Introduction to GWAS and MR

Introduction to genetic concepts and Genome-Wide Association Studies (GWAS)

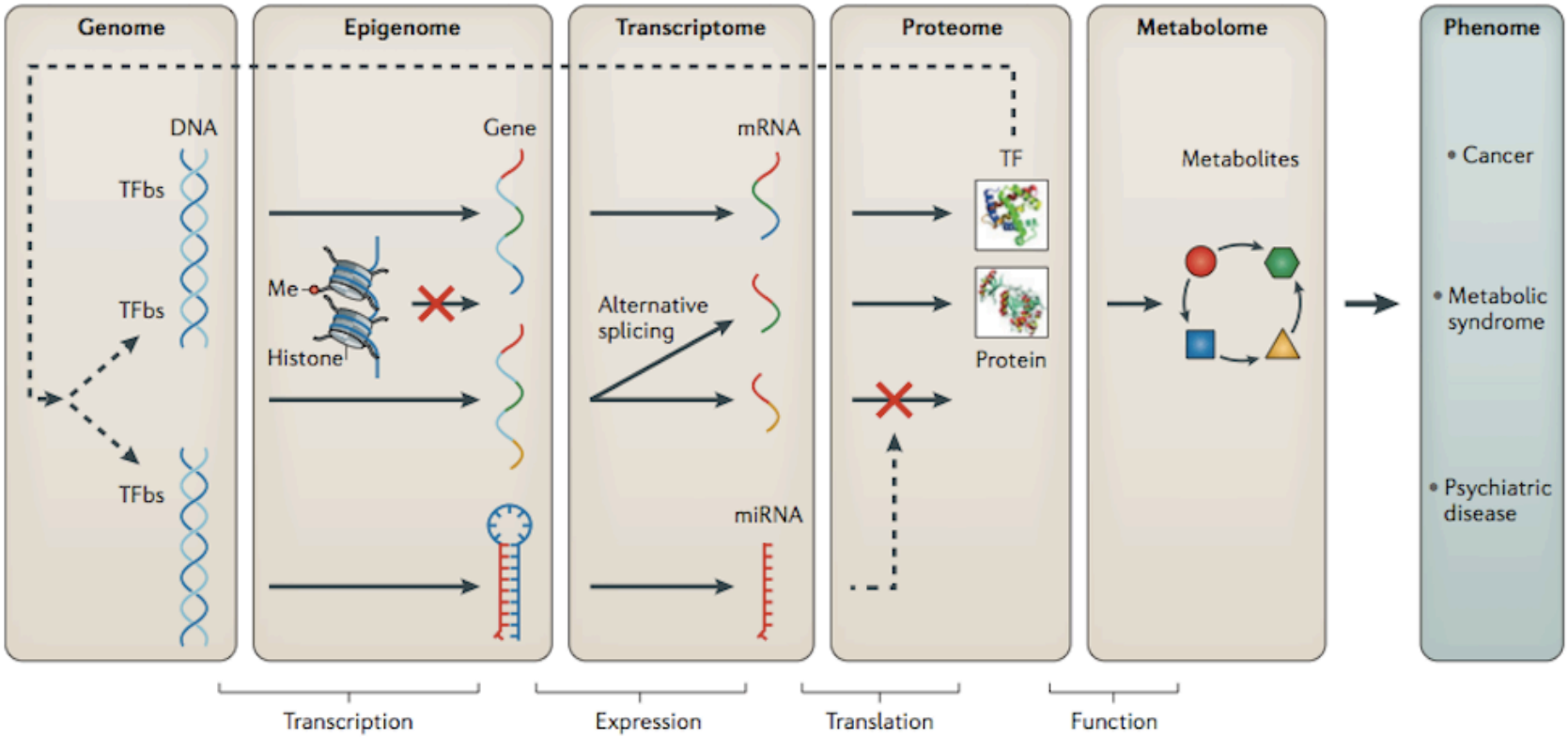
Overview

- **Day 1:**
- Genetic terms and concepts: HWE, population stratification
- Introduction to GWAS
  - Way to determine genetic variants associated with a trait
- Tutorial on GWAS
- Need plink2 and R installed before the tutorial

- **Day 2:**
- Genetic terms and concepts: pleiotropy
- Introduction to Mendelian randomization (MR)
  - Method that uses genetic variants to determine the relationship between 2 traits
- Tutorial on MR
- Need R installed before the tutorial and a TwoSampleMR token

DNA to Phenotype

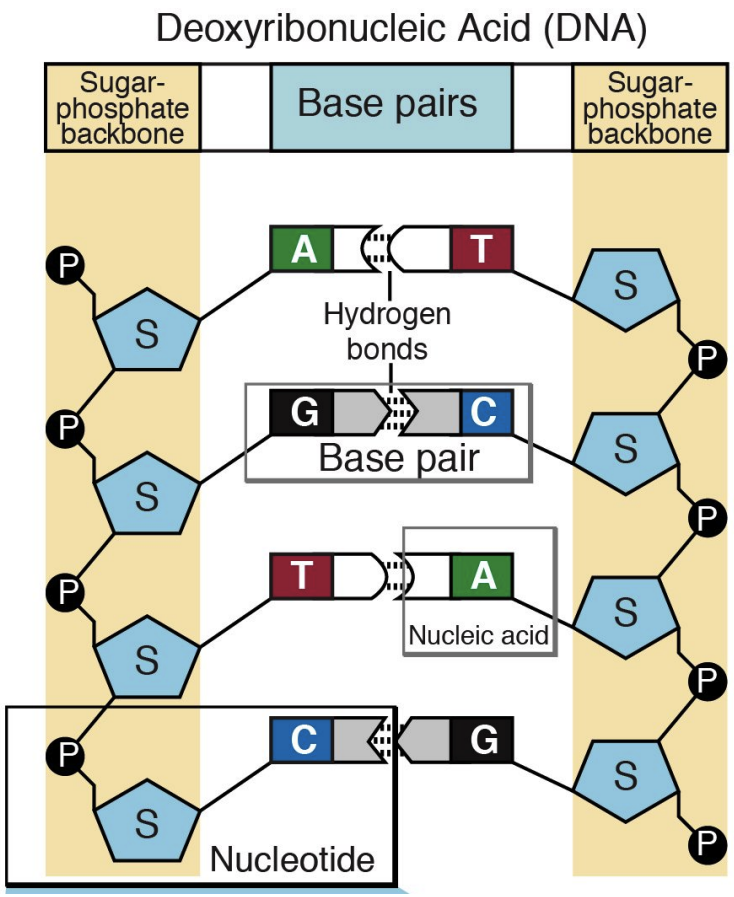
- DNA— [transcription] —> RNA — [translation] —> Protein
- DNA is like the recipe & protein dictates function



- Transcription factor (TF) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA (mRNA), by binding to a specific DNA sequence
  - TFbs= transcription factor binding sites
- Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. (2015) Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews. 16:85-98.

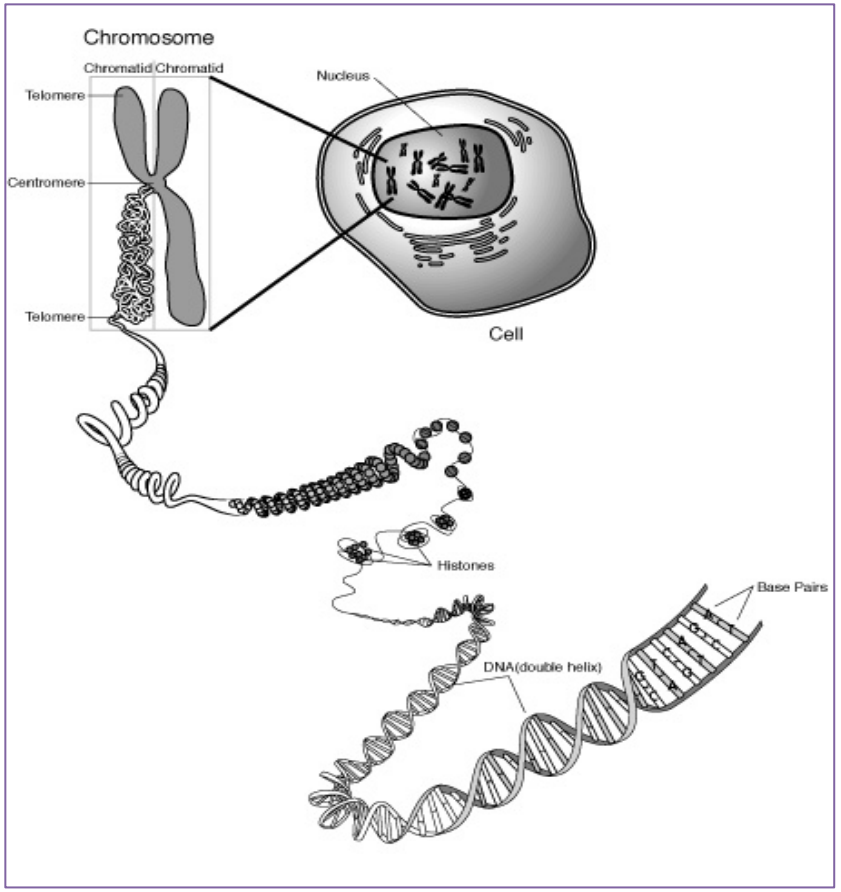
Deoxyribonucleic Acid (DNA)

Each ‘ladder rung’ has 2 complementary base pairs



A Adenine
T Thymine
C Cytosine
G Guanine

Each chromosome is made up of two strands of deoxyribonucleic acid (DNA) in a double helix arrangement

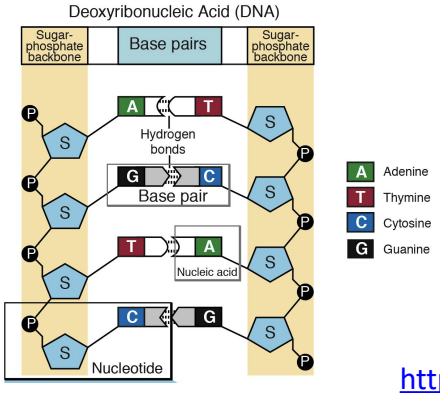


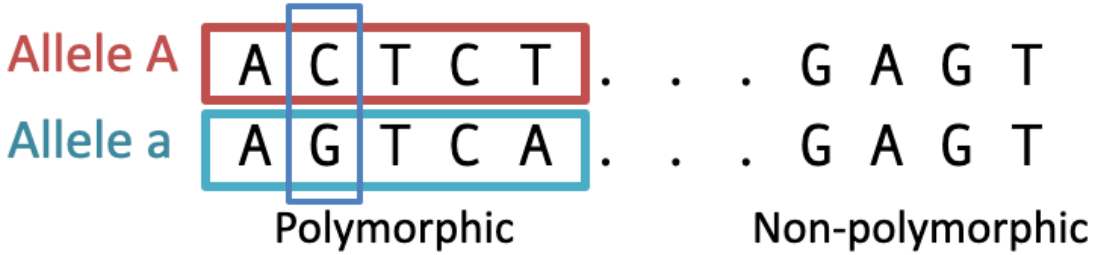
http://knowgenetics.org/nucleotides-and-bases/

NHGRI

Genes

- The human **genome** is made up of 3 billion base-pairs.
- **Genes** are linear stretches of DNA. The human genome contains about 20,000 protein coding genes and thousands of non-protein coding genes
- **Locus:** Particular location in the genome. (Plural: Loci)
- **Alleles / Variants:** Different forms (i.e. different DNA sequences) of the same gene or genetic locus. Many different types of variation.
- **Polymorphic:** Polymorphic loci have several different alleles. At other loci, there is no variation from person to person.
- **Genetic Marker:** Random variable providing information on DNA sequence at a particular locus; many different types of loci and types of information. Markers are inherently categorical.

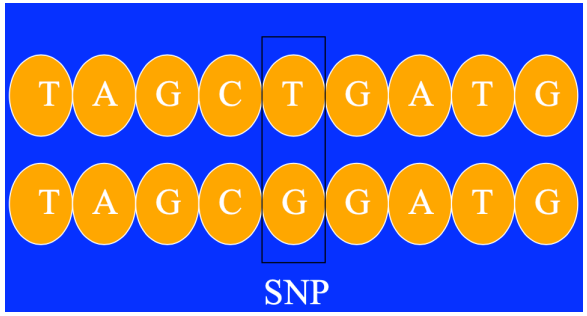




http://knowgenetics.org/nucleotides-and-bases/

SNPs

- **SNP:** single nucleotide polymorphism
  - DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered
  - On average, a SNP exists about every 100-300 base pairs. About 12 millions SNPs on a genome
- **Terms:** genetic variants, markers (more general), SNPs
- **Goal:** find SNPs associated with trait



Terms to Know

- **Genotype:** Pair of alleles at a locus (i.e. DD, Dd, dd or 2,1,0 [#D])
- **Heterozygote:** genotype with different alleles on the two chromosomes (i.e. Dd or 1)
- **Homozygote:** genotype with the same alleles (i.e. DD, dd or 2/0)
- **Dominant:** the allele of a gene that masks or suppresses the expression of an alternate allele
  - [**DD** (Homozygote) &**Dd** (Heterozygote)](trait) vs [**dd** (Homozygote)](no trait)
- **Recessive:** an allele that is masked by a dominant allele
  - [**DD**(Homozygote)](trait) vs [**Dd** (Heterozygote) & **dd**(Homozygote)](no trait)

Using Regression Models to Specify Mode of Inheritance

Code genotype as a variable X; coding can be chosen to represent different genetic models for a general Y;

$$E(Y|X) = \beta_0 + \beta_1 * X$$

Ways of coding X to represent different modes of inheritance:

Recessive	
X	G
1	DD
0	dd/Dd

Dominant	
X	G
1	DD/Dd
0	dd

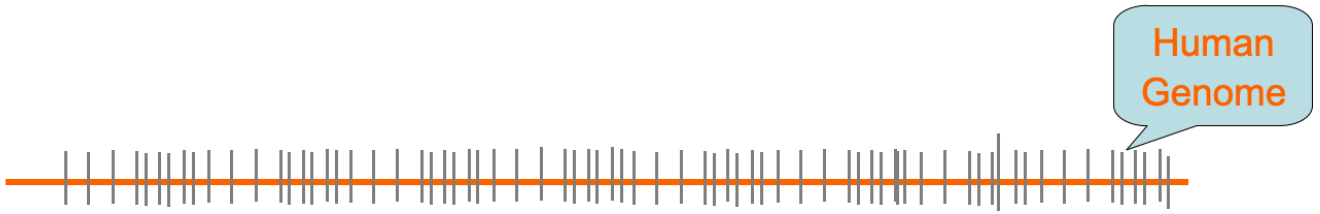
Additive	
X	G
2	DD
1	Dd
0	dd

Co-Dominant (need two indicator variables to code 3 groups): $E(Y|X) = \beta_0 + \beta_1 * X1 + \beta_2 * X2$

X1	X2	G
1	0	DD
0	1	Dd
0	0	dd

Genome-wide association study

Definition: Association analysis performed with a panel of polymorphic markers adequately spaced to capture most of the linkage disequilibrium information in the entire genome in the study population. Usually: 500k - 4m SNPs



- Test each SNP for association with disease phenotype
- SNPs may not be independent due to linkage disequilibrium (LD)
- LD:

GWAS

- Test each SNP for an association with the outcome separately
  - $g(E[Y])=\beta_0+\beta_{snp}*X+\beta_C C$ where the SNP X=0,1,2
  - Most common models: linear or logistic regression
- C are covariates
  - Adjust for precision variables (age, etc)
  - Adjust for confounders (population stratification)
    - Include PCs as covariates in model
- Need to decide on genetic coding
  - Recessive, dominant, co-dominant, additive
  - Additive model is most common (X=0,1,2)
  - The co-dominant test is not a favorite with geneticists
    - Because it doesn't require that probabilities should be monotone in number of disease alleles

GWAS

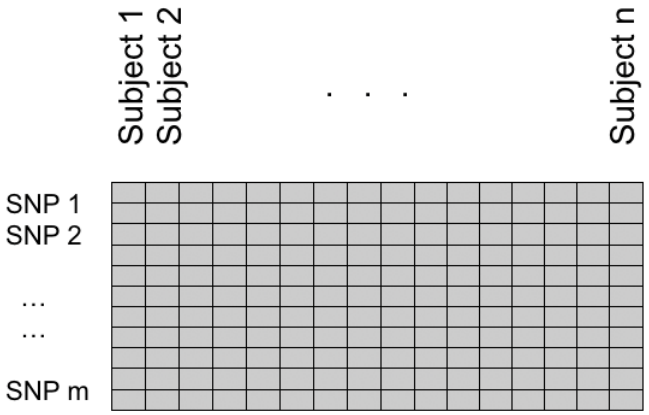
- For 500k - 4m SNPs, test each SNP for an association with the outcome separately
- Running 500k-4m regressions (not independent due to LD)
- Computationally efficient packages (plink), but don't check model assumptions or source of association
- Source of association:
  - Causal association: SNP influences disease susceptibility
  - Linkage disequilibrium (LD): SNP associated with other nearby SNPs that influence susceptibility
  - Population stratification: Confounding, phenotypes and allele frequencies differ between populations (genetic drift)
- Theory is simple, but there can be a lot of issues due to both biological and statistical assumptions and considerations

Analysis outline

- Step 1: Genotyping/ Data cleaning / QC
- Step 2: Association test between genotype and phenotype
  - Analyze each SNP separately
  - Often logistic or linear regression using additive model, including important covariates
  - Control for population substructure
    - Add principal components as covariates
- Step 3: Visualize and check results
  - Use Q-Q plots to visualize obvious difficulties or issues
  - Use Manhattan plots to visualize results
    - Multiple testing issue
      - Use Bonferroni: cutoff of 0.05/ #SNPs ($\sim 5*10^{-8}$)

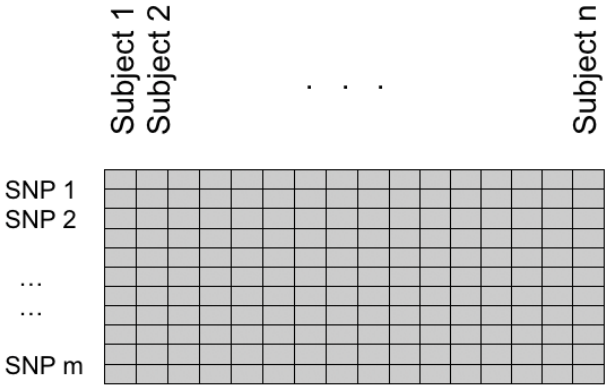
Variety of Quality Assessment Techniques

- Sample quality
  - Missing call rate over SNPs
  - Median genotype confidence score
- SNP quality
  - Missing call rate over samples
  - Duplicate sample discordance
  - *Hardy-Weinberg Equilibrium
- Genotyping batch quality
  - Missing call rate of samples in a batch
  - Allelic frequency difference relative to a pool of other batches



Variety of Quality Assessment Techniques

- Sample identity
  - Planned duplicate sample check
  - Relatedness (relatives)
- Case versus control confounding
  - *Population Stratification
  - Missing call-rate differences
- Preliminary association tests
  - *QQ plots can help identify systematic issues
  - *Manhattan signal plot



Hardy Weinberg Equilibrium (HWE)

Theorem: Allele frequencies in a population remain constant if no evolutionary forces exist.

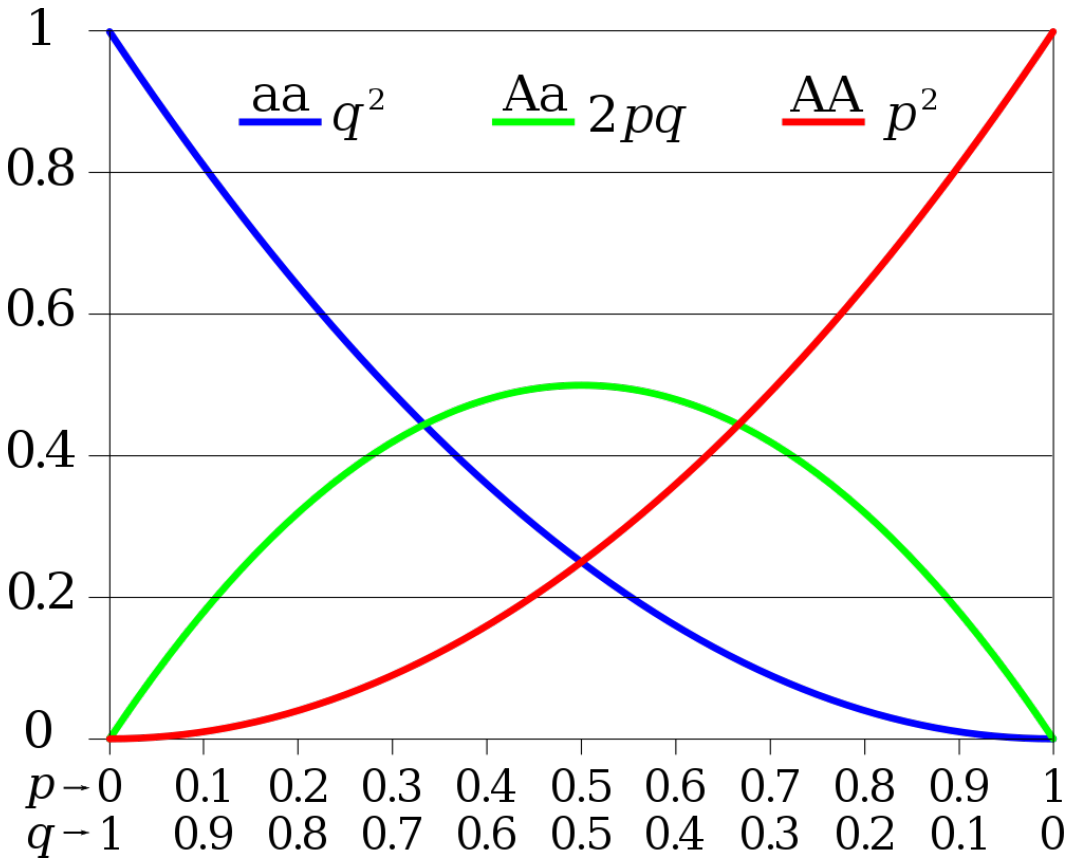
Requirements for Hardy-Weinberg equilibrium:

- Random mating
- No Inbreeding
- Large population
- Discrete generations
- No mutation
- No migration
- No selection

Departures from HW equilibrium provide a mechanism to study evolution

Hardy Weinberg Equilibrium (HWE)

Rule: If you know allele frequency, use HWE to calculate genotype probabilities (p=P(A) and q=P(a); q=1-p)



Johnuniq
https://commons.wikimedia.org/wiki/File:Hardy-Weinberg.svg

When is HWE is useful?

The **failure** of HWE can reveal a lot about sample features:

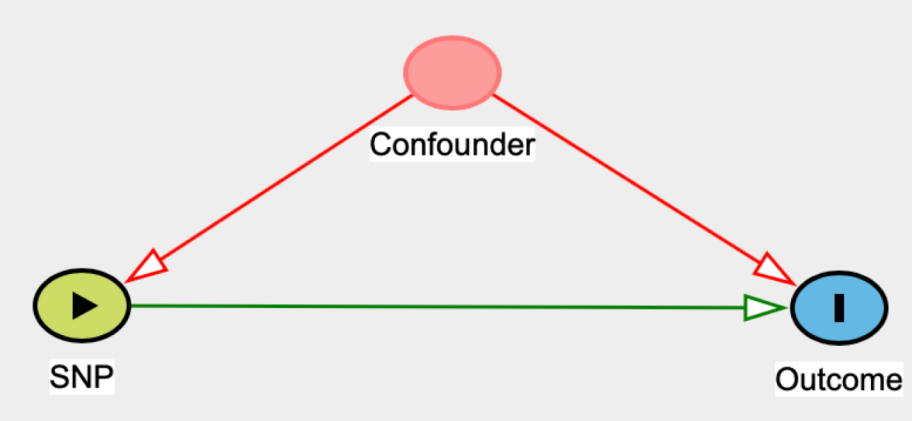
- Selection of subjects related to genotype
- Population Substructure
- Genotyping errors

- How to detect failure of HWE:
  - Estimate allele frequencies from genotypes
  - Compute **expected** genotype frequencies assuming HWE holds
  - Use Pearson Chi-Square test

Genotyping Errors & HWE

- Assumptions underlying Hardy–Weinberg equilibrium:
  - Random mating, large population (infinite), discrete generations, no inbreeding, no migration, no gene flow, no admixture, no mutation, no selection
- Assume that HWE should hold in the sample b/c
  - Genetic effect sizes are small
  - Majority of the SNPs are under the null-hypothesis
  - No selection
- Use HWE to check for genotyping errors
  - Remove all SNPs that have strong evidence of being out of HWE from the analysis

Confounding



- Confounding: major threat to the validity of non-randomized studies
- Classical definition confounding: A variable is a confounding variable if it satisfies two conditions:
  - It is a risk factor for the outcome
  - It is associated with exposure (SNP), but not a consequence of the exposure (not a mediator)
- Failure to control for confounding can lead to bias
- What is a potential confounding variable for a genetic study?
  - Population substructure

Population Substructure

- Features of a population which result in variation of allele and/or genotype frequencies across individuals in a population
  - https://www.genecards.org/cgi-bin/carddisp.pl?gene=PPARA

PPARA (Peroxisome Proliferator Activated Receptor Alpha) is a Protein Coding gene. Associated with Alcoholic Cardiomyopathy and Fatty Liver Disease.

MAF differs by ancestry

MAF(minor allele freq=0.05)

MAF=0.0006

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
Latino	649	11522	28	0.05633
East Asian	361	8618	8	0.04189
Other	10	904	0	0.01106
European (Finnish)	22	6606	1	0.00333
South Asian	42	16440	2	0.002555
European (Non-Finnish)	73	66602	0	0.001096
African	6	10294	0	0.0005829
Total	**1163**	**120986**	**39**	**0.009613**

Population Substructure: Population Stratification

Native American Heritage (with number of great- grandparents)	**$Gm^{3;5;13;14}$ %**	% Diabetes age adjusted
0	69%	18.5%
4	45%	28.6%
8	.01%	39.2%

Adapted from Knowler, 1988

- **$Gm^{3;5,13,14}$ allele frequency is a marker of Native American Heritage**
- Strong correlation between $Gm^{3;5,13,14}$ allele frequency across strata of Native American heritage (# of great- grandparents)
- Different degrees of Native American and European Hispanic ancestry in Gila River community
- GM allele lies on a locus of the human immunoglobulin G gene
- Failure to adjust for confounding by population stratification
  - Spurious inverse association between $Gm^{3,5,13,14}$ and non-insulin-dependent diabetes mellitus among residents of the Gila River Community

Population Substructure: What to do?

1. Match or stratify on ancestry
   - Self report may not be accurate
   - Ethnicity [i.e. “race”] may not be good surrogate for ancestry
   - Difficult to match admixed ancestry subjects
2. Genomic control
   - Adjust using multiple unlinked markers:
   - Estimate “inflation factor” and adjust accordingly
3. Population Components (PCA)
   - Infer population substructure using ‘continuous axes of variation’
   - MOST popular method
4. Linear mixed effect models
5. Use family-based controls
   - Siblings (conditional logistic)
   - Case-parent “pseudocontrols” (TDT, FBAT etc.)

Method 2: Genomic Control

- Statistical method that is used to control for the confounding effects of population stratification
  - Devlin and Roeder (1999)
- Idea: Most SNPs are not associated with the trait
- Consider the genomic inflation factor (lambda)
  - lambda is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median
    - Quantifies the extent of the bulk inflation and the excess false positive rate
  - Expresses the deviation of the distribution of the observed test statistic compared to the distribution of the expected test statistic

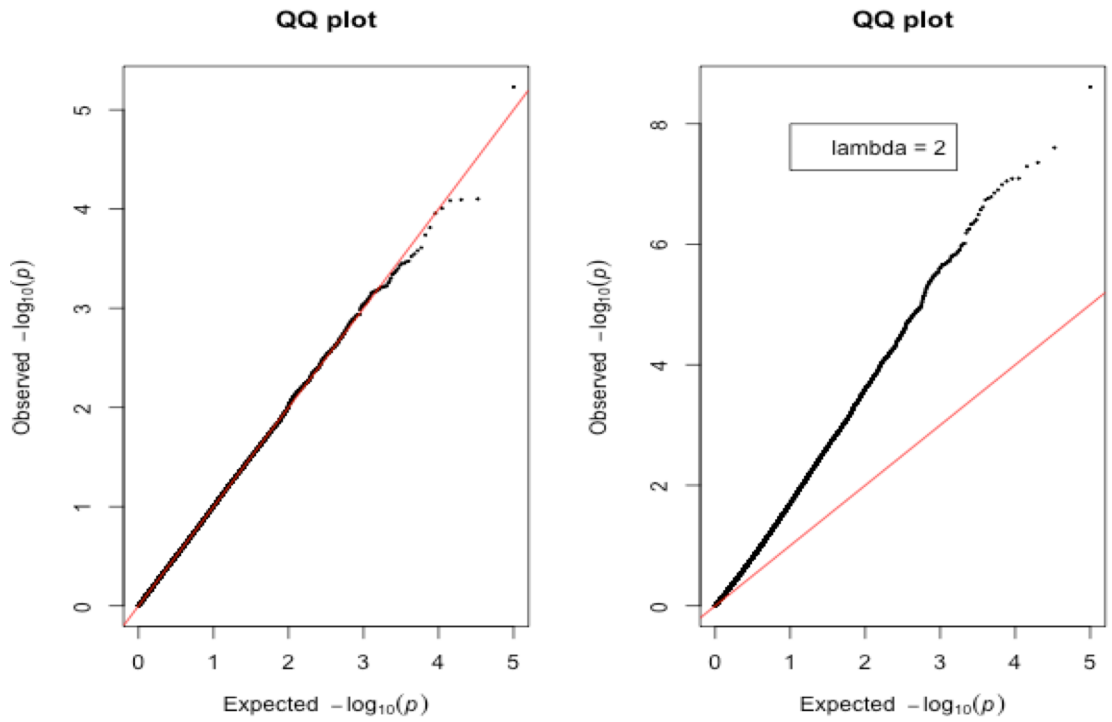
Method 2: Genomic Control

- High genomic inflation factors are caused by
  - Population stratification
  - Strong linkage disequilibrium (LD) between SNPs
  - Strong association between SNPs and phenotypes
  - Systematic bias
  - Other sources

- Then, adjust for lambda
  - May not completely adjust for population stratification
  - Not as popular of an approach
  - lambda used to check GWAS

How to detect genomic inflation?

- Genomic Inflation factor
  - Estimate *λ to check whether λ > 1*
  - *Basis for genomic control*
- Q-Q plots
  - p-values follow an uniform distribution between 0 and 1
  - Plot observed p-values vs quantiles of an uniform distribution



Method 3: Principle Components (PC)

- With a GWAS, get 500K or more markers; essentially enough to nail down an individual's genetic signature
  - Do not use imputed SNPs
- Use var-cov matrix of individual genetic variation, denoted by C
- Use PC to determine 'major axes' or continuous axes of genetic variation
- Adjust for population stratification by including the top PCs as covariates in the model
- PC explanation without equations below
  - https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/

GWAS Analysis: QQ Plots

- Compute p-values for each SNP
  - Running a regression (linear, logistic, etc) for each SNP
  - Under the null, the p-values follow a uniform distribution
- Q-Q plot: Compare the distribution of the observed p-values vs expected from the Uniform distribution
  - Inflation is indicative of population stratification or cryptic relatedness or other issues
  - The genomic inflation factor (λ) is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median
  - Quantifies the extent of the bulk inflation and the excess false positive rate ($\lambda < 1.1$)

Q-Q Plots

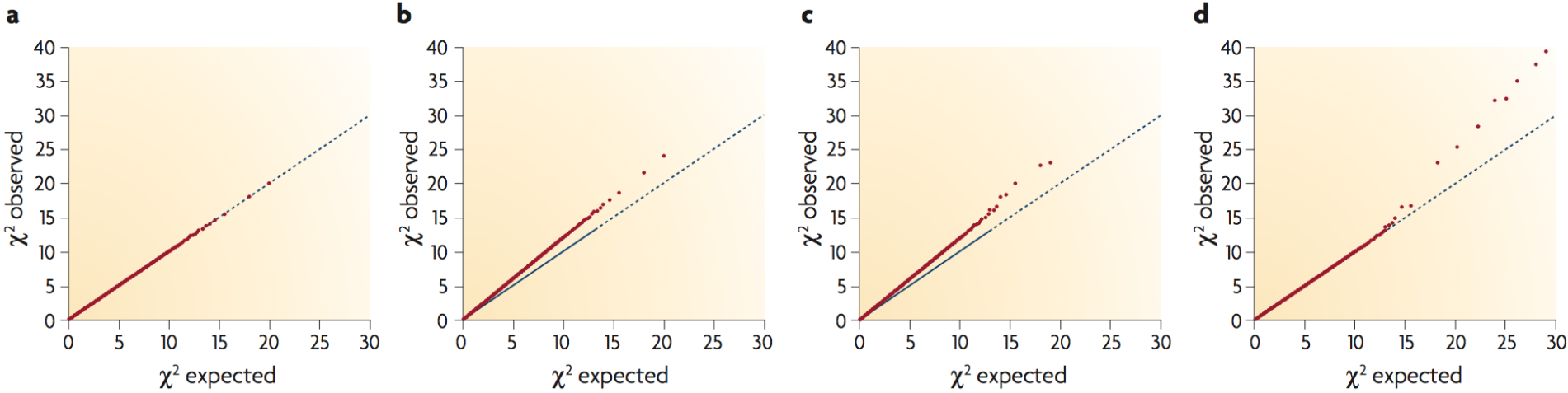
- Plot test statistics or -log10(p)

-Panel a, observed data conforms closely to expectation
Little evidence for association

-Panel b, inflation of the observed findings across the distribution
Indicative of population stratification or cryptic relatedness

-Panel c there is similar evidence of population substructure
But some suggestion of an excess of strong associations

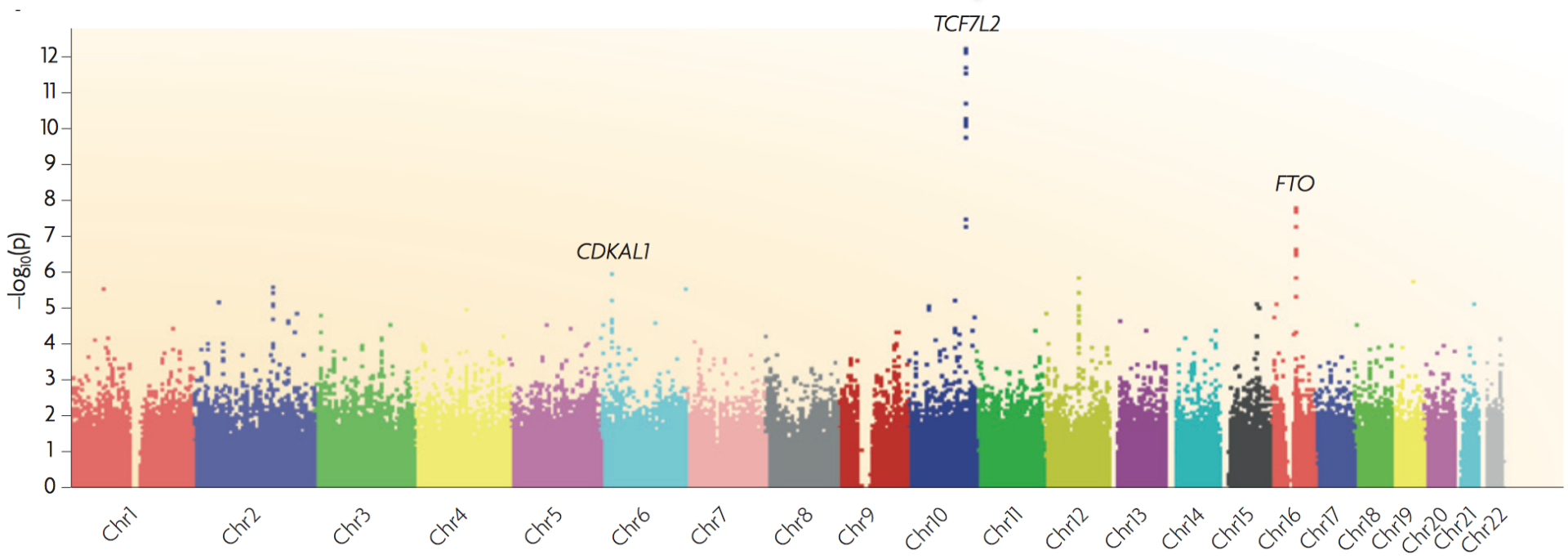
-Panel d there is little evidence of substructure
But compelling evidence for an excess of disease associations



McCarthy MI et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008 May;9(5):356-69

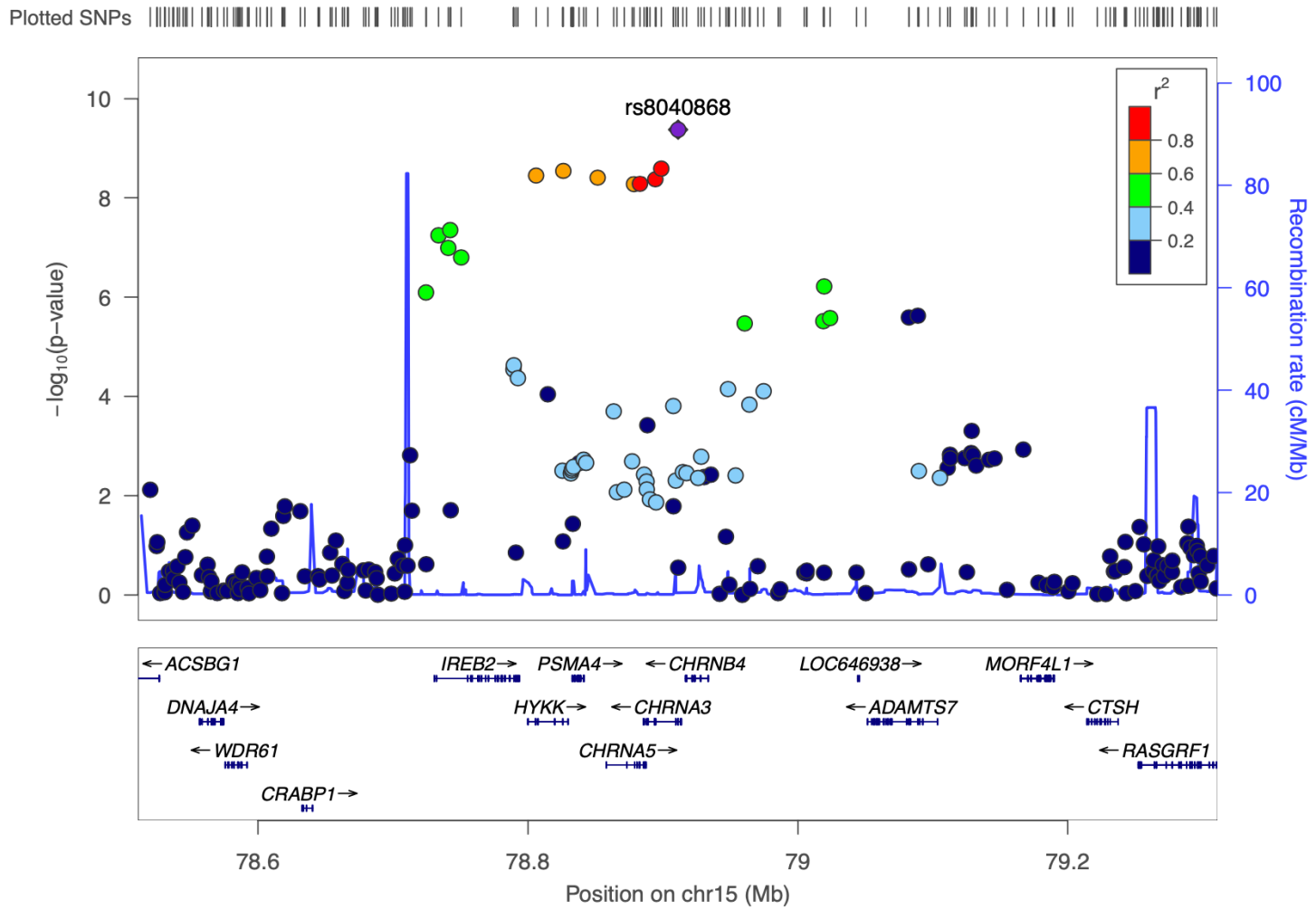
Manhattan Plots

- Visualize results and check for issues
- Due to LD, expect a few significant SNPs at a locus



Regional Plot

- Visualize results and check for issues
- Due to LD, expect SNPs in LD will lead SNP



GWAS: COPDGene Study

- Genetic Epidemiology of COPD (COPDGene)
  - Multi-center case-control study designed to identify genetic determinant of COPD and COPD-related phenotypes
- Recruited COPD cases and controls
  - Ancestry: European (~6,600) and African (~3,300)
  - Ages 45 to 80
  - At least 10 pack-years of smoking history
- Is this a normal population?
  - Careful with generalizations to larger population
  - Careful with selection bias

GWAS: pack-years

- Outcome: pack-years of cigarette smoking
- Adjust: age, sex, genetic ancestry (as summarized by principal components)

OXFORD JOURNALS Nicotine & Tobacco Research

Nicotine Tob Res. 2019 Jun; 21(6): 714–722. PMCID: PMC6528143
Published online 2018 May 15. doi: 10.1093/ntr/nty095 PMID: 29767774

Common and Rare Variants Genetic Association Analysis of Cigarettes per Day Among Ever-Smokers in Chronic Obstructive Pulmonary Disease Cases and Controls

Sharon M Lutz, PhD,[1] Brittni Frederiksen, PhD,[2] Ferdouse Begum, PhD,[3] Merry-Lynn N McDonald, PhD,[4] Michael H Cho, MD,[4,5] Brian D Hobbs, MD,[4,5] Margaret M Parker, PhD,[4] Dawn L DeMeo, MD,[4,5] Craig P Hersh, MD,[4,5] Marissa A Ehringer, PhD,[6] Kendra Young, PhD,[2] Lai Jiang, PhD,[3] Marilyn G Foreman, MD,[7] Greg L Kinney, PhD,[2] Barry J Make, MD,[8] David A Lomas, MD,[9] Per Bakke, PhD,[10] Amund Gulsvik, PhD,[10] James D Crapo, MD,[8] Edwin K Silverman, MD,[4,5] Terri H Beaty, PhD,[3] John E Hokanson, PhD,[2] and ECLIPSE and COPDGene Investigators

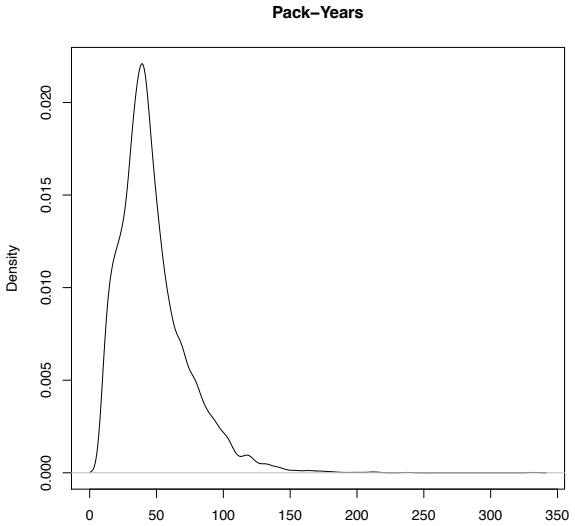
GWAS: Format Phenotype File

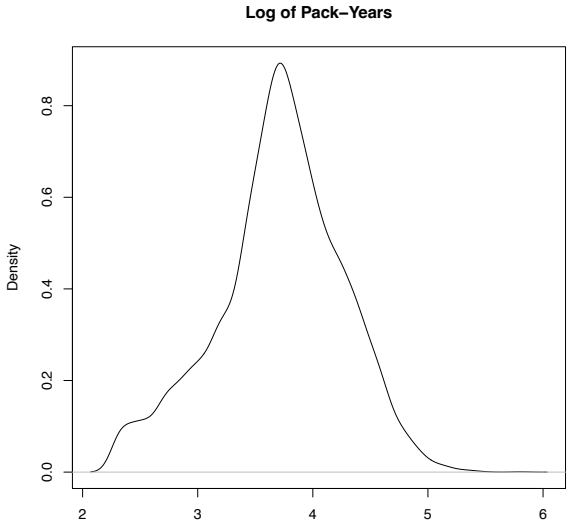
- Format phenotype file (col1-2: FID, IID)

```
# Load phenotype file
copdgene <-read.table("/Users/sharon/Data/Final10000_Dataset_12MAR13.txt", sep="\t",header=T,
 na.strings="")
names(copdgene)[1]<-"IID"

# Load PC file
pca = read.csv("/Users/sharon/Data/nhw-pcs.csv",header=T)
names(pca)[1]<-"IID"

# Merge files
d2<-merge(copdgene,pca,by="IID")

# create matrix of needed traits
matP<-
 cbind(d2[,1:2],d2[,1:2],d2[,"Age_Enroll"],d2[,"gender"],d2[,"ATS_PackYears"],log(d2[,"ATS_PackYears"])
 ,d2[,"PC1"],d2[,"PC2"],d2[,"PC3"],d2[,"PC4"],d2[,"PC5"])

matP2<-matP[,c(1,3:ncol(matP))]
colnames(matP2)<-c("FID","IID","center","age","sex","py","pyl","PC1","PC2","PC3","PC4","PC5")

# Output the phenotype file to use for plink
write.table(matP2,file="/Users/sharon/Data/copdgeneCIGnhw.txt",quote=F,row.names=F)

#display first 5 lines
matP2[1:5,]
     FID    IID center  age sex   py      pyl     PC1     PC2     PC3     PC4     PC5
1 10002K 10002K    BWH 63.6   2 45.0 3.806662  0.0051 -0.0059 -0.0159  0.0230  0.0143
2 10004O 10004O    BWH 68.7   1 26.4 3.273364 -0.0049 -0.0134  0.0022  0.0154 -0.0012
3 10005Q 10005Q    NJC 54.5   2 40.5 3.701302 -0.0066 -0.0032  0.0052 -0.0107  0.0098
4 10006S 10006S    BWH 62.4   2 52.0 3.951244 -0.0037 -0.0059 -0.0030 -0.0093  0.0051
5 10009Y 10009Y    NJC 69.1   1 66.0 4.189655 -0.0023  0.0074 -0.0041  0.0004 -0.0031
```

Check Model Fit

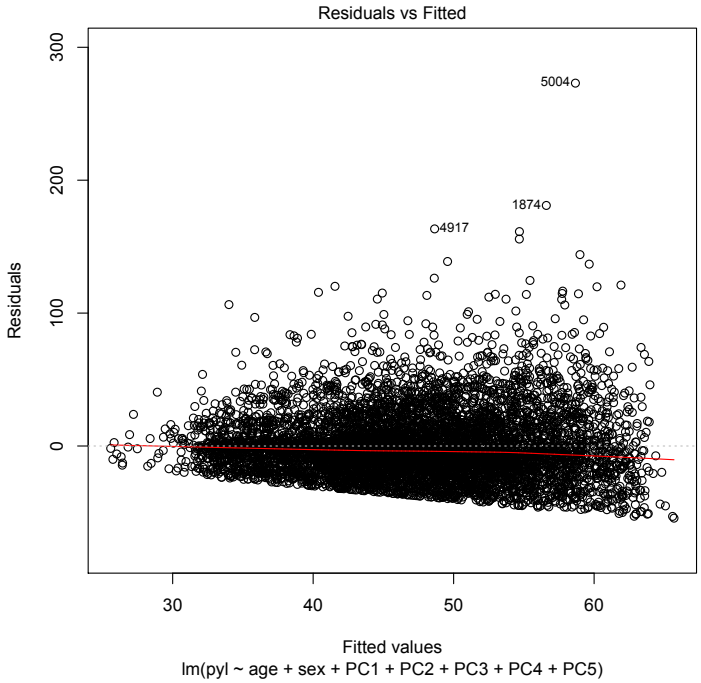
- Pack-years vs log of pack-years

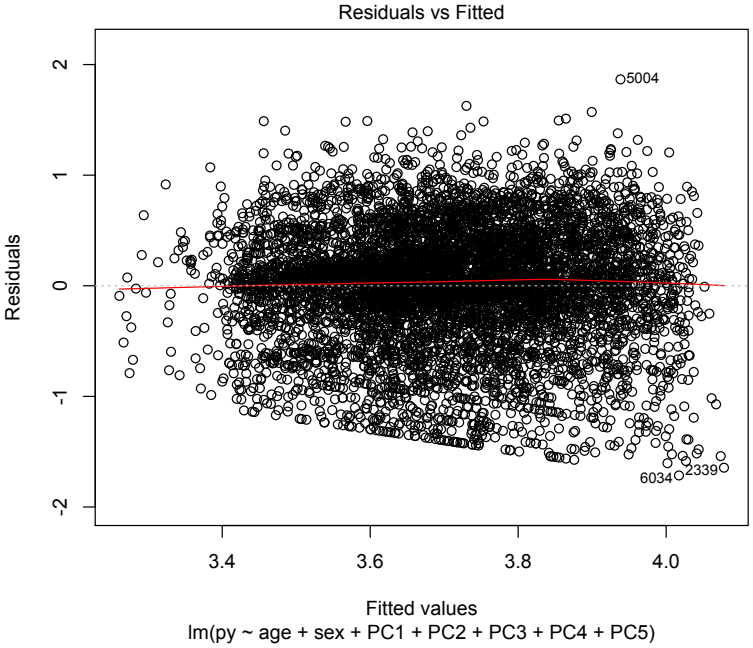


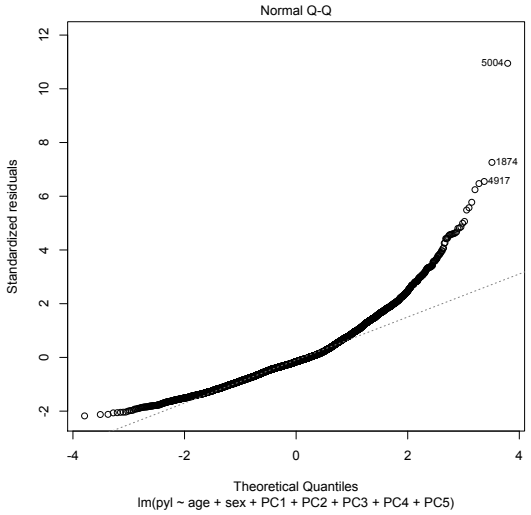


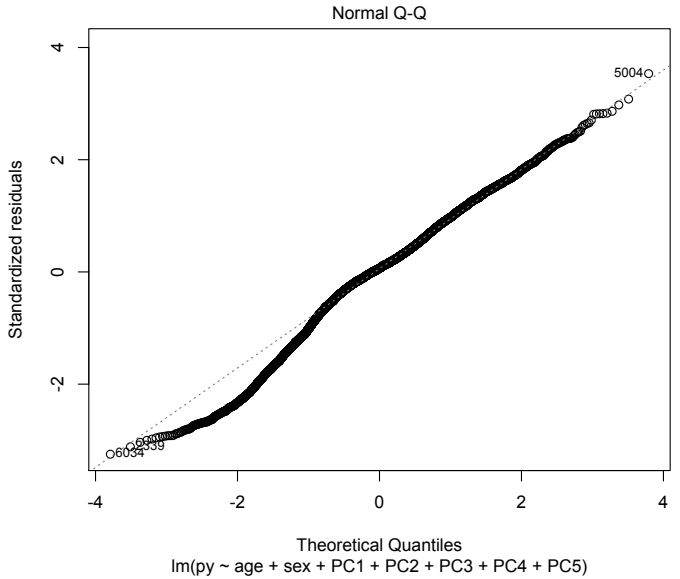

```
# check model fit
modelPY<-lm(py~age+sex+PC1+PC2+PC3+PC4+PC5,data=matP2)
summary(modelPY)
pdf("diagPY.pdf")
plot(modelPY)
dev.off()
pdf("denPY.pdf")
plot(density(d2[,"ATS_PackYears"],na.rm=T),main="Pack-Years",xlab="")
dev.off()

modelPYL<-lm(pyl~age+sex+PC1+PC2+PC3+PC4+PC5,data=matP2)
summary(modelPYL)
pdf("diagPYL.pdf")
plot(modelPYL)
dev.off()
pdf("denPY.pdf")
plot(density(log(d2[,"ATS_PackYears"]),na.rm=T),main="Log of Pack-Years",xlab="")
dev.off()
```

Pack-Years vs Log Pack-Years









Plink File Format

- Either text-format files or binary files
- Recommend binary files
- Because reading large text files can be time consuming
- Text PLINK data consist of two files:
  - (1) Information on the individuals and genotypes (*.ped)
  - (2) Information on the genetic markers (*.map)
- Binary PLINK data consist of three files:
  - (1) Binary file that contains individual identifiers (IDs) and genotypes (*.bed)
  - (2) 2 text files that contain information on the individuals (*.fam) and on the genetic markers (*.bim)

Plink File Format

- Binary outcome: 1/2 not 0/1
- *.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

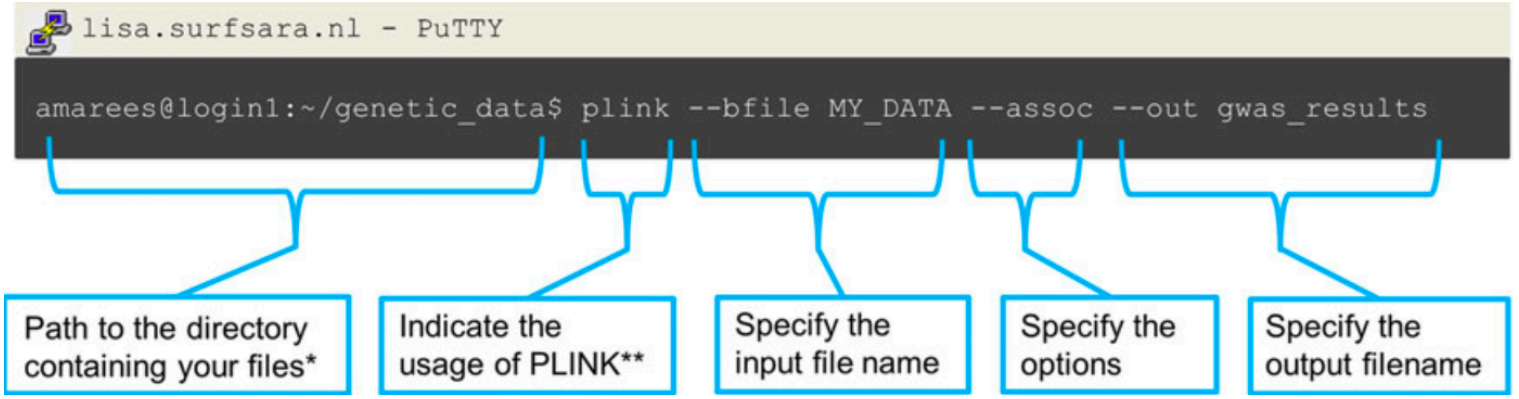
FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

Marees AT, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27:e1608.

Plink: Input

- --linear or --logistic for linear or logistic regression
- --maf includes only SNPs above MAF threshold
- --hwe excludes markers which deviate from Hardy– Weinberg equilibrium
- Recommend setting a low threshold
  - Serious genotyping errors often yield extreme p-values (1e-50)
  - Genuine SNP-trait associations can be expected to deviate slightly from Hardy-Weinberg equilibrium
  - Dangerous: threshold that filters out too many variants
- http://zzz.bwh.harvard.edu/plink/thresh.shtml



```
Sharons-MacBook-Pro:~ sharon$ cd
Sharons-MacBook-Pro:~ sharon$ pwd
/Users/sharon
Sharons-MacBook-Pro:~ sharon$ ./plink --noweb --bfile   /Users/sharon/Data/CG10kNhwHg19Clean_v2_Mar2
013 --maf 0.01 --hwe 0.001  --linear --pheno  /Users/sharon/Data/copdgeneCIGnhw.txt --pheno-name pyl
  --covar /Users/sharon/Data/copdgeneCIGnhw.txt  --covar-name age,sex,PC1,PC2,PC3,PC4,PC5 --adjust
--out  /Users/sharon/Data/copdgenePYLnhw
PLINK v1.90b6.8 64-bit (15 Feb 2019)           www.cog-genomics.org/plink/1.9/
(C) 2005-2019 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to /Users/sharon/Data/copdgenePYLnhw.log.
Options in effect:
  --adjust
  --bfile /Users/sharon/Data/CG10kNhwHg19Clean_v2_Mar2013
  --covar /Users/sharon/Data/copdgeneCIGnhw.txt
  --covar-name age,sex,PC1,PC2,PC3,PC4,PC5
  --hwe 0.001
  --linear
  --maf 0.01
  --noweb
  --out /Users/sharon/Data/copdgenePYLnhw
  --pheno /Users/sharon/Data/copdgeneCIGnhw.txt
  --pheno-name pyl

Note: --noweb has no effect since no web check is implemented yet.
16384 MB RAM detected; reserving 8192 MB for main workspace.
630860 variants loaded from .bim file.
6670 people (3493 males, 3177 females) loaded from .fam.
6670 phenotype values present after --pheno.
Using 1 thread (no multithreaded calculations invoked).
--covar: 7 out of 10 covariates loaded.
Before main variant filters, 6670 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.998687.
--hwe: 3027 variants removed due to Hardy-Weinberg exact test.
244 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
627589 variants and 6670 people pass filters and QC.
Phenotype data is quantitative.
Writing linear model association results to
/Users/sharon/Data/copdgenePYLnhw.assoc.linear ... done.
--adjust: Genomic inflation est. lambda (based on median chisq) = 1.00722.
--adjust values (627589 variants) written to
/Users/sharon/Data/copdgenePYLnhw.assoc.linear.adjusted .
```

Note: lambda is close to 1 which is good. There does not appear to be issues due to population substructure.

Check Results: Command Line

- 9 SNPs reach genome wide significance (p-value <5*10^-8)

```
Sharons-MacBook-Pro:~ sharon$ head -11 /Users/sharon/Data/copdgenePYLnhw.assoc.linear.adjusted
 CHR         SNP      UNADJ         GC       BONF       HOLM   SIDAK_SS   SIDAK_SD     FDR_BH     FDR_BY
  15   rs8040868  4.193e-10  4.838e-10  0.0002632  0.0002632  0.0002631  0.0002631  0.0002632   0.003665
  15  rs12914385  2.556e-09  2.912e-09   0.001604   0.001604   0.001603   0.001603  0.0004152   0.005783
  15    rs931794  2.849e-09  3.244e-09   0.001788   0.001788   0.001787   0.001787  0.0004152   0.005783
  15   rs8034191   3.54e-09  4.023e-09   0.002221   0.002221   0.002219   0.002219  0.0004152   0.005783
  15   rs2036527  3.908e-09   4.44e-09   0.002453   0.002453    0.00245    0.00245  0.0004152   0.005783
  15   rs1051730  4.222e-09  4.793e-09   0.002649   0.002649   0.002646   0.002646  0.0004152   0.005783
  15  rs16969968  5.205e-09    5.9e-09   0.003266   0.003266   0.003261   0.003261  0.0004152   0.005783
  15    rs951266  5.293e-09  5.999e-09   0.003322   0.003322   0.003316   0.003316  0.0004152   0.005783
  15  rs17483929  4.466e-08  4.987e-08    0.02803    0.02803    0.02764    0.02764   0.003114    0.04337
  15  rs17483721  5.692e-08  6.347e-08    0.03572    0.03572    0.03509    0.03509   0.003572    0.04975
```

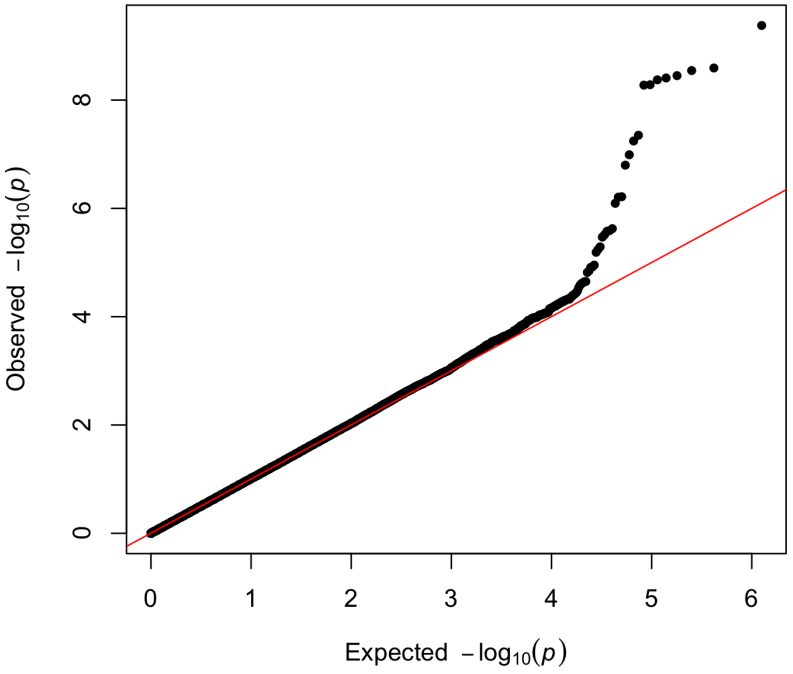
- Check SNPs: rs8040868 CHRNA3 chr15:78618839
- https://wakegen.phs.wakehealth.edu
- https://www.ncbi.nlm.nih.gov/snp/
- Multiple testing
  - α = 0.05 doesn't work for 500,000 or more tests
  - Multiple ways to correct for multiple testing:
    - Bonferonni, Holm, Sidak, etc (Next lecture)

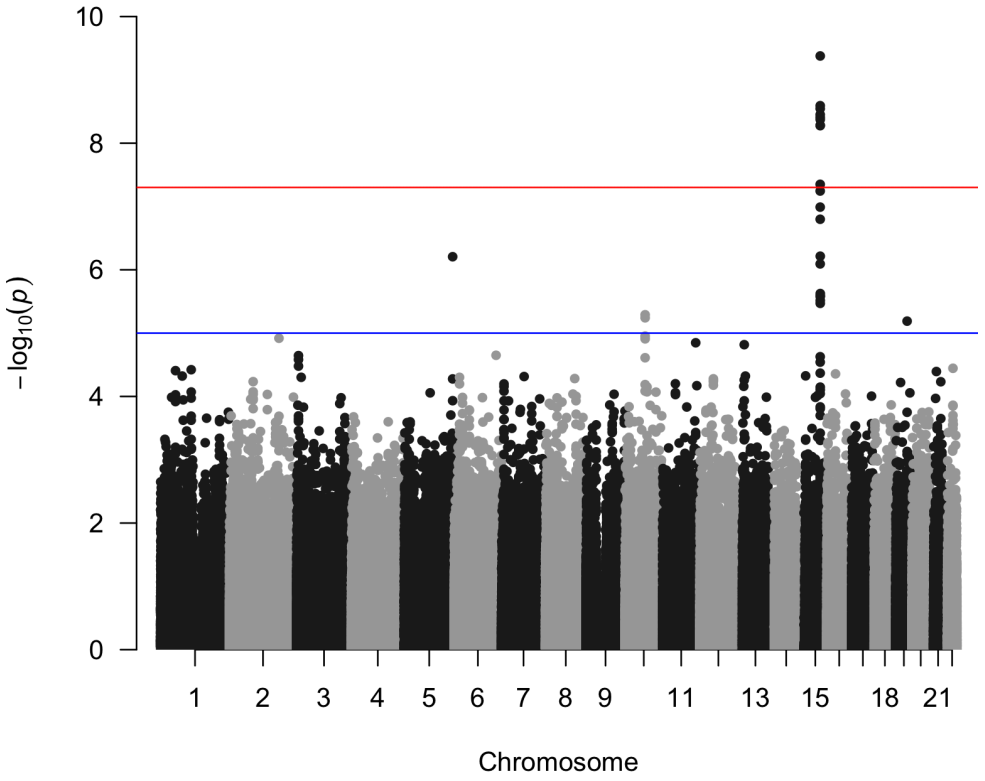
Plot Results: QQ & Manhattan

```
library(qqman)

py<-read.table("/Users/sharon/Data/copdgenePYLnhw.assoc.linear.adjusted",header=T)
dim(py) #[1] 627589     10
py[1:2,]
#  CHR        SNP     UNADJ        GC      BONF      HOLM  SIDAK_SS  SIDAK_SD    FDR_BH    FDR_BY
#1  15  rs8040868 4.193e-10 4.838e-10 0.0002632 0.0002632 0.0002631 0.0002631 0.0002632 0.003665
#2  15 rs12914385 2.556e-09 2.912e-09 0.0016040 0.0016040 0.0016030 0.0016030 0.0004152 0.005783

pdf("qqPY.pdf")
qq(py$UNADJ)
dev.off()

pyr<-read.table("/Users/sharon/Data/copdgenePYLnhw.assoc.linear",header=T)
dim(pyr)
#[1] 5020712       9
pyr[1:10,]
#   CHR        SNP     BP A1 TEST NMISS      BETA     STAT         P
#1    1 rs12562034 768448  A  ADD  6538 -0.007074  -0.4874 6.260e-01
#2    1 rs12562034 768448  A  age  6538  0.012280  16.5500 2.620e-60
#3    1 rs12562034 768448  A  sex  6538 -0.179500 -13.7000 3.827e-42
#4    1 rs12562034 768448  A  PC1  6538 -3.425000  -6.4020 1.642e-10
#5    1 rs12562034 768448  A  PC2  6538 -1.382000  -2.5900 9.628e-03
#6    1 rs12562034 768448  A  PC3  6538  0.699100   1.3090 1.905e-01
#7    1 rs12562034 768448  A  PC4  6538  0.082980   0.1552 8.767e-01
#8    1 rs12562034 768448  A  PC5  6538  0.243900   0.4576 6.472e-01
#9    1 rs12124819 776546  G  ADD  6593  0.003586   0.3460 7.294e-01
#10   1 rs12124819 776546  G  age  6593  0.012240  16.5600 2.145e-60
ss<-seq(from=1,to=nrow(pyr),by=8)
pyr2<-pyr[ss,]
pyr2<-na.omit(pyr2)
#   CHR        SNP     BP A1 TEST NMISS      BETA    STAT      P
#1    1 rs12562034 768448  A  ADD  6538 -0.007074 -0.4874 0.6260
#9    1 rs12124819 776546  G  ADD  6593  0.003586  0.3460 0.7294
#17   1 rs11240777 798959  A  ADD  6670  0.006593  0.5989 0.5492
#25   1  rs4970383 838555  A  ADD  6643 -0.012090 -1.1290 0.2588
#33   1  rs4475691 846808  T  ADD  6668 -0.003065 -0.2638 0.7920
#41   1  rs7537756 854250  G  ADD  6668 -0.004123 -0.3626 0.7169
#49   1 rs13302982 861808  A  ADD  6664  0.020390  0.7271 0.4672
#57   1  rs1110052 873558  G  ADD  6643  0.003368  0.3381 0.7353
#65   1  rs2272756 882033  A  ADD  6668 -0.003593 -0.3369 0.7362
#73   1  rs3748597 888659  T  ADD  6669  0.007407  0.3691 0.7121

manhattan(pyr2, chr = "CHR", bp = "BP", p = "P", snp = "SNP")
```

Just want TEST==ADD -SNPs

Plot Results: QQ & Manhattan

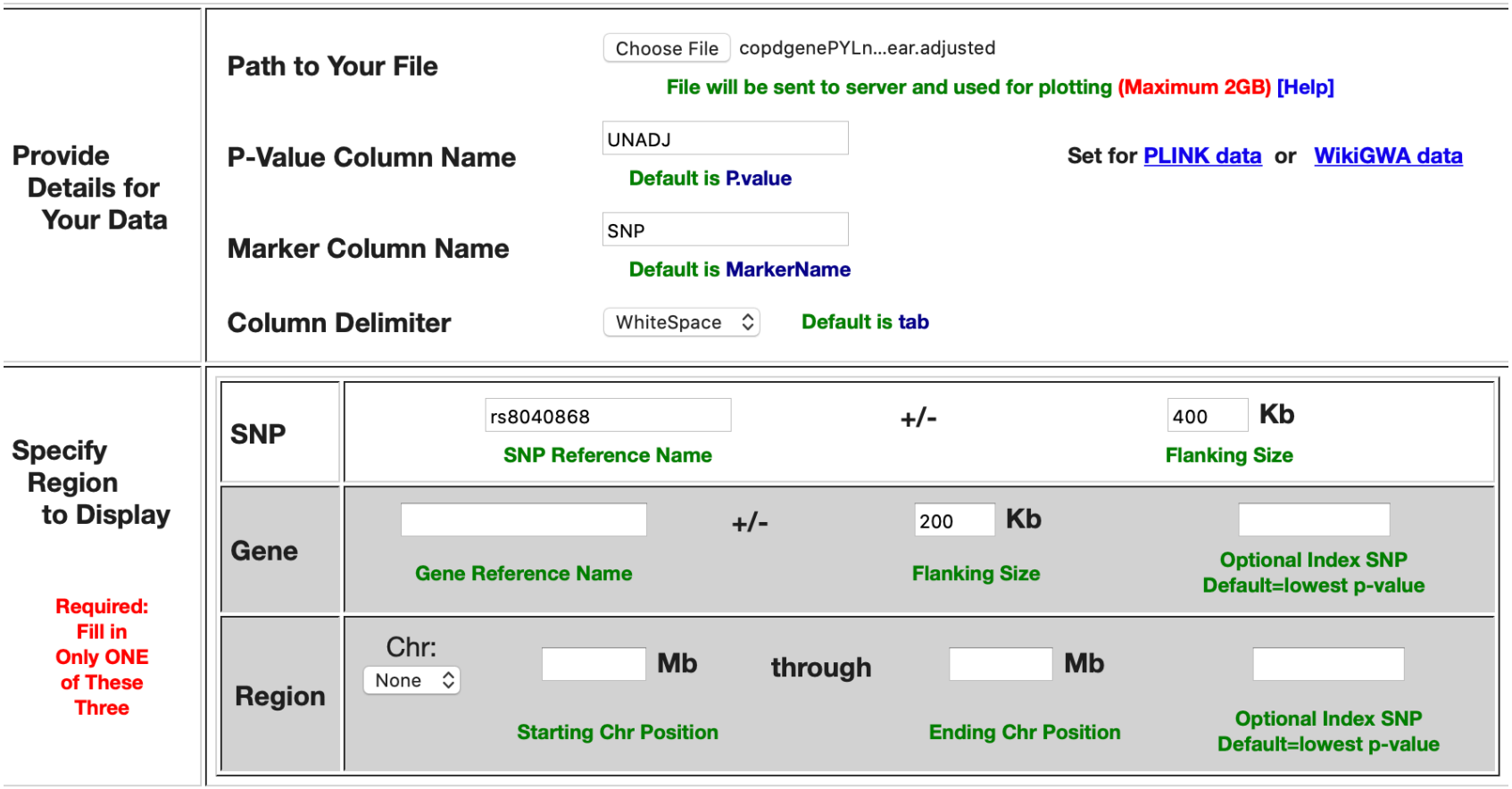




- QQ plot looks good- shows there are some significant hits but probably not issues with population stratification (remember: lambda =1.007)
- Manhattan plot shows that there is one region on chromosome 15 associated with pack-years

Plot Results: Regional Plots

- http://locuszoom.sph.umich.edu/locuszoomjs.php
- Make sure that you are using the right reference panel



Regional Plot

- Looks good: Others SNPs in LD