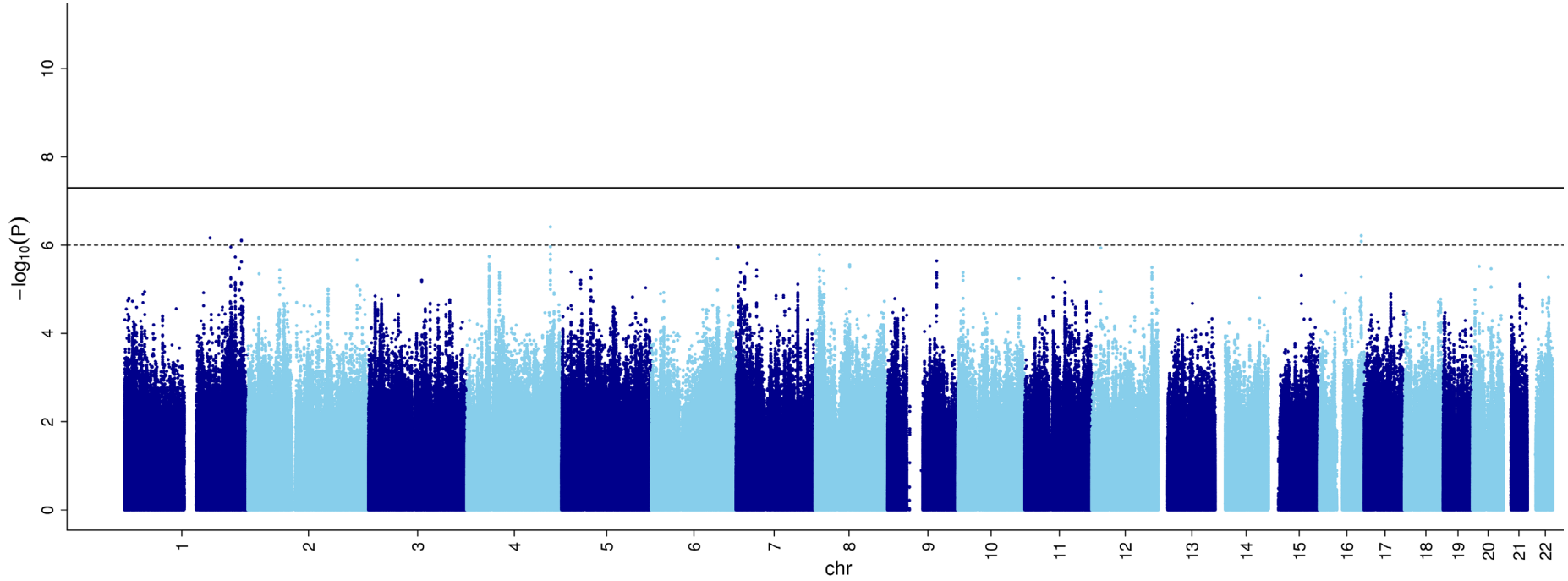
- Plots the -log10 of the association p-value for each SNP against the genomic coordinates.
- The strongest associations will have the smallest p-values and the –log10 of these p-values will have the highest height in the plot.

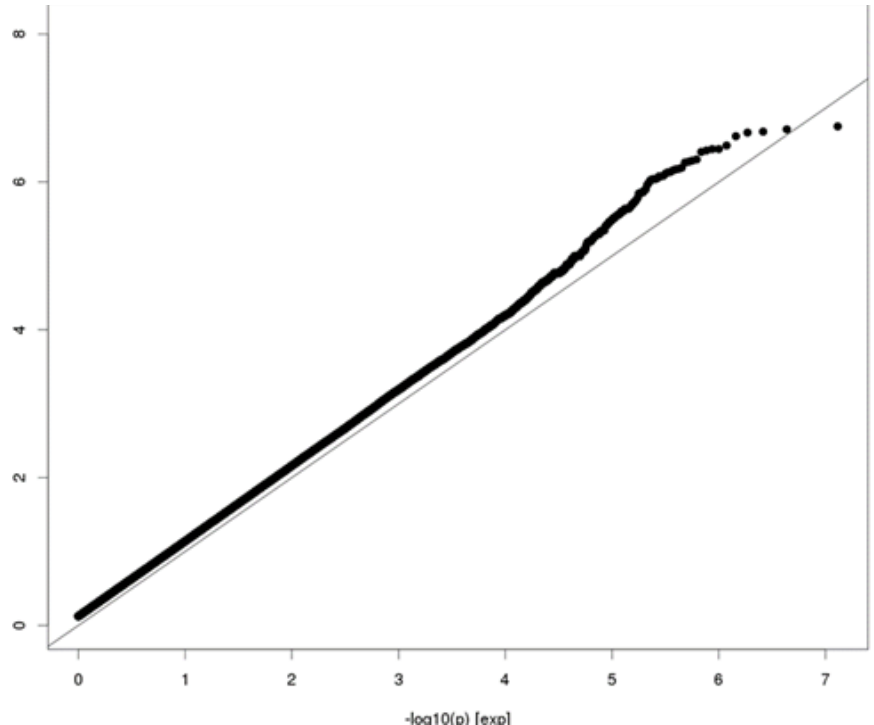


What are these telling us?

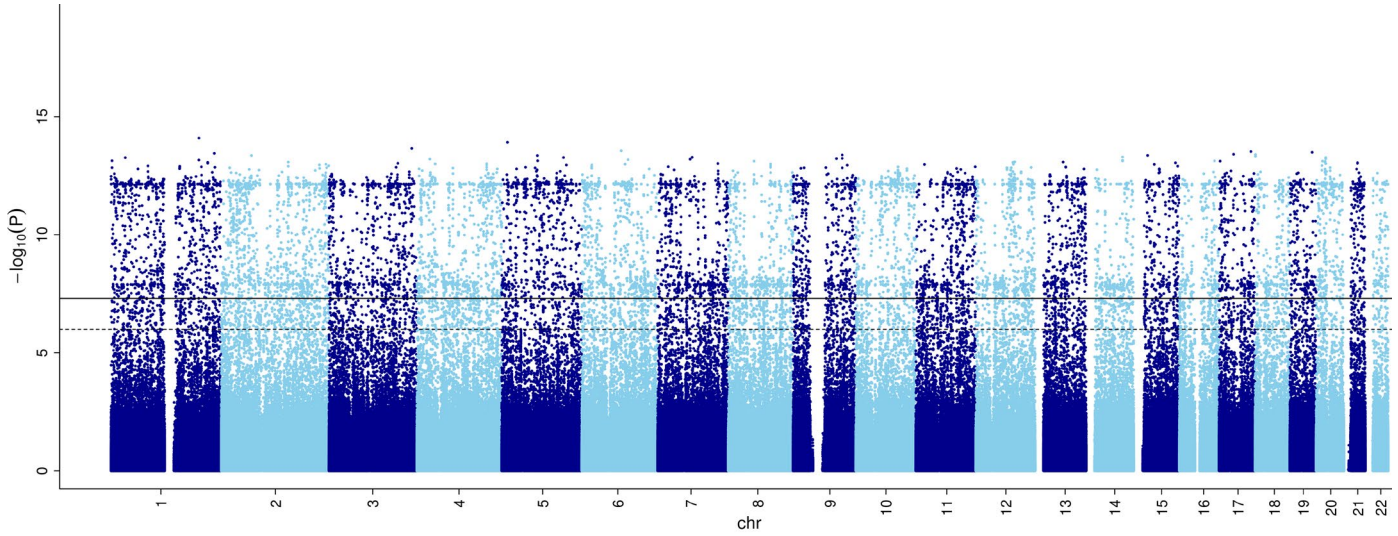


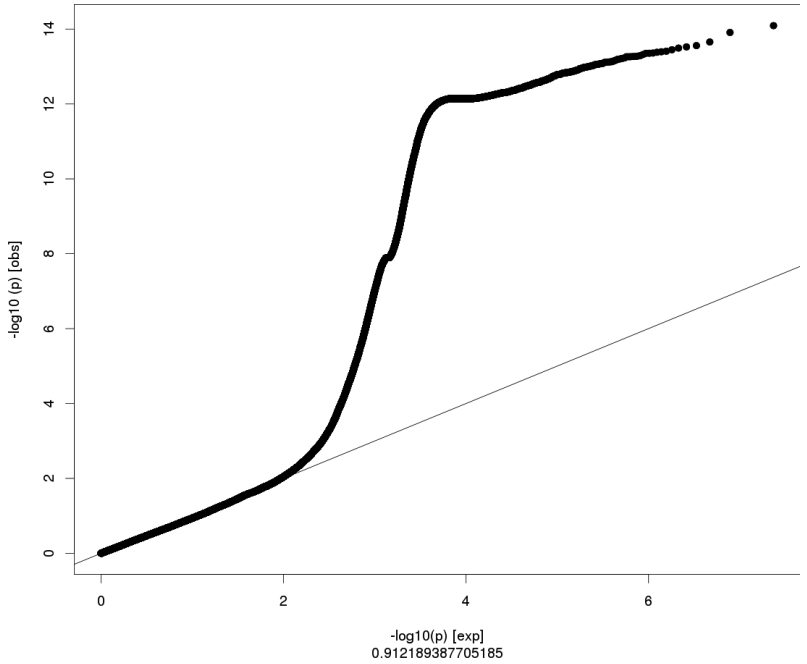



−log10(P)
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
chr

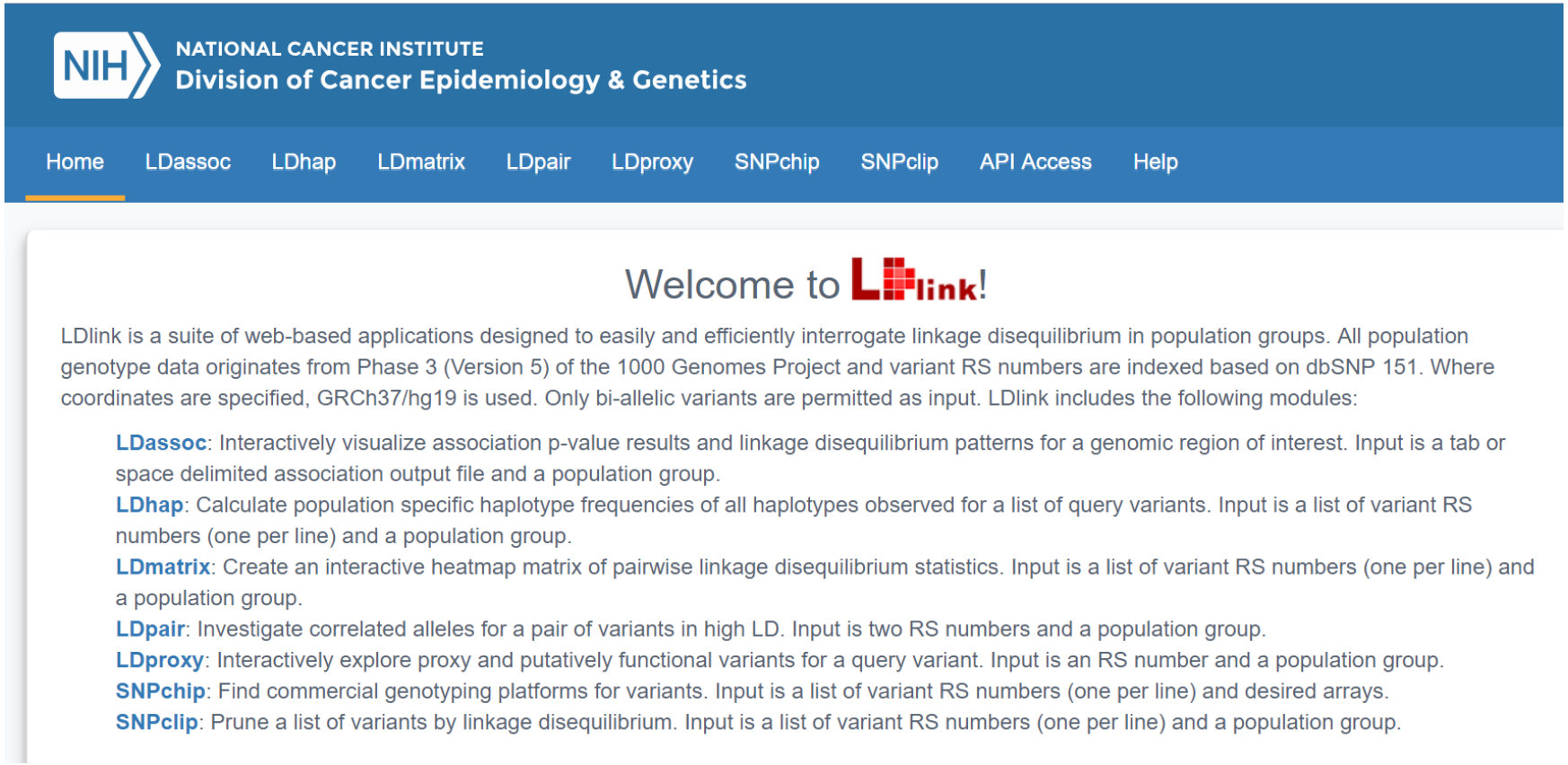

0
2
4
6
8
0
1
2
3
4
5
6
7
-log10(p) [exp]

And these? What are these telling us?





Regional plot



https://ldlink.nci.nih.gov/

NATIONAL CANCER INSTITUTE
Division of Cancer Epidemiology & Genetics

Home | LD Tools | API Access | Citations | Version History | Documentation

LDassoc Tool

Interactively visualize association p-value results and linkage disequilibrium patterns for a genomic region of interest.

prostate_example.txt Browse

Use example GWAS data

Chromosome: chr column

Position: pos column

P-Value: p column

Variant

rs7837688

± 500000 base pair window

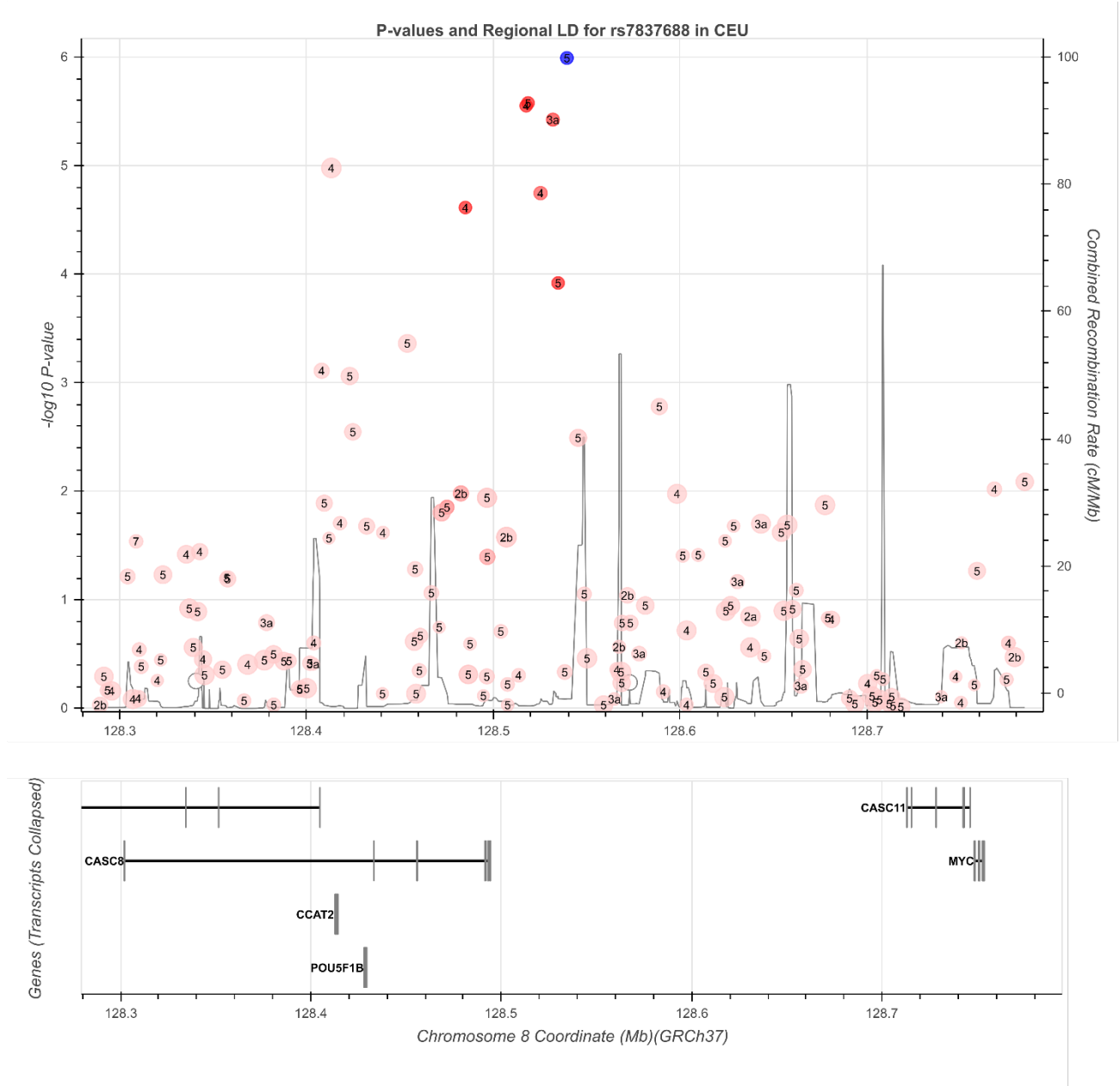
CEU

Calculate

LD Measure: R^2 D'

Collapse transcripts: Yes No

RegulomeDB annotation: Yes No


P-values and Regional LD for rs7837688 in CEU
-log10 P-value
Combined Recombination Rate (cM/Mb)
128.3
128.4
128.5
128.6
128.7
Genes (Transcripts Collapsed)
CASC8
CCAT2
POU5F1B
CASC11
MYC
Chromosome 8 Coordinate (Mb)(GRCh37)

Today's practical data are freely available from:

ARTICLE

Genetic Control of Human Brain Transcript Expression in Alzheimer Disease

Jennifer A. Webster,[1,2,3,16] J. Raphael Gibbs,[4,5,16] Jennifer Clarke,[6] Monika Ray,[7] Weixiong Zhang,[7,8] Peter Holmans,[9] Kristen Rohrer,[4] Alice Zhao,[4] Lauren Marlowe,[4] Mona Kaleem,[4] Donald S. McCorquodale III,[10] Cindy Cuello,[10] Doris Leung,[4] Leslie Bryden,[4] Priti Nath,[4] Victoria L. Zismann,[1,2] Keta Joshipura,[1,2] Matthew J. Huentelman,[1,2] Diane Hu-Lince,[1,2] Keith D. Coon,[1,2,11] David W. Craig,[1,2] John V. Pearson,[1,2] NACC-Neuropathology Group,[12] Christopher B. Heward,[13,17] Eric M. Reiman,[1,2,14] Dietrich Stephan,[1,2,14] John Hardy,[4,5] and Amanda J. Myers[10,15,*]

Source:

https://med.miami.edu/faculty/amanda-myers-phd

https://xzmxbgsv808roffneicreq.on.drv.tw/www.lfun/LFUN/LFUN/DATA.html

Set up – we'll go through the first few commands together

We will use the text editor, the terminal, R Studio, and the browser.

Copy all the files by typing in your working directory:

```
cp -r /faculty/luke/2023/gwas2/ ./
```

The practical is "Day2_GWAS_Practical_INSTRUCTIONS.pdf", and is available on the file sharing site.

A plain text version is also on the file sharing site

The rubric is "Day2_GWAS_Practical_RUBRIC.pdf", and is available on the file sharing site.

Some commands in the instructions are missing (indicated with XXX) – it's up to you & your group to determine what to include there.

Questions are in the instructions – some are about the specific results (what is N?) while others are asking you to think about what the results are telling you (QQ plot look OK?)