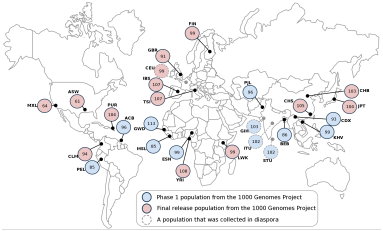
In search of the "opera loving gene"

Vivian Link

Is there a genetic basis to loving opera music?



Genotypes:

Subset of 1000 Genomes project

(Largest public catalogue of human genotype data)

Phenotypes:

Minutes per year of listening to opera

Part 0: Download folder “handout” from blackboard (5 min)

- Which file contains the phenotypes?
- Which file contains the genotypes?
- Which file contains variant information?

Part 1: Test SNPs for association with phenotype using plink2 (5 min)

1) Run association test with the following command:

```
plink2 --bfile subset --glm allow-no-covars --out subset
```

2) Look at resulting file

Part 1: Test SNPs for association with phenotype using plink2 (5 min)

1) Run association test with the following command:

```
plink2 --bfile subset --glm allow-no-covars --out subset
```

2) Look at resulting file

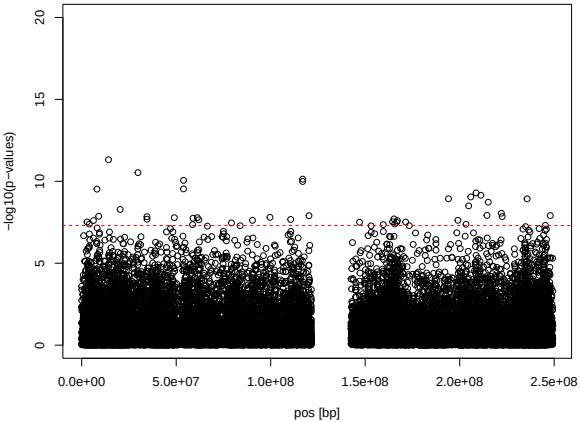
subset.PHENO1.glm.linear

#CHROM	POS	ID	REF	ALT	A1	TEST	OBS_CT	BETA	SE	T_STAT	P	ERRCODE	
1	10177	rs367896724	A	AC	AC	ADD	301	28.7964	10.0658	2.86081	0.00452339	.	
1	11008	rs575272151	C	G	G	ADD	301	1.21975	19.9121	0.0612568	0.951196	.	
1	13116	rs62635286	T	G	G	ADD	301	20.2637	14.5949	1.38841	0.166046	.	
1	13273	rs531730856	G	C	C	ADD	301	12.8326	13.9448	0.920242	0.358188	.	
1	14599	rs531646671	T	A	A	ADD	301	-3.18051	14.2043	-0.223912	0.822979	.	
1	15820	rs2691315	G	T	T	ADD	301	-25.7153	8.18301	-3.14253	0.00184297	.	
1	15903	rs557514207	G	GC	G	ADD	301	34.6933	9.38815	3.69544	0.000261007	.	
1	54712	rs568927205	T	TTTTC	T	ADD	301	7.9405	10.8861	0.729416	0.466318	.	
1	54716	rs569128616	C	T	T	ADD	301	29.3231	10.4421	2.80815	0.00531056	.	

Part 2: Make manhattan plot of association results of chr 1 (10 min)

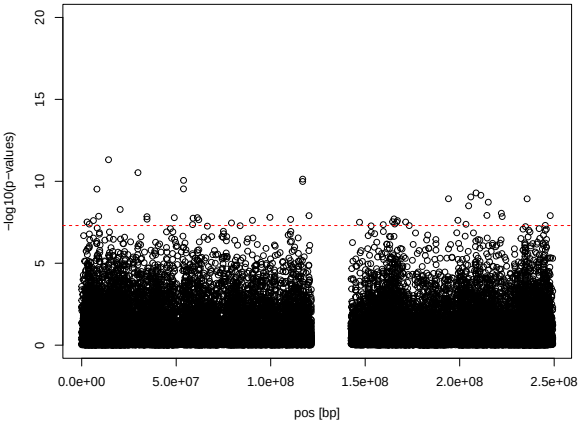
1) Read in the results table
2) Use function `plot_manhattan_chr_plink2()` to plot results for chromosome 1

Association Results chr 1



- What is the hole in the middle?
- How many peaks are there?

Association Results chr 1



- What is the hole in the middle?
- How many peaks are there?
- Does this seem realistic?

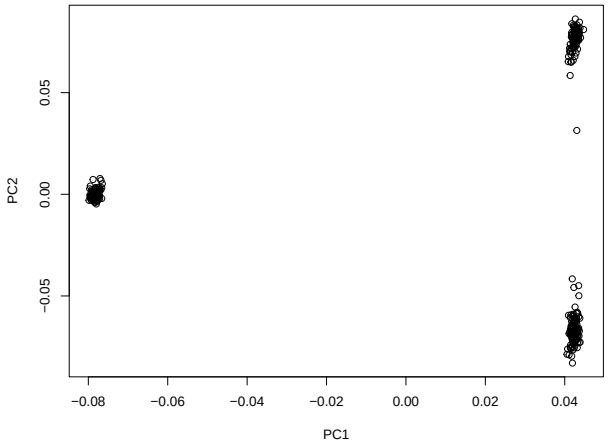
Part 3: make PCA for the subset (10 min)

1) Make a PCA with 10 PCs with plink2 using the following command:

```
plink2 --bfile subset --pca 10 --out subset
```

2) Read file `subset.eigenvec` into R
3) Plot the loadings of the first two PCs for every individual (x-axis is PC1 and y-axis is PC2)

PCA



Why might the PCA look like this?

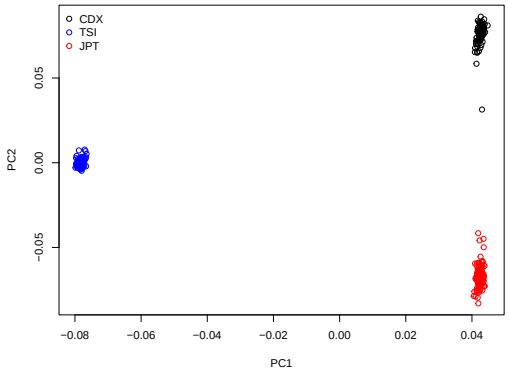
Part 3: make PCA for the subset (10 min)

1) Make a PCA with 10 PCs with plink2 using the following command:

```
plink2 --bfile subset --pca 10 --out subset
```

2) Plot the loadings of the first two PCs for every individual
3) Read the information from file `subset.fam` into R
4) Color the points in PCA according to the first column in the fam file

PCA



Can you say something about the relative positions of the populations?

(Hint: TSI are Italians, CDX are chinese and JPT are Japanese)

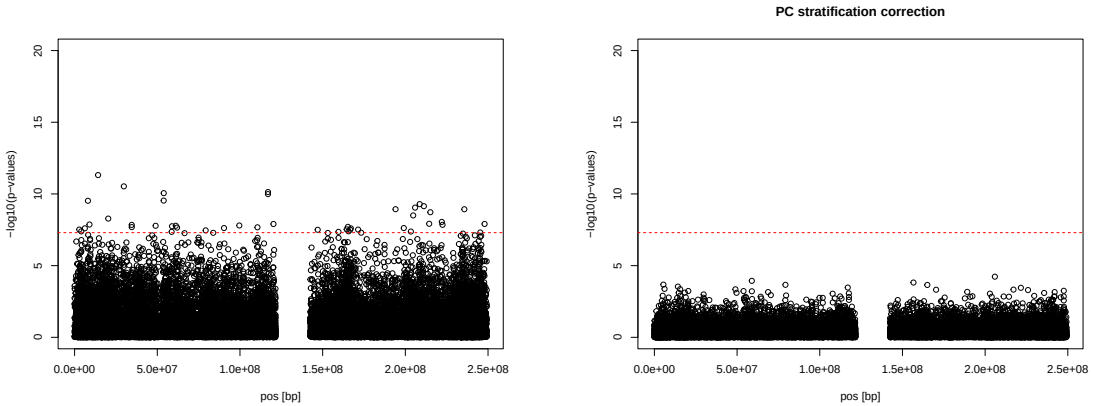
Part 4: Run association tests while correcting for PCs (5 min)

1) Run plink2 with the following command:

```
plink2 --bfile subset glm --out subset_withPCCorrection --covar
subset.eigenvec
```

2) Read resulting file subset_withPCCorrection.PHENO1.glm.linear into R
3) Plot manhattan plot for chromosome 1 using function `plot_manhattan_chr_plink2()`

Association results



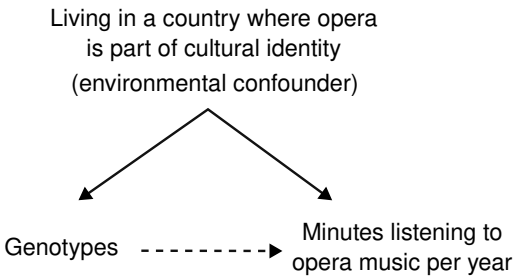
- All the hits disappear when we correct for population structure
- I simulated the phenotypes with an environmental component that correlates with population structure

Simulating Phenotypes

```
simulate_phenotypes_environmental_confounding <- function(name_fam, offset){
  fam <- read.table(name_fam)
  colnames(fam) <- c("FID", "IID", "PID", "MID", "Sex", "Phenotype")
  num_inds <- length(fam$FID)
  phenotypes <- rnorm(n=num_inds, mean=300, sd=100)
  names(phenotypes) <- fam$IID
  TSI <- fam$IID[fam$FID == "TSI"]
  phenotypes[TSI] <- phenotypes[TSI] + offset

  return(phenotypes)
}
```

Associations were caused by confounding factor



- I simulated random phenotypes and then added 100 to the individuals from TSI
- Alleles that happen to be more frequent in TSI correlate with phenotype
- Conditioning on the population removes association between genotypes and phenotypes

How can we know if we removed all the bias?

Since we know there are no real associations, how should our p-values be distributed?

How can we know if we removed all the bias?

Since we know there are no real associations, how should our p-values be distributed?

$\rightarrow$ uniformly

How can we check if the p-values follow the correct distribution?

How can we know if we removed all the bias?

Since we know there are no real associations, how should our p-values be distributed?
$\rightarrow$ uniformly

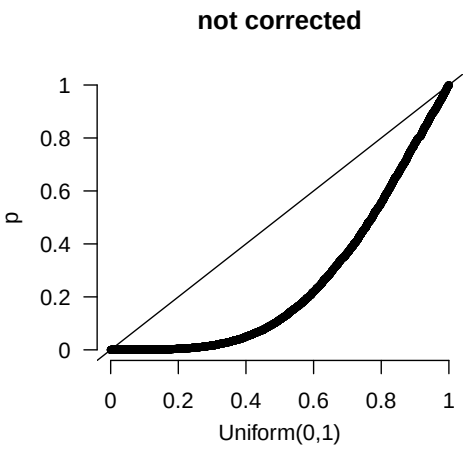
How can we check if the p-values follow the correct distribution?
$\rightarrow$ Quantile-Quantile plot

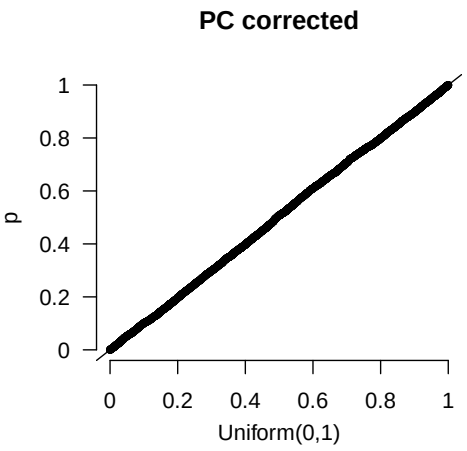
A Q-Q plot compares the quantiles of your dataset with the quantiles of a theoretical distribution (like a uniform distribution)

Part 5: Check p-value distribution (5 min)

Use function `plot_qq()` to plot the p-value distributions for the association tests run without correction and run with PC correction

p-value distributions





Correcting for structure using the GRM (GCTA)

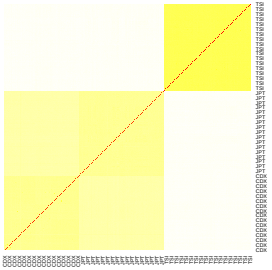
Model:

GCTA tests for association using a linear mixed model, where the GRM is a random effect. In PCA correction, the PCs are fixed effects

Correcting for structure using the GRM (GCTA)

Model:

GCTA tests for association using a linear mixed model, where the GRM is a random effect. In PCA correction, the PCs are fixed effects



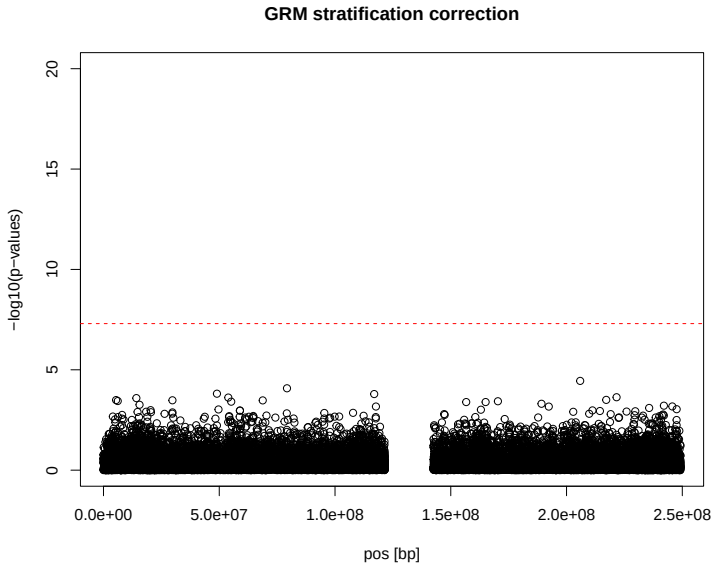
Part 6: Use GRM to correct for population structure using GCTA (15 min)

1) Download GCTA from
   https://yanglab.westlake.edu.cn/software/gcta/#Download
2) Extract the files into your working directory
3) Run GCTA using this command:

```
gcta-1.94.1-linux-kernel-3-x86_64/gcta-1.94.1 --mlma-loco --bfile
subset --out subset --pheno subset.fam --mpheno 4
```

4) Plot manhattan plot for chr1 with the function `plot_manhattan_chr_GCTA()`

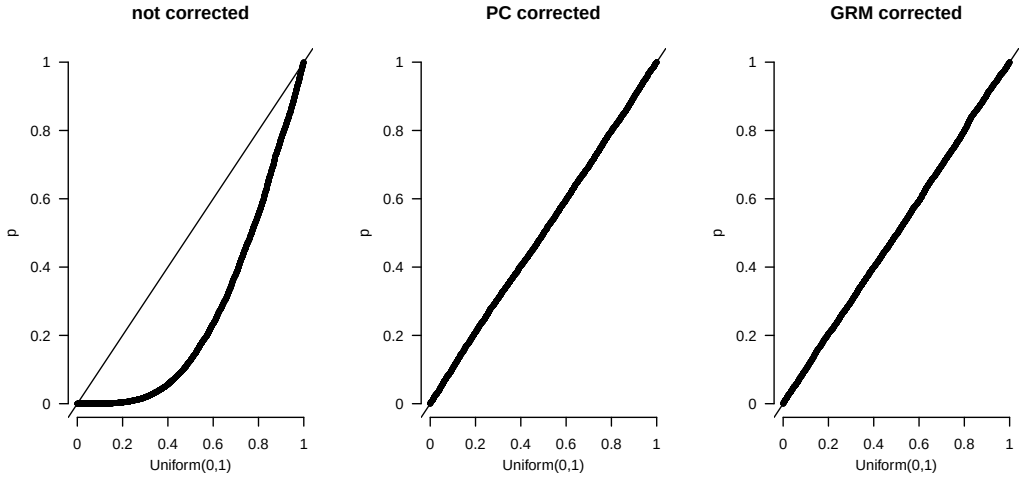
Part 6: Use GRM to correct for population structure using GCTA



Part 7: check p-value distributions

Use function `plot_qq()` to plot the p-value distributions for all three association tests (no correction, PC correction and GRM correction)

p-value distributions



Other ways to simulate population stratification?

Other ways to simulate population stratification?

Genetic confounding

- Choose random variants on chromosome 1 to be causal
- Associate effect size to each of them
- Simulate the phenotype for each individual based on his genotypes at the causal variants and their effect sizes
- Test variants of other chromosomes for association

$\rightarrow$ The phenotype mean will be different for each population and variants for which the allele frequencies show a similar pattern will look like they explain the phenotypic differences