Bioinformatics Nanocourse
Genome-Wide Association Studies

He Zhang
He.Zhang@UTSouthwestern.edu
Bioinformatics Core Facility
4/27/2017

- Part I - GWAS study design
- Part II - Population stratification
- Part III - Genetic relationship

The principal goals of design for association studies

- Minimize systematic bias
  - If a marker is truly unassociated with a trait, tests of association should not reject the null hypothesis of no association any more than expected

- Maximize power
  - If a marker is truly associated with a trait, tests should have a good chance to reject the null hypothesis

Two primary classes of phenotypes

- Case/control trait
  - Case group affected by a disease
  - Healthy control group
  - Coronary artery disease, type II diabetes, Crohn's disease

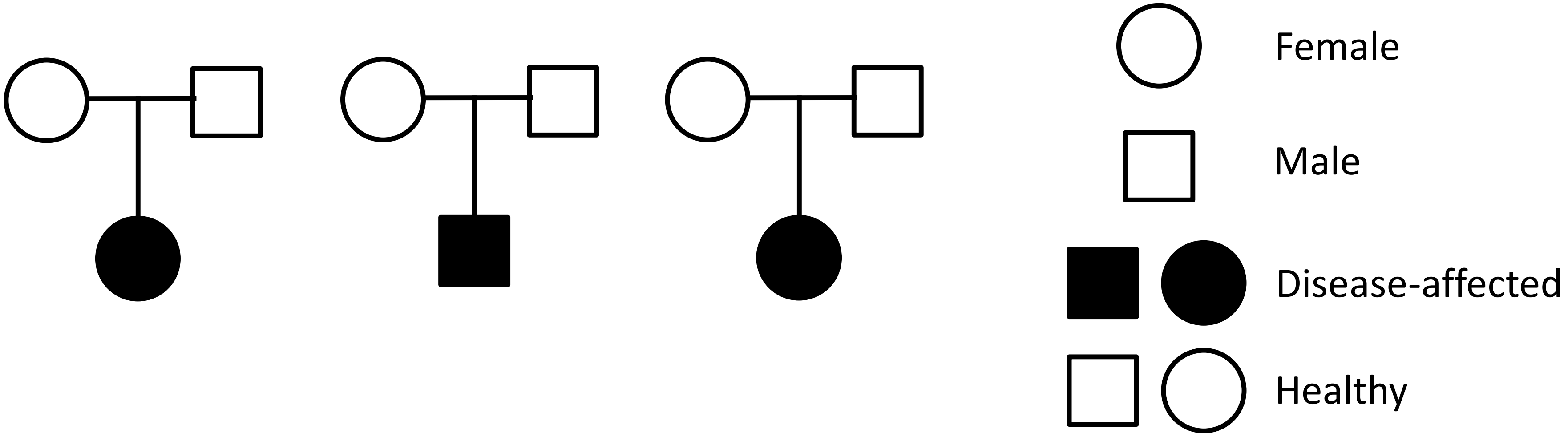
- Quantitative trait
  - Continuous value
  - Body mass index (BMI), Plasma high-density lipoproteins (HDL) level, blood pressure

Population-based design

- Case and controls are unrelated
- Susceptible to population stratification bias
- Easier to collect

Family-based design

- Cases and controls are related: parents, sibs etc
  - Commonly used design: case-parent trios
- Not susceptible to population stratification bias
- Not easy to collect
- Not appropriate for late-onset diseases



Case selection

- Improve study power through enrichment for specific disease-predisposing alleles
- Minimize phenotypic heterogeneity
  - Trait or disease was defined clearly
  - Medical diagnosis for cases were clear
  - You may focus on extreme (early age of onset) and/or familial cases

Control selection

- Controls should be selected from the same populations with cases

- Controls should also have clear diagnoses of the disease, if possibly

Systematic bias – population stratification

- Population stratification bias - cases and controls are not from the same population

- The genetic and environmental backgrounds for cases and controls may differ simply as a result of selection bias

- You should be careful when you use controls who were recruited and genotyped for a previous study

Systematic bias – relatedness

- If subjects are closely related, then their genotypes will be correlated, and the usual test statistics (which assume independence) will be inflated

- This is particularly a concern when only cases with a positive family history of disease are enrolled

Systematic bias – other selection bias

- Age and sex can also be confounders if the genotype frequency in the source population varies with age and gender

- A gene may be associated with a known behavioral risk factor (e.g., smoking, alcohol use) which will increase the risk of a disease

Systematic bias - batch effect

- Bias may be caused by the differences between cases and controls in DNA collection, storage, and genotyping methods
  - Samples of cases and controls were prepared by different people
  - Different kits or protocols were used for cases and controls
  - Cases and controls were genotyped in different batches

How to minimize systematic bias?

- Select samples with comprehensive medical records
- Select controls that matched with cases in age, sex, ethnicity, and cofounding behavioral factors
- Avoid using close-relatives if you don’t conduct a family-based study.
- Process a case and its matched control (if possible) in the same batch during DNA collection, storage, and genotyping.
- Adjust for possible confounding factors in association test.

Statistical power

- The power in a hypothesis test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true
- H0: The variant was not associated with trait
- H1: The variant was associated with trait

Power is determined by many factors

- Disease prevalence
- Risk-allele frequency
- Genotype relative risk (odds ratio/effect size)
- Number of cases
- Number of controls
- Significance level
  - Determined by number of markers tested

Multi-stage designs

- A typical two-stage design
  - First stage
    - Identify potential disease-associated variants
  - Second stage
    - a subset of variants were retyped in additional samples

- Multi-stage designs have been seen as an effective way of retaining power while reducing genotyping costs.

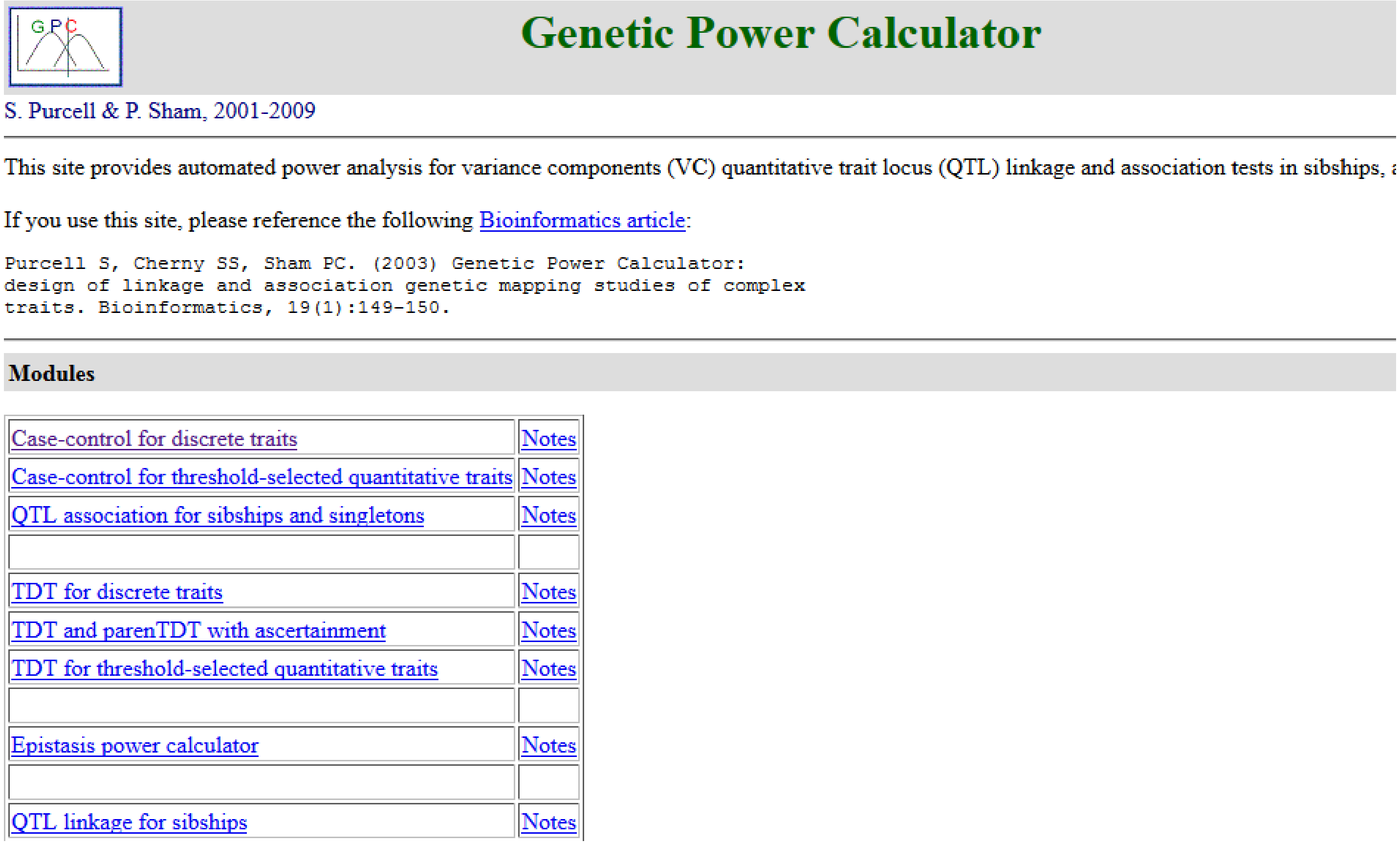
Multi-stage designs

- The substantial price differential between commodity and custom genotyping means that those cost benefits can be less dramatic than comparisons of genotype numbers alone would suggest.

- Winner’s curse effect
  - The original study will typically overestimate the true effect size.

Power estimation

- Genetic Power Calculator
  - http://pngu.mgh.harvard.edu/~purcell/gpc/



Genetic Power Calculator

S. Purcell & P. Sham, 2001-2009

This site provides automated power analysis for variance components (VC) quantitative trait locus (QTL) linkage and association tests in sibships,

If you use this site, please reference the following Bioinformatics article:

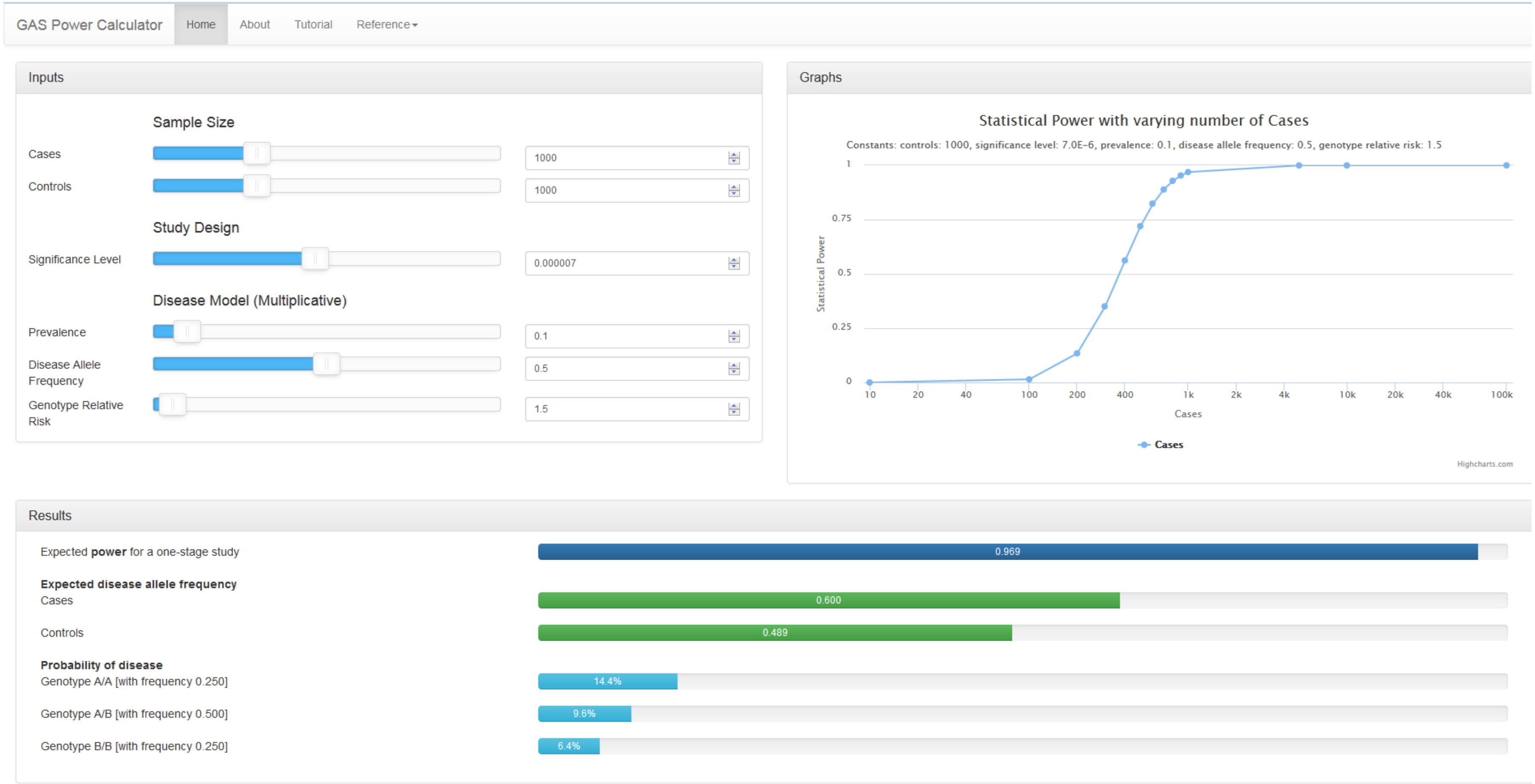
```
Purcell S, Cherny SS, Sham PC. (2003) Genetic Power Calculator:
design of linkage and association genetic mapping studies of complex
traits. Bioinformatics, 19(1):149-150.
```

Modules

Case-control for discrete traits	Notes
Case-control for threshold-selected quantitative traits	Notes
QTL association for sibships and singletons	Notes
TDT for discrete traits	Notes
TDT and parenTDT with ascertainment	Notes
TDT for threshold-selected quantitative traits	Notes
Epistasis power calculator	Notes
QTL linkage for sibships	Notes

Power estimation

- CaTS - Power Calculator for Two Stage Association Studies
  - http://csg.sph.umich.edu/abecasis/cats/


GAS Power Calculator
Home
About
Tutorial
Reference ▾
Graphs
Statistical Power with varying number of Cases
Constants: controls: 1000, significance level: 7.0E-6, prevalence: 0.1, disease allele frequency: 0.5, genotype relative risk: 1.5
Statistical Power
1
0.75
0.5
0.25
0
10
20
40
100
200
400
1k
2k
4k
10k
20k
40k
100k
Cases
● Cases
Highcharts.com

Inputs

Sample Size

Cases: 1000

Controls: 1000

Study Design

Significance Level: 0.000007

Disease Model (Multiplicative)

Prevalence: 0.1

Disease Allele Frequency: 0.5

Genotype Relative Risk: 1.5

Results

Expected **power** for a one-stage study: 0.969

Expected disease allele frequency

Cases: 0.600

Controls: 0.489

Probability of disease

Genotype A/A [with frequency 0.250]: 14.4%

Genotype A/B [with frequency 0.500]: 9.6%

Genotype B/B [with frequency 0.250]: 6.4%

Genotyping methods

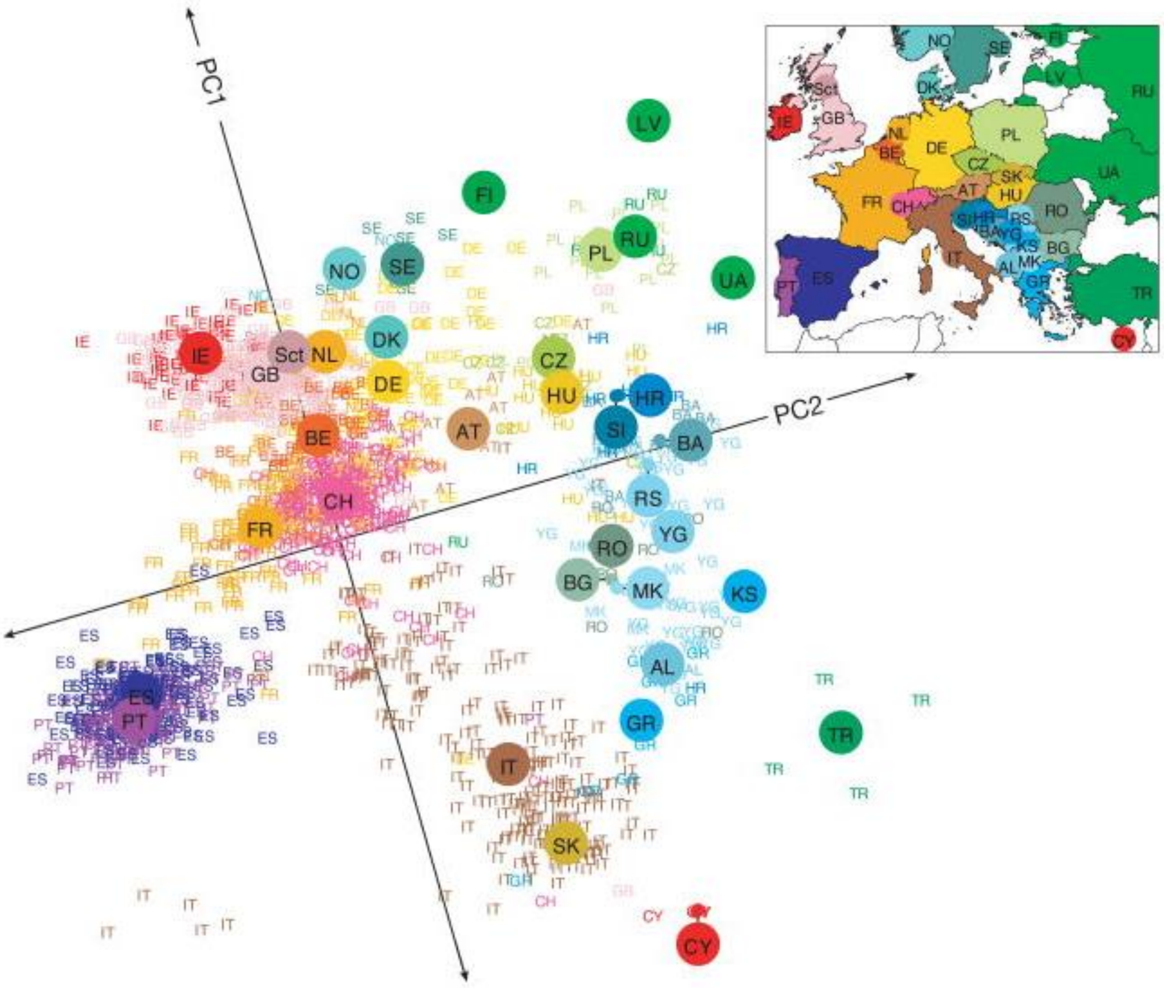
- Array
  - Up to several million markers
  - Only be able to detect markers in array
  - Cheaper than Sequencing
  - whole-genome, exome, CNV, custom markers, and etc

- Sequencing
  - More comprehensive catalog of variants
  - Be able to discover novel variants
  - Expensive
  - Whole-genome, exome, SV-seq, targeted sequencing, and etc

Quick summary

- Different designs for GWAS
- Sample selection
- Systematic bias
- Power estimation
- Genotyping methods

Part II - Population stratification



Genes mirror geography within Europe (Novembre et al, Nature, 2008)

What is population stratification

- Systematic difference in allele frequencies between subpopulations in a population possibly due to different ancestry rather than association of genes with the phenotype

- The cause of population stratification is nonrandom mating between groups
  - Physical separation : e.g. African and European
  - Mating based on proximity or culture

Population stratification may be a problem for GWAS

- If allele frequency vary between populations and disease prevalence also differs, association studies can produce misleading results

- Confounding
  - Higher chance of false positive association findings

- Reduced Power
  - Lower chance of detecting true effects

Genetics of chopstick use

Chinese, n = 2000

$\chi^2 = 0$, $P = 1$

	Use of chopsticks			
Allele	Yes	No	Total	
A1	900	100	1000	90%
A2	900	100	1000	90%
Total	1800	200	2000	

Genetics of chopstick use

Chinese, n = 2000
$\chi^2 = 0$, $P = 1$

	Use of chopsticks			
Allele	Yes	No	Total	
A1	900	100	1000	90%
A2	900	100	1000	90%
Total	1800	200	2000	

European, n = 2000
$\chi^2 = 0$, $P = 1$

	Use of chopsticks			
Allele	Yes	No	Total	
A1	180	1620	1800	10%
A2	20	180	200	10%
Total	200	1800	2000	

Genetics of chopstick use

Chinese, n = 2000
$\chi^2 = 0$, $P = 1$

	Use of chopsticks			
Allele	Yes	No	Total	
A1	900	100	1000	90%
A2	900	100	1000	90%
Total	1800	200	2000	

European, n = 2000
$\chi^2 = 0$, $P = 1$

	Use of chopsticks			
Allele	Yes	No	Total	
A1	180	1620	1800	10%
A2	20	180	200	10%
Total	200	1800	2000	

Chinese + European, n = 4000
$\chi^2 = 486$, $P = 10^{-107}$

	Use of chopsticks			
Allele	Yes	No	Total	
A1	1080	1720	2800	39%
A2	920	280	1200	77%
Total	2000	2000	4000	

How to identify potential population stratification?

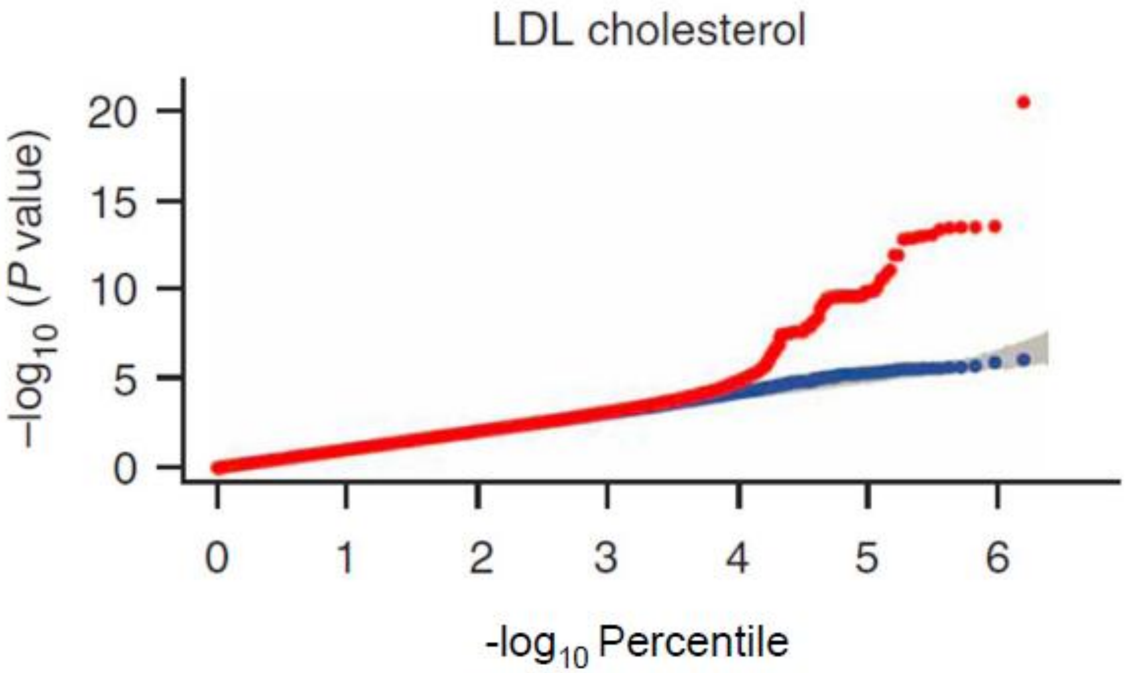
- The quantile-quantile (Q-Q) plot is an easy way to assess potential confounding factors, including population stratification

- Principal Components Analysis (PCA) and MultiDimensional Scaling (MDS) are the most commonly methods to infer population stratification

Q-Q plot : A useful diagnostic

- QQ plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- Population stratification may show as inflation

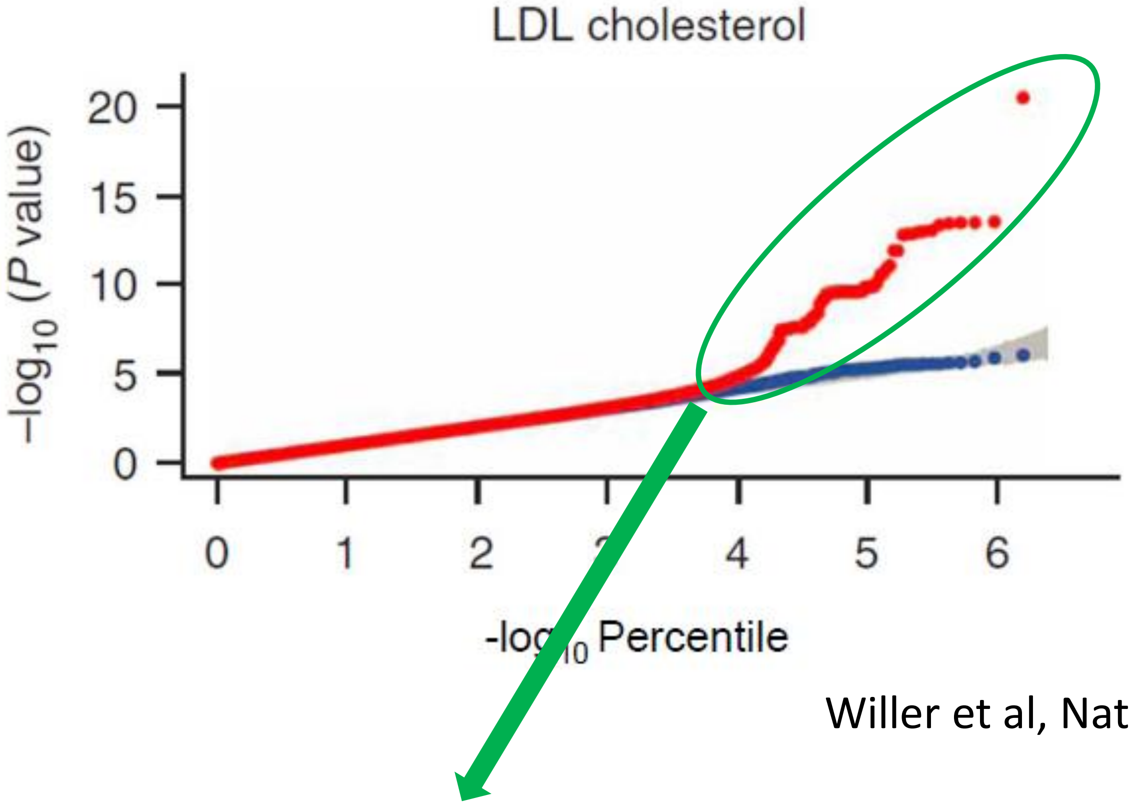
Q-Q plot : A useful diagnostic



Willer et al, Nature Genetics, 2008

Comparison of expected and observed p-values in a study of LDL cholesterol for all markers (red) and for markers in regions not known to impact LDL levels (blue)

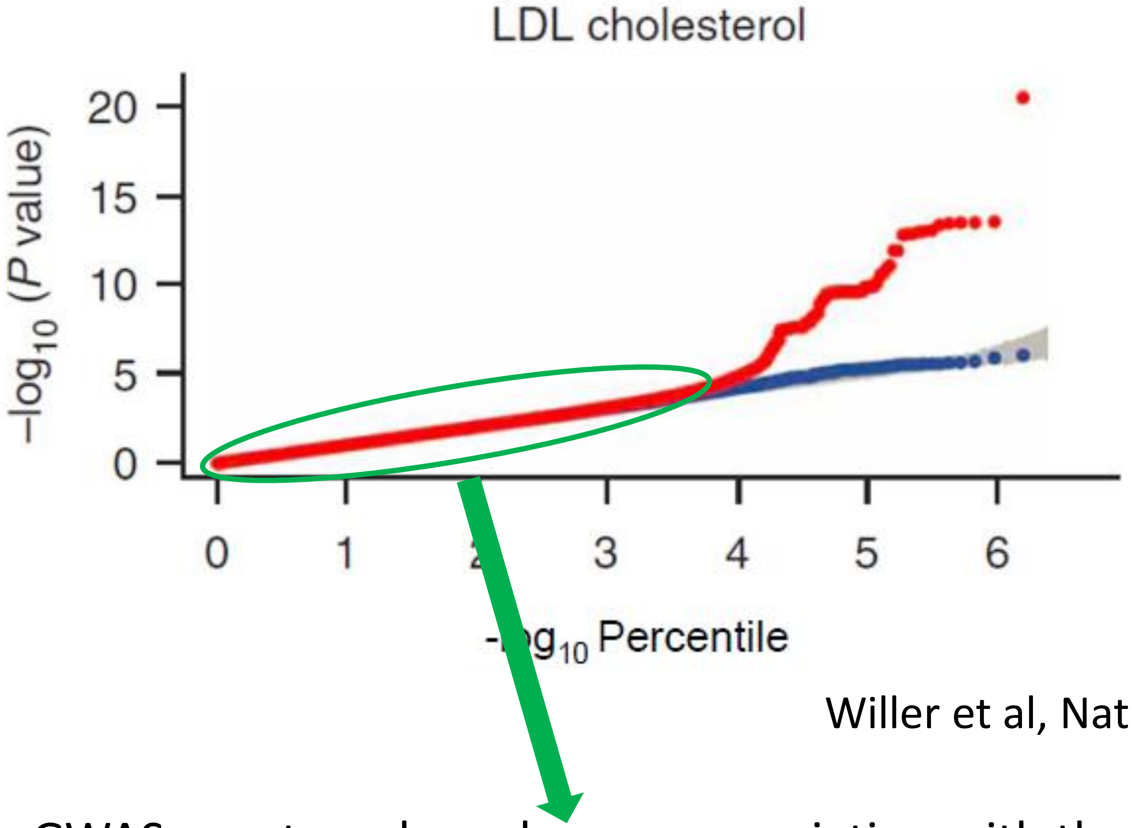
Q-Q plot : A useful diagnostic



Willer et al, Nature Genetics, 2008

In GWAS, only a small subset of markers are expected to show association with any particular trait.

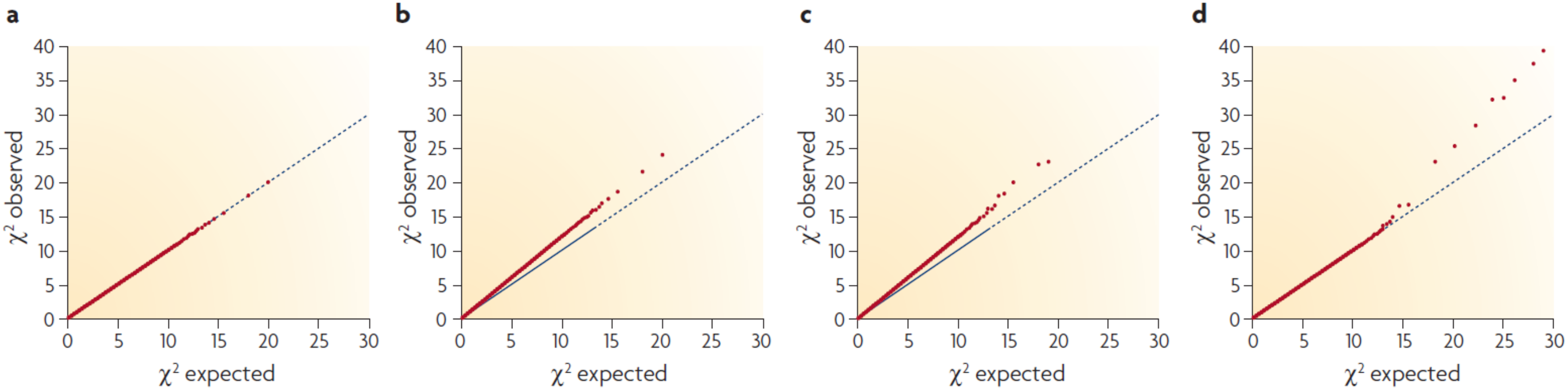
Q-Q plot : A useful diagnostic



Willer et al, Nature Genetics, 2008

In GWAS, most markers show no association with the trait and, therefore, very similar observed and expected p-values

Q-Q plot : A useful diagnostic



a. the observed data conforms closely to expectation little evidence for association.
b. inflation of the observed findings across the distribution is seen, indicative of population stratification or cryptic relatedness.
c. there is similar evidence of population substructure, but some suggestion of an excess of strong associations
d. there is little evidence of substructure, but compelling evidence for an excess of disease associations

McCarthy et al, Nat Rev Genet, 2008

Principal Components Analysis (PCA)

- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

- First few PCs may explain a large proportion of variance
  - The first PC has the largest possible variance
  - The second PC has the second largest possible variance
  - …

- PCA is useful for dimension reduction

PCA for genotype data

	SNP1	SNP2	SNP3	…	SNPn
Individual 1	0	1	2	…	1
Individual 2	1	0	0	…	0
Individual 3	0	0	0	…	0
…	…	…	…	…	…
Individual m	2	0	1	…	0

PCA

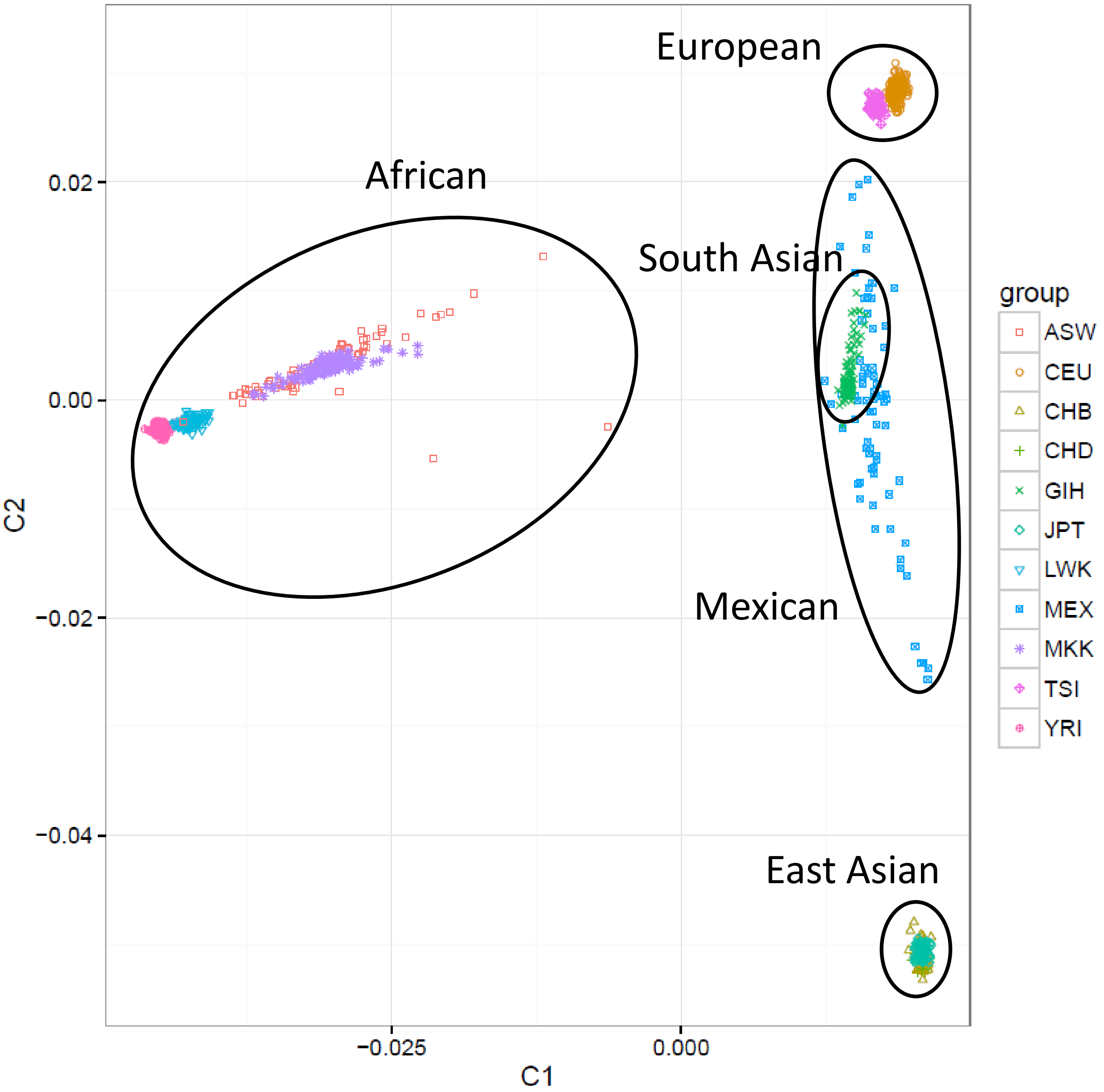
	PC1	PC2	PC3	…	PCn
Individual 1	0.019	0.002	-0.041	…	0
Individual 2	-0.033	0.015	0.037	…	0
Individual 3	-0.016	-0.003	0.019	…	0
…	…	…	…	…	…
Individual m	0.005	0.027	0.004	…	0

PCs are useful to identify possible population stratification

- Check the positions of cases and controls in PCs plot to identify possible bias caused by population stratification

- Projecting PCs to available population studies (e.g. Hapmap, 1000 Genomes Project, Human Genome Diversity Project) can help confirm the ethnicity of each samples and identify systematic errors

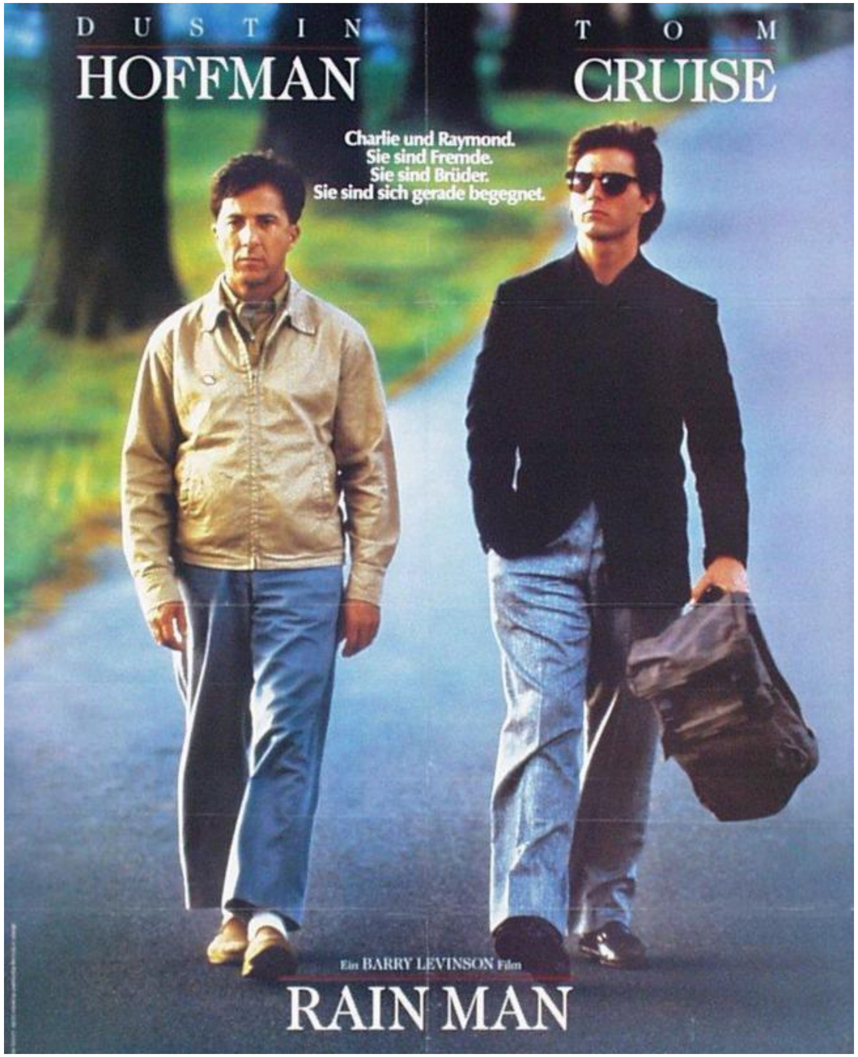
PCs estimated from Hapmap3 SNPs



What you can do if you find population stratifications in the data?

- Drop obvious outliers
- Match the cases and controls according to PCs
- Adjust for PCs in the association tests
- Use ‘genomic control’ factor to correct the observed test statistics
- Use family based controls
- If you find samples from different ethnic groups
  - Perform association tests for different ethnic groups, then
  - Then perform a meta-analysis

Part III – Genetic relationship



Cryptic relatedness

- Some members of a population-based study might actually be close relatives, but not known to the investigator

- Cryptic relatedness is likely to be a far more important confounder than population structure

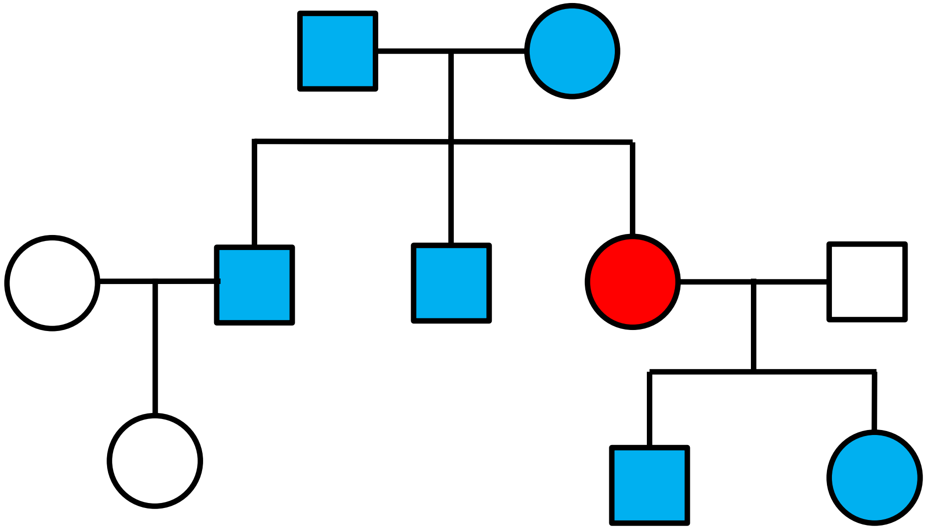
Verifying relationships is crucial

- Genetic analyses require relationships to be specified
  - Family-based design: relationship between samples (Pedigree) must be clear
  - Population-based design: samples should be unrelated

- Mis-specified relationships lead to misleading results
  - Inflated Type I error (false positive)
  - Decreased power

First-degree relatives

- A first-degree relative includes the individual's parents, full siblings, or children
- A pair of first degree relatives shares about 50% of their DNA



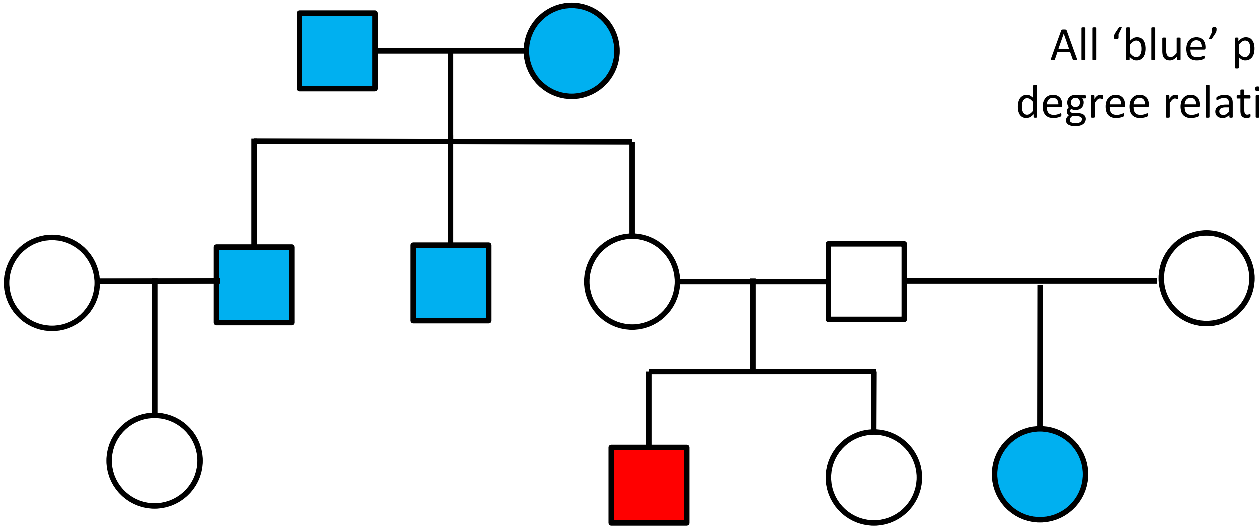
All ‘blue’ person are the first degree relatives of the ‘red’ person

○ Female

□ Male

Second-degree relatives

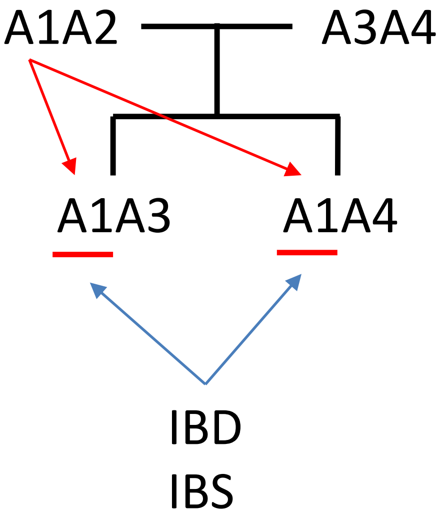
- A second-degree relative includes the individual's grandparents, grandchildren, aunts, uncles, nephews, nieces or half-siblings
- A pair of first degree relatives shares about 25% of their DNA



All ‘blue’ person are the second degree relatives of the ‘red’ person

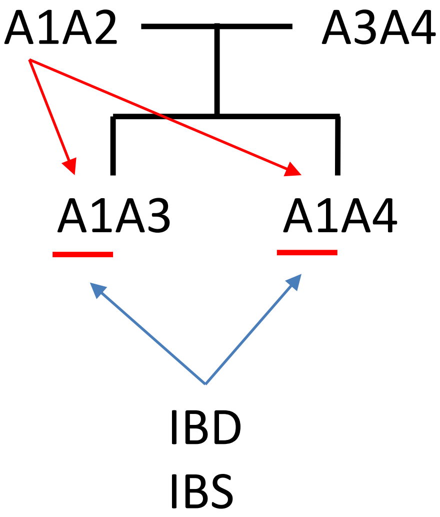
Identical by Descent (IBD)
Identical by State (IBS)

- Two alleles are IBS if they have identical nucleotide sequences
- Two alleles are IBD if they are descended from the same ancestral allele



Identical by Descent (IBD)
Identical by State (IBS)

- Two alleles are IBS if they have identical nucleotide sequences
- Two alleles are IBD if they are descended from the same ancestral allele

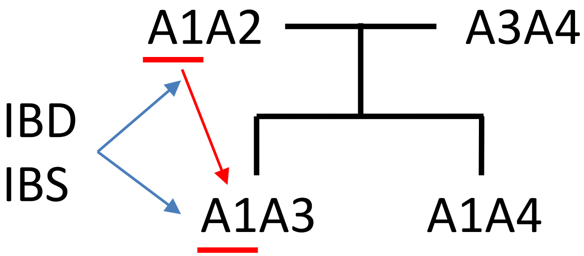


A1A2 A3A4

A1A3 A1A4

IBD

IBS



A1A2 A3A4

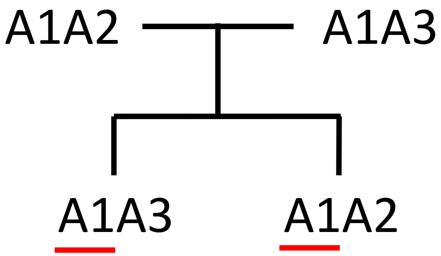
IBD

IBS

A1A3 A1A4

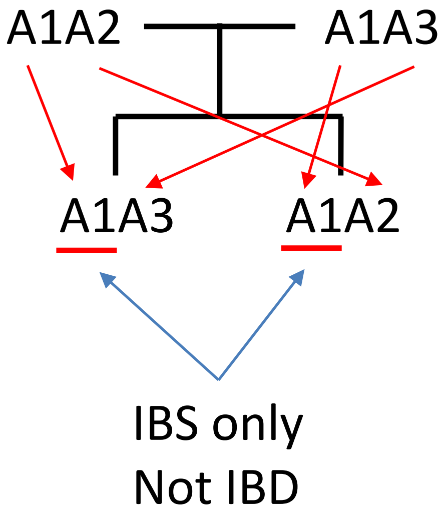
Identical by Descent (IBD)
Identical by State (IBS)

- Two alleles are IBS if they have identical nucleotide sequences
- Two alleles are IBD if they are descended from the same ancestral allele



Identical by Descent (IBD)
Identical by State (IBS)

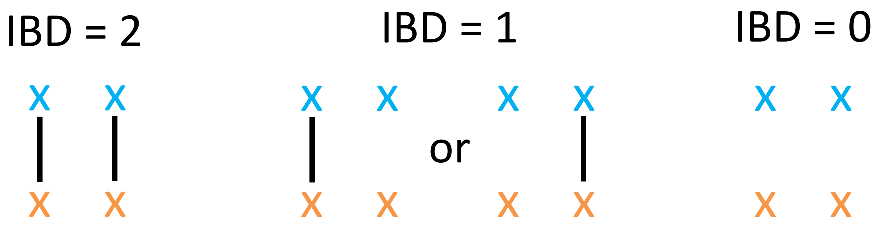
- Two alleles are IBS if they have identical nucleotide sequences
- Two alleles are IBD if they are descended from the same ancestral allele



IBD for close relatives

- Consider a single locus in two individuals. There are four alleles.

Individual 1	x	x
Individual 2	x	x



IBD = 2

IBD = 1

or

IBD = 0

- P(IBD = 2) : probability of individuals sharing two alleles IBD
- P(IBD = 1) : probability of individuals sharing only one allele IBD
- P(IBD = 0) : probability of individuals sharing no allele IBD

- Coefficient of relationship (r) : the fraction of alleles that the two individuals shared IBD

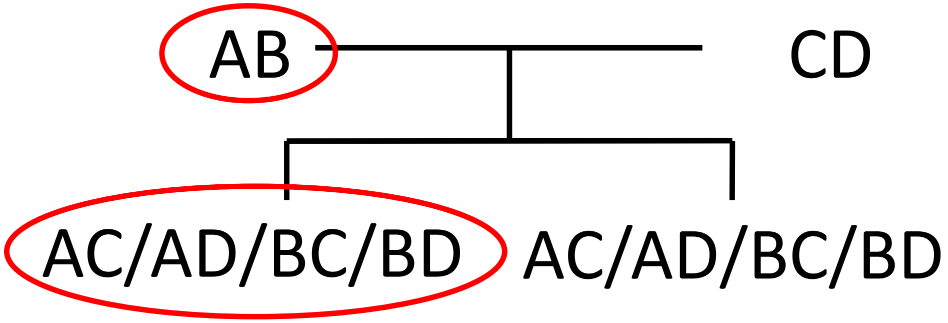
$$r = \frac{P(IBD = 1)}{2} + P(IBD = 2)$$

Probability of IBD for Monozygotic (identical) twins

- P(IBD = 0) = 0
- P(IBD = 1) = 0
- P(IBD = 2) = 1

- r = 0 / 2 + 1 = 1

Probability of IBD for parent-offspring



AB

CD

AC/AD/BC/BD

AC/AD/BC/BD

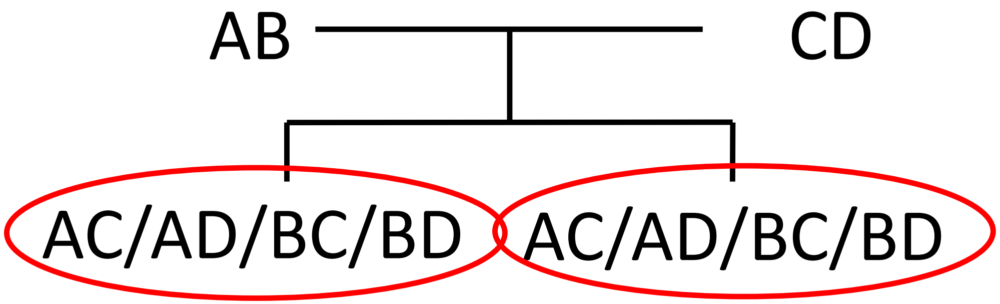
		Parent
		AB
offspring	AC	1
	AD	1
	BC	1
	BD	1

- P(IBD = 0) = 0
- P(IBD = 1) = 1
- P(IBD = 2) = 0

- r = 1 / 2 + 0 = 0.5

IBD	**0**	**1**	**2**
P	0%	100%	0%

Probability of IBD for full-sibs



AB

CD

AC/AD/BC/BD

AC/AD/BC/BD

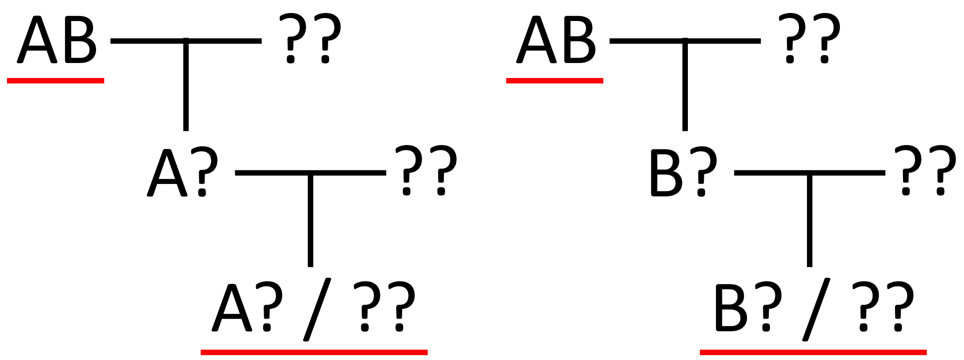
		Sib1			
		AC	AD	BC	BD
Sib2	AC	2	1	1	0
	AD	1	2	0	1
	BC	1	0	2	1
	BD	0	1	1	2

- P(IBD = 0) = 0.25
- P(IBD = 1) = 0.5
- P(IBD = 2) = 0.25

- r = 0.5 / 2 + 0.25 = 0.5

IBD	**0**	**1**	**2**
P	25%	50%	25%

Probability of IBD for grandparent-grandchild



AB ??

A? ??

A? / ??

AB ??

B? ??

B? / ??

		Grand parent
		AB
offspring	A?	1
	??	0
	B?	1
	??	0

- P(IBD = 0) = 0.5
- P(IBD = 1) = 0.5
- P(IBD = 2) = 0

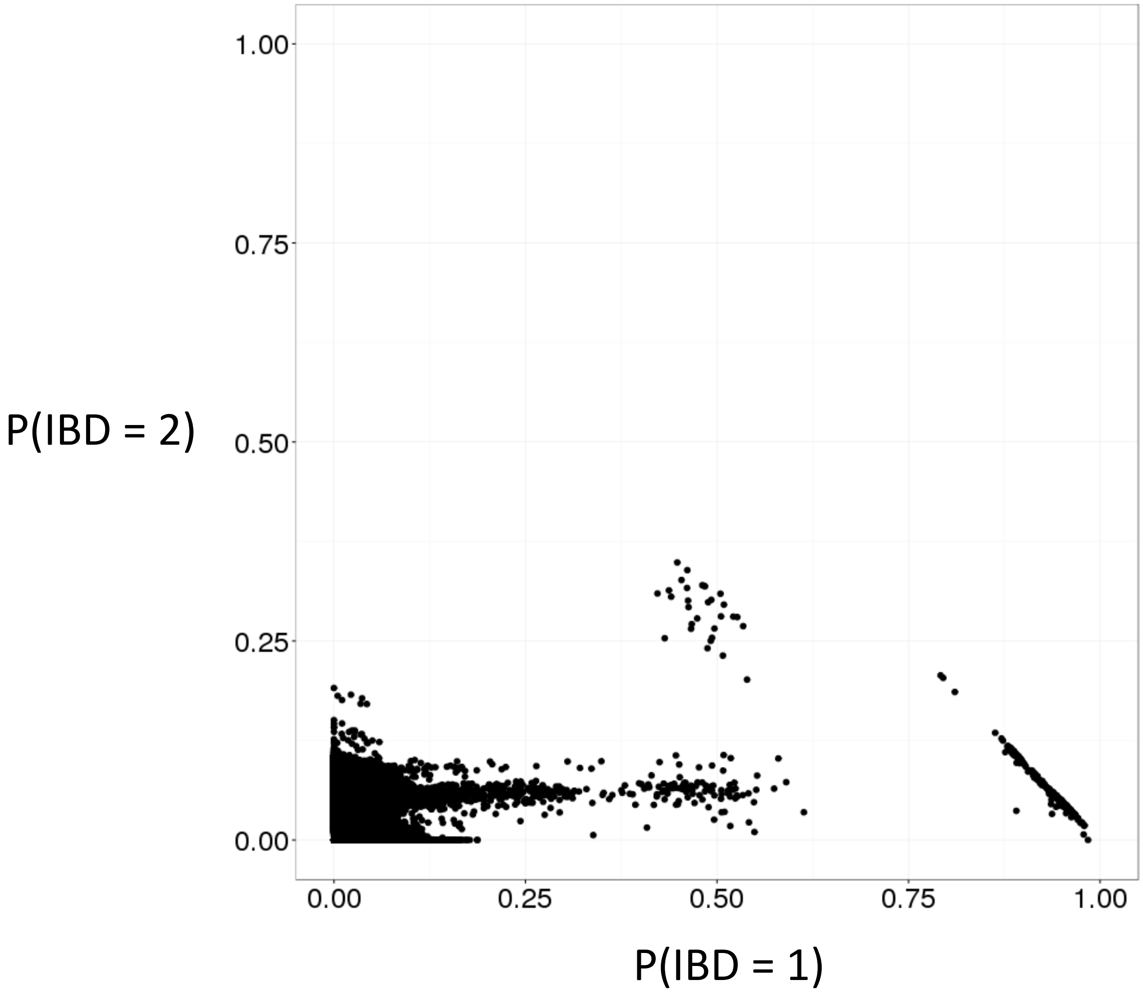
- r = 0.5 / 2 + 0 = 0.25

IBD	**0**	**1**	**2**
P	50%	50%	0%

Probability of IBD for relatives

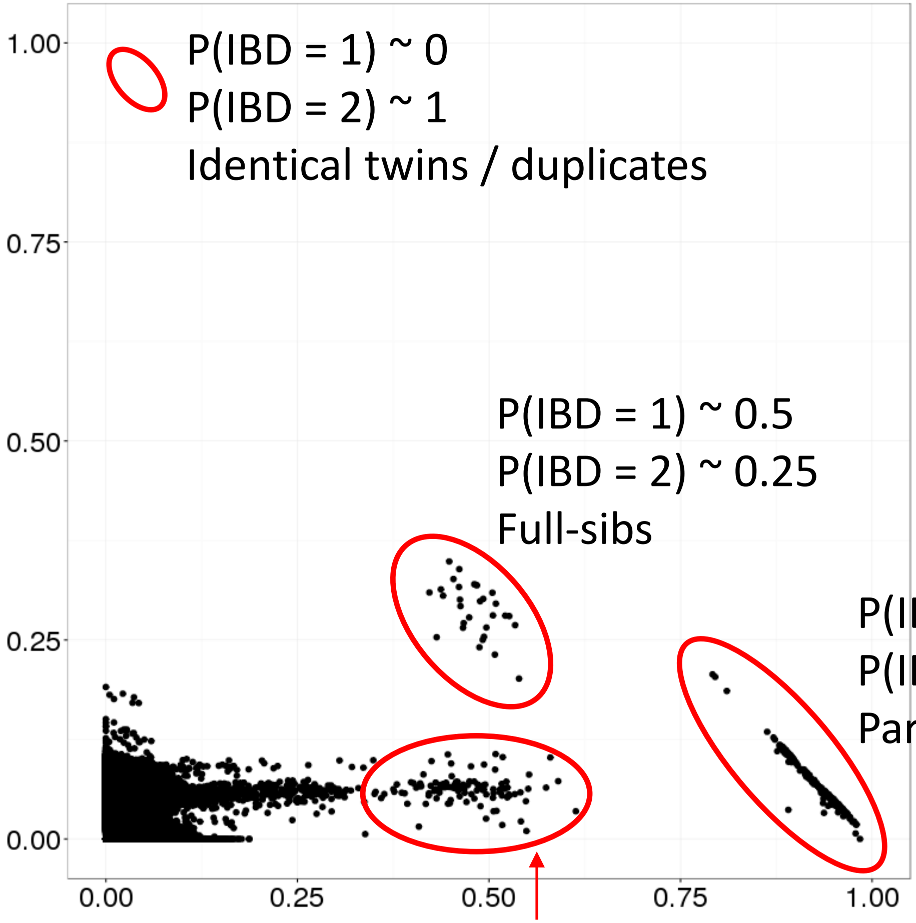
Relationship	P(IBD = 0)	P(IBD = 1)	P(IBD = 2)	r
Identical twins	0	0	1	1
Parent-offspring	0	1	0	0.5
Full sibs Dizygotic twins	0.25	0.5	0.25	0.5
Half sibs Uncle(aunt)-nephew(niece) Grandparent-grandchild	0.5	0.5	0	0.25
First cousins Great grandparent-great grandchild	0.75	0.25	0	0.125
Unrelated	1	0	0	

IBD estimated from Hapmap3 SNPs



IBD estimated from Hapmap3 SNPs

P(IBD = 2)



P(IBD = 1) ~ 0
P(IBD = 2) ~ 1
Identical twins / duplicates

P(IBD = 1) ~ 0.5
P(IBD = 2) ~ 0.25
Full-sibs

P(IBD = 1) ~ 1
P(IBD = 2) ~ 0
Parent-offspring

Half sibs
Uncle(aunt)-nephew(niece)
Grandparent-grandchild

P(IBD = 1) ~ 0.25
P(IBD = 2) ~ 0

What you can do if you find close relatives in your data?

- For each pair/group of close relatives, keep only one of them.

- If you find many close relatives (>10% of all samples), you can try the method accounting for relatedness
  - EMMAX, GEMMA, FaST-LMM, ...

Take Home Messages

- Before genotyping
  - Select cases with comprehensive medical records and clear diagnosis
  - Select controls that matched with cases in age, sex, ethnicity, and other possible cofounding factors
  - Avoid using known close-relatives if you don’t conduct a family-based study.
  - Process a case and its matched control (if possible) in the same batch during DNA collection, storage, and genotyping.

Take Home Messages

- After genotyping
  - Estimate PCs to check population stratification
  - Estimate genetic relationship
  - Remove some cases/controls if necessary (e.g. PCs outliers, close relatives)
  - Adjust for possible confounding factors (e.g. age, sex, PCs …) in association test.
  - Try the methods accounting for population stratification and cryptic relatedness