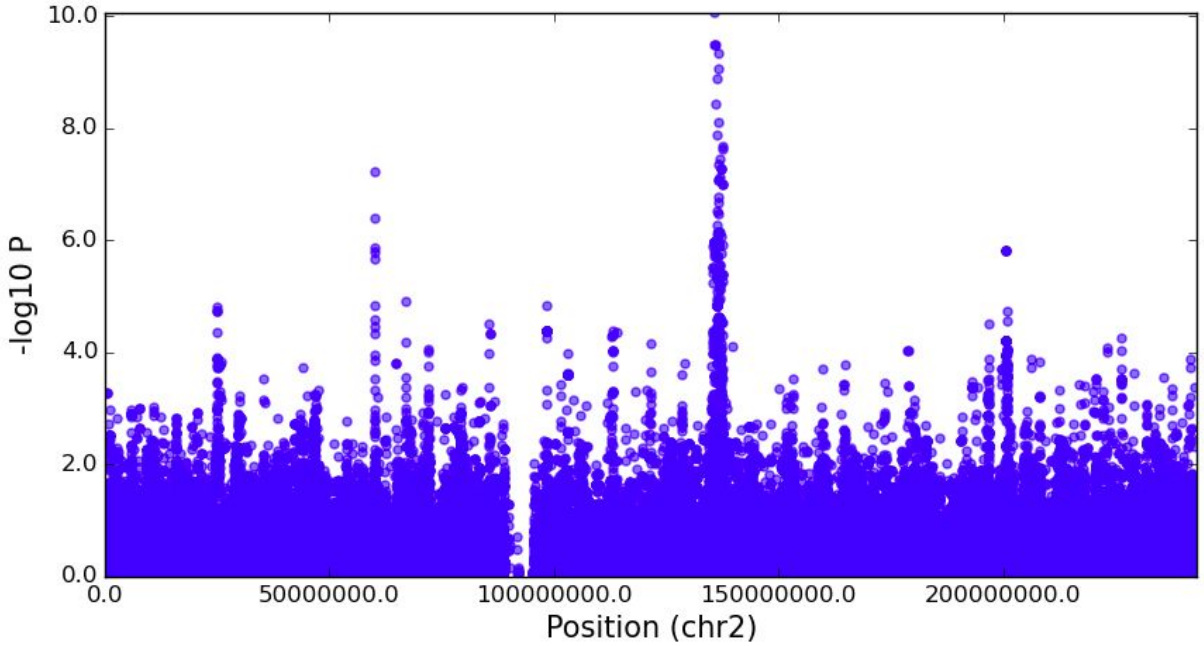
Part 1: A basic GWAS (4 points)

Exercise 1

SNPs with pvalue “NA” could not be tested because there was no variation at those positions in our dataset

Manhattan plot:



“Vertical lines” represent many SNPs in strong linkage with the “causal” SNP.

Exercise 2

192 SNPs have p<10^-4, using e.g. the command
cat /oasis/projects/nsf/csd524/mgymrek/ps3/results/ps3_gwas.assoc.linear.tab | grep -v NA | awk '($9<10**-5)' | wc -l

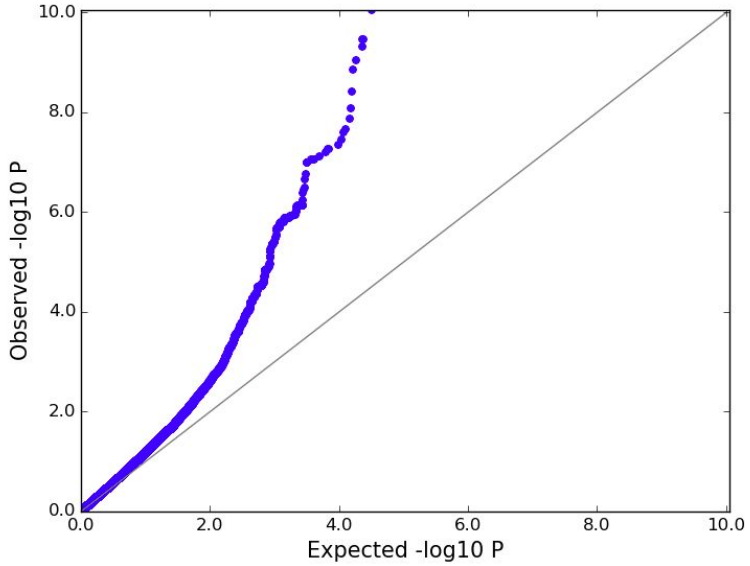
Using the clump command, e.g.

```
# Get independent signals
plink \
    --bfile ${PREFIX} \
    --clump ${USERPREFIX}.assoc.linear --clump-field P \
    --clump-p1 0.0001 \
    --clump-p2 0.01 \
    --clump-r2 0.5 \
    --clump-kb 250 \
```

--out ${USERPREFIX}

We find around 41 different signals
After removing the outlier region, we find 22 signals.

Exercise 3

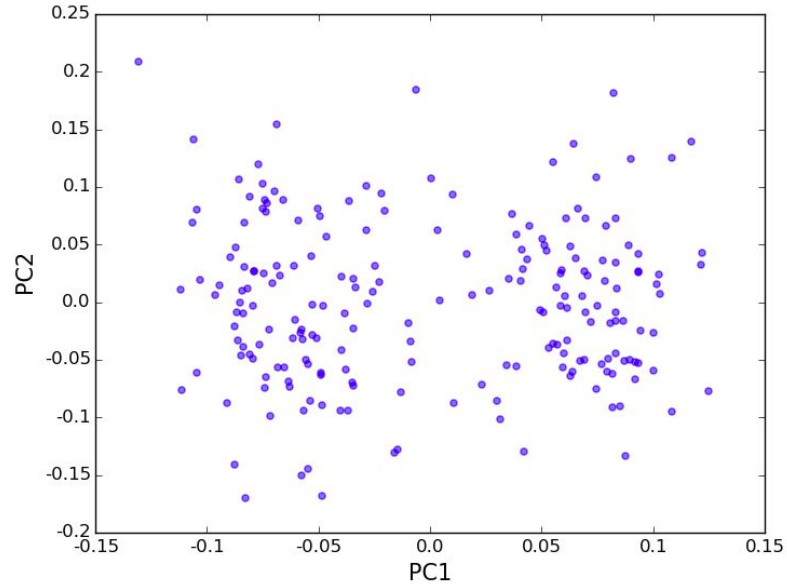


Strong departure from the diagonal, suggests inflated p-values. Perhaps there is a confounding factor (e.g. population structure) we are not controlling for).

Exercise 4
TMEM163, ACMSD, CCNT, RAB3GAP1, ZRANB3, R3HDM1, UBXN4, **LCT,** MCM6, DARS, CXCR4
Yes, we have seen LCT, which controls lactase persistance. The relevant mutation shows strong differences between populations. It also lies on long haplotypes due to recent positive selection, which could explain the many many apparent signals falling in that region. This suggests the signal is due to confounding by population structure, and is not truly associated with height.

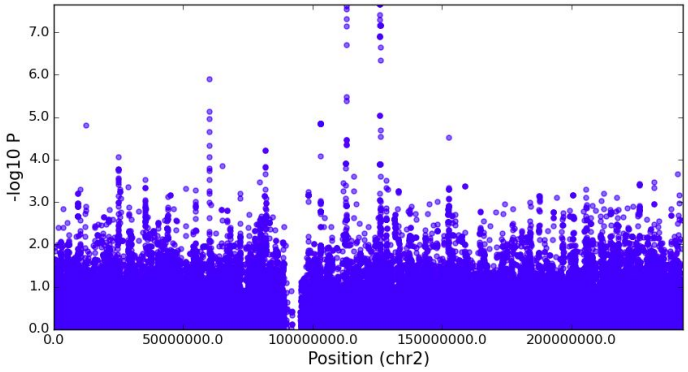
Part 2: Confounding by population structure (3 points)

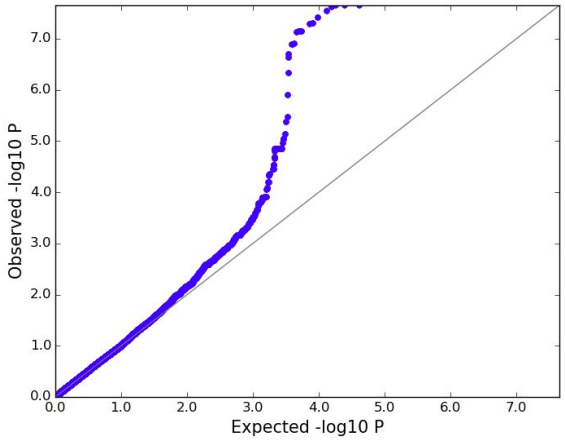
Exercise 1



PC1 separates CEU from TSI

Exercise 2





Exercise 3

The signal at LCT is now gone. The qqplot shows much less inflation in the p-values, suggesting we remove much of the confounding population structure. There are new hits on the Manhattan plot that weren't visible before.

Pros: remove variation that is simply correlated with ancestry, and not causal in the trait
Cons: This could remove true signals. Imagine there is a variant that is fixed at allele A in CEU and fixed at allele B in TSI, and this variant really is correlated with height. If we control for population group, this signal would be masked.

Part 3: Predicting eye color (3 points)

NOTES:

- **Many people switched the directions. E.g. the blue probability was reported as brown, and vice versa. I took off only 0.5 pts for this, since the Irisplex paper appears to have a mistake in their notation. Thus if you were following directly from the paper this could happen. If your answers matched the answer key *exactly* except for mixing brown/blue up, I didn't count off.**
- **Many people had close to but not quite the right answers. For that I took off only 0.5 for the whole problem.**

Exercise 1

NA12249 has 0.968458 chance to have blue eyes.
NA20509 has 0.97542 chance to have brown eyes.
NA12750 has 0.27 chance for blue, 0.50 chance for "other", 0.23 chance for brown eyes

Exercise 2

CEU
blue: 0.631302930401
brown: 0.238685614717
other: 0.130011454881
TSI
blue: 0.250104760778
brown: 0.583218459929
other: 0.166676779293

CEU more likely to have blue eyes

Exercise 3

Effect size gives the change in the log odds ratio for each minor allele.

- **I was hoping for the precise definition given here, but nobody reported this. However I did not count off for for descriptive answers.**

Exercise 4

It falls in an intron of HERC2. It is not a coding region and doesn't seem to be an annotated transcription factor. It is thought that this mutation lowers promoter activity of a nearby gene OCA2.