Bioinformatics Nanocourse Genetic variations and diseases

He Zhang
He.Zhang@UTSouthwestern.edu
Bioinformatics Core Facility
5/31/2018

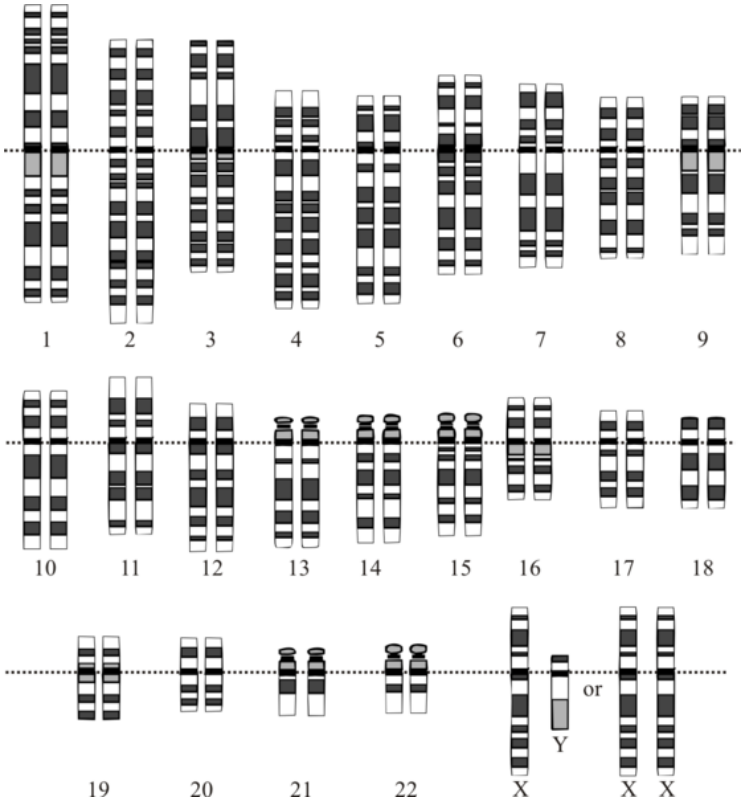
Outline

- Genetic variations in human populations
- Genetic variations and diseases
- Population stratification
- Association study

Human reference genome

- The human genome is the complete set of nucleic acid sequence for humans (Homo sapiens)

- Haploid human genome
  - 22 autosomes
  - X chromosome
  - Y chromosome



Human reference genome

- Human reference genome does not correspond to any actual human individual.

- Genome Reference Consortium human genome (build 37) is mosaic haploid genome derived from 13 anonymous volunteers.
  - One male accounts for 66% of the total

- The latest human reference genome (GRCh38) integrated data from other projects to improve the completeness, but still have gaps covering ~5% of the genome.

Human reference genome

- Human reference genome doesn't represent a 'healthy' genome.
- GRCh37 reference genome have 15 rare variants known to increase the risk of a variety of diseases including type 1 diabetes and hypertension (Chen & Butte, 2011).
- Reference allele is not always 'good' allele.
  - *APOC3* A43T mutation is a protective mutation to coronary Heart Disease
- Reference allele is not always minor allele.

How many variants in human genomes

- A typical genome differs from the reference human genome at 4.1 million to 5.0 million sites

	AFR	AMR	EAS	EUR	SAS
Samples	661	347	504	503	489
Mean coverage	8.2	7.6	7.7	7.4	8
SNPs	4.31M	3.64M	3.55M	3.53M	3.60M
Indels	625k	557k	546k	546k	556k
Large deletions	1.1k	949	940	939	947
CNVs	170	153	158	157	165
Inversions	12	9	10	9	11
Nonsynonymous	12.2k	10.4k	10.2k	10.2k	10.3k

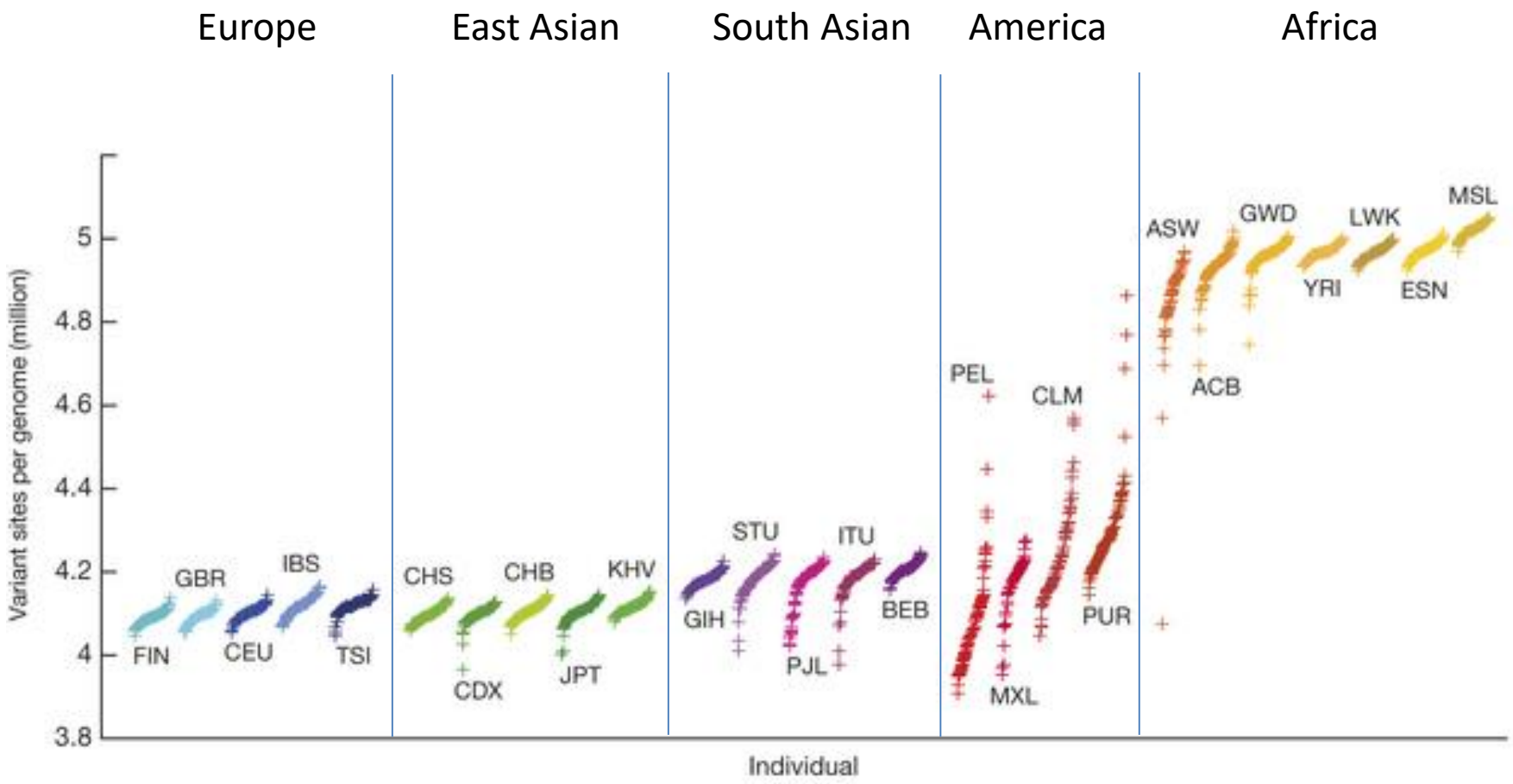
The 1000 Genomes Project Consortium, 2015

Loss of function variants in human genome

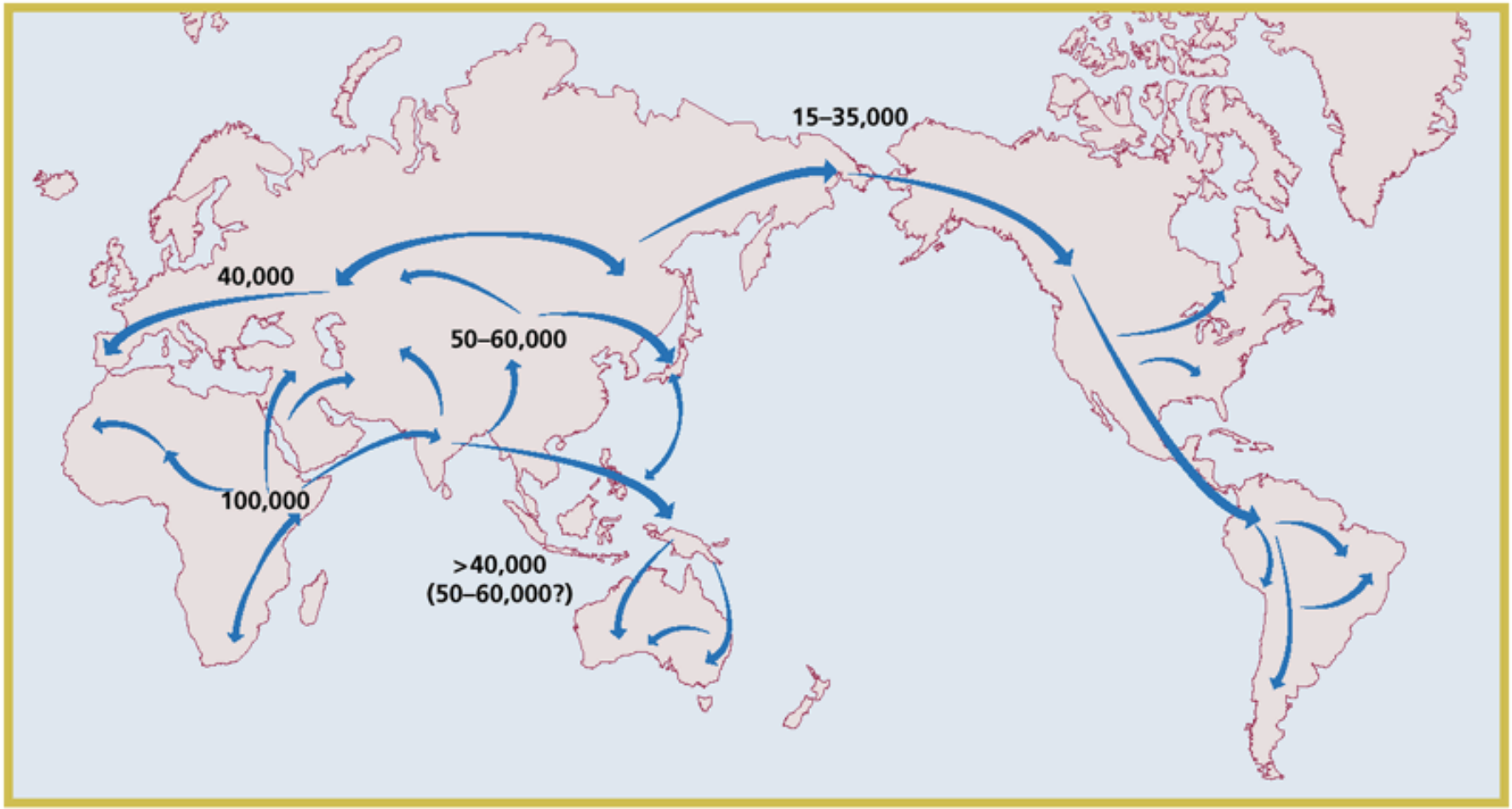
- human genomes typically contain ~100 genuine loss of function (LoF) variants with ~20 genes completely inactivated

	CEU	CHB+JPT	YRI
Stop	26.2 (5.2)	27.4 (6.9)	37.2 (6.3)
Splice	11.2 (1.9)	13.2 (2.5)	13.7 (1.9)
Frameshift indel	38.2 (9.2)	36.2 (9.0)	44.0 (8.0)
Large deletion	28.3 (6.2)	26.7 (5.9)	26.6 (5.5)
Total	103.9 (22.5)	103.5 (24.3)	121.5 (21.7)

Genetic diversity in different populations



Modern human originated from Africa



L. Luca Cavalli-Sforza & Marcus W. Feldman, 2003

Genetic variation exists between populations

- Founder effect and past small population size (increasing the likelihood of genetic drift) may have had an important influence in neutral differences between populations.

- Natural selection may confer an adaptive advantage to individuals in a specific environment if an allele provides a competitive advantage.

- Genetic drift will cause some neutral mutations fixed or disappeared randomly in a population.

Hardy Weinberg Equilibrium (HWE)

- Allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other disturbing factors.

- Suppose:

  f(A) = p

  f(B) = q

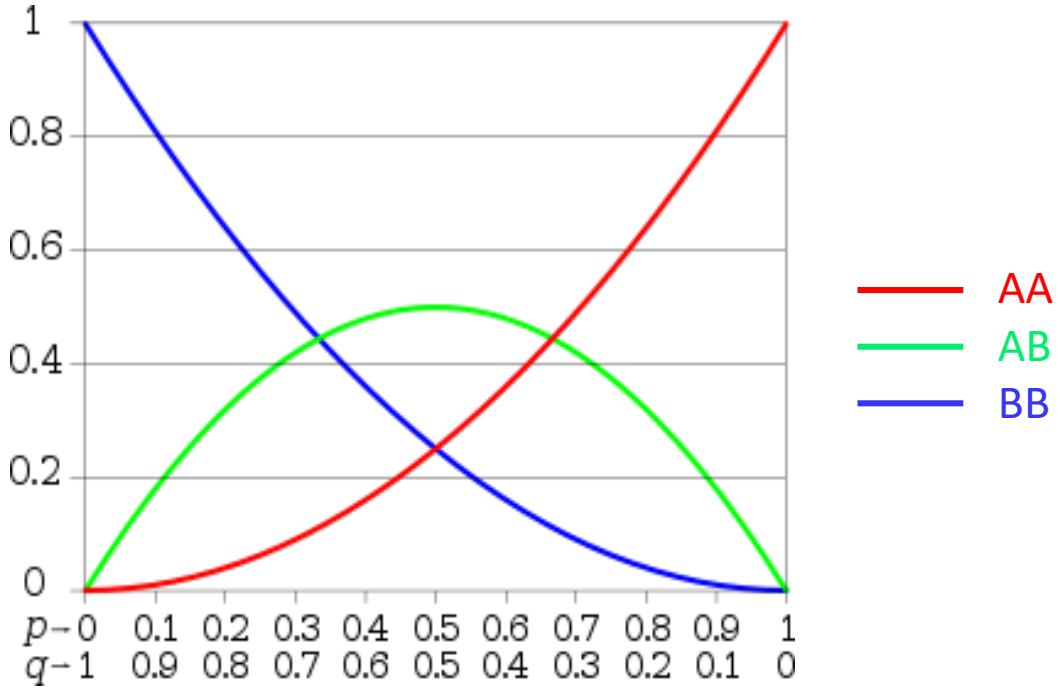
  p + q = 1

- HWE

  $f(AA) = p^2$

  f(AB) = 2pq

  $f(BB) = q^2$



Assumptions of Hardy Weinberg Equilibrium

- Organisms are diploid
- Only sexual reproduction occurs
- Generations are nonoverlapping
- Mating is random
- Population size is infinitely large
- Allele frequencies are equal in the sexes
- There is no migration, mutation or selection

Hardy Weinberg Equilibrium

- HWE is quite robust
  - In most cases, it only take one generation of random mating to return a population to new equilibrium
- HWE does occur in nature
  - Most of the genetic variations in most populations are in HWE (1000 Genomes Project and Hapmap)
- HWE can be used to identify population stratifications, genotyping errors, batch effects,, etc.

Genetic variants and health

- Most of the variants in human genome don't affect health
- A typical human genome contains ~100 loss of function (LoF) variants with ~20 genes completely inactivated (Daniel MacArthur, et al, 2012).
- Variants found in healthy individuals will fall into several overlapping categories
  - Severe recessive disease alleles in the heterozygous state
  - Alleles that are less deleterious but nonetheless have an impact on phenotype and disease risk
  - Benign LoF variation in redundant genes
  - Variants that do not seriously disrupt gene function

Genetic disorder

- A genetic disorder is a disease caused in whole or in part by a change in the DNA sequence away from the normal sequence.

- Genetic disorders can be caused by
  - Mutation(s) in one gene (monogenic disorder),
  - Mutations in multiple genes (multifactorial inheritance disorder)
  - A combination of gene mutations and environmental factors
  - Damage to chromosomes (changes in the number or structure of entire chromosomes, the structures that carry genes).

Monogenetic disorders

- Monogenetic disorders (single-gene disorders, Mendelian disorders) are caused by mutations in a single gene.
- These are usually rare diseases.
- The mutation may be present on one or both chromosomes
- Over 4000 human diseases are caused by single-gene defects
  - Sickle cell disease
  - Cystic fibrosis

Multifactorial inheritance disorders

- Multifactorial inheritance disorders are caused by a combination of variations in different genes, often acting together with environmental factors.

- The effect of each variant/gene was usually small

- Many common diseases including cardiovascular disease, diabetes, and most cancers are examples of such disorders.

Chromosome disorders

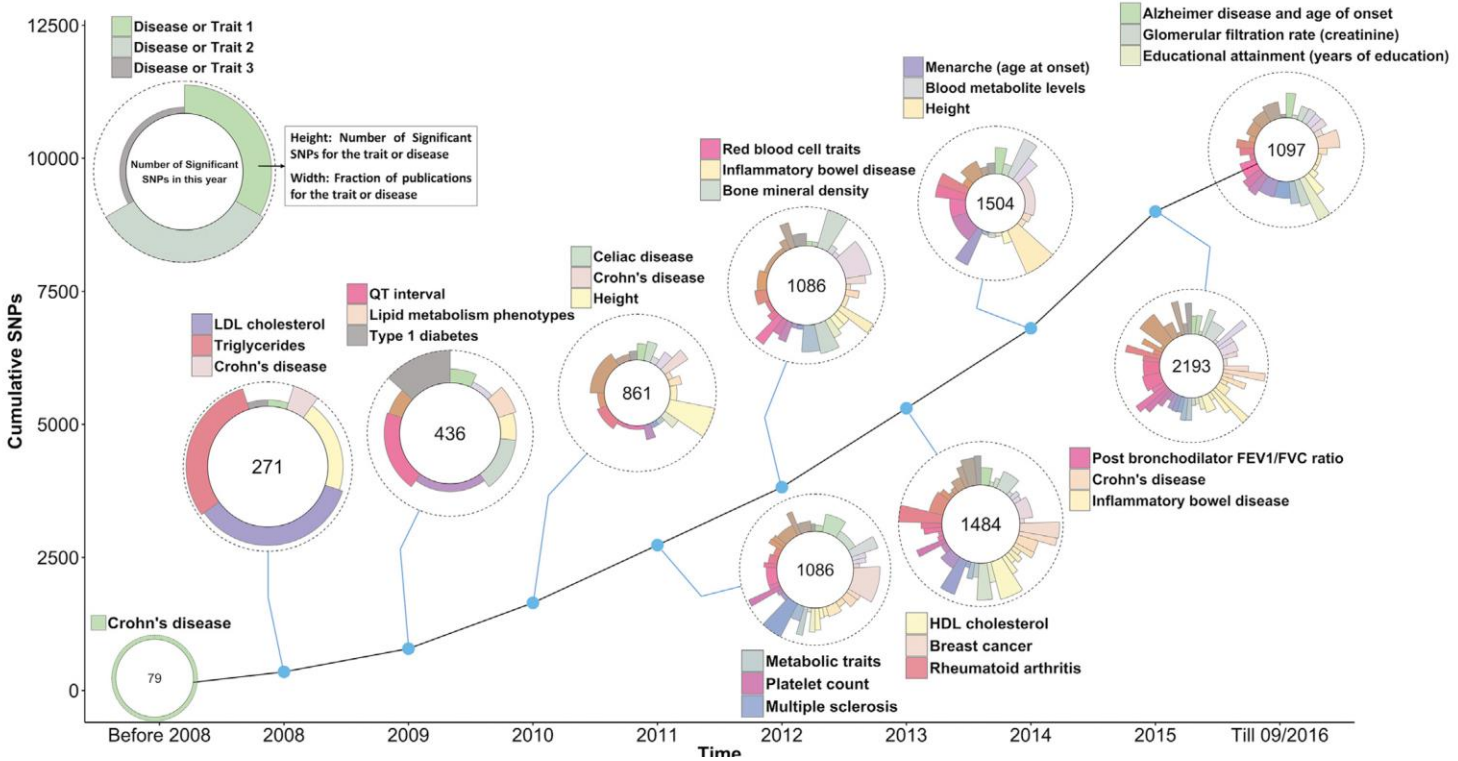
- Chromosome disorders are caused by an excess or deficiency of the genes that are located on chromosomes, or by structural changes within chromosomes.

- Down syndrome is caused by an extra copy of chromosome 21 (called trisomy 21)

- Prader-Willi syndrome is caused by the absence or non-expression of a group of genes on chromosome 15.

Association study is a powerful to identify the genetic variants behind diseases study

- Association study is an approach used in genetic research to test the correlation between disease status and genetic variations



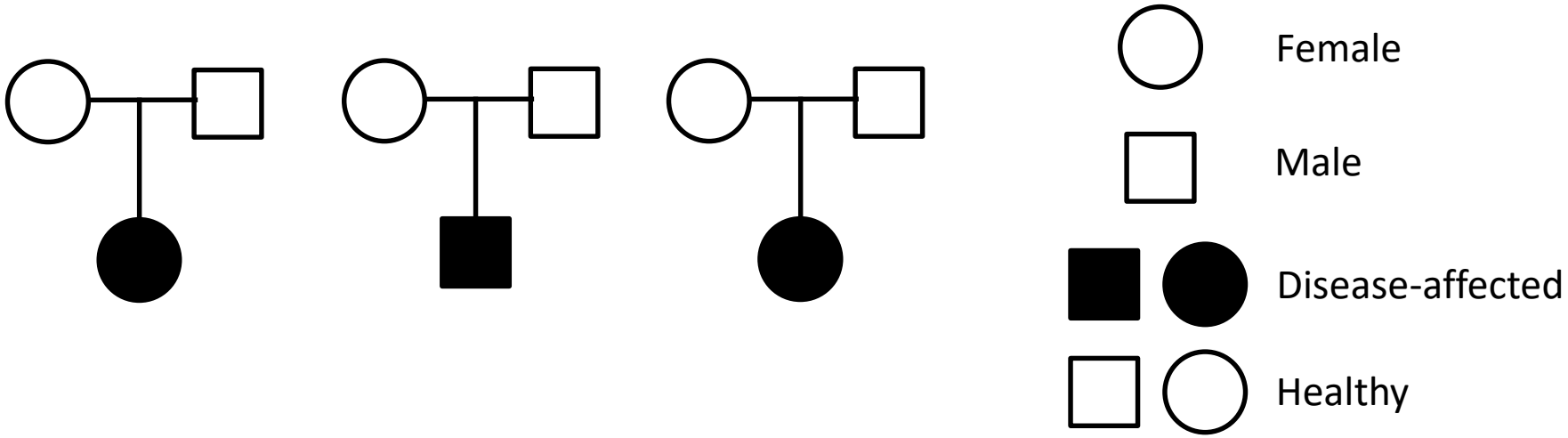
Visscher et al, 2017

Population-based design

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

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



Two primary classes of phenotypes

- Case/control traits (binary trait)
  - Case group affected by a disease
  - Healthy control group
  - Coronary artery disease, type II diabetes, Crohn's disease
- Quantitative traits (continuous traits)
  - Continuous value
  - Body mass index (BMI), Plasma high-density lipoproteins (HDL) level, blood pressure

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

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

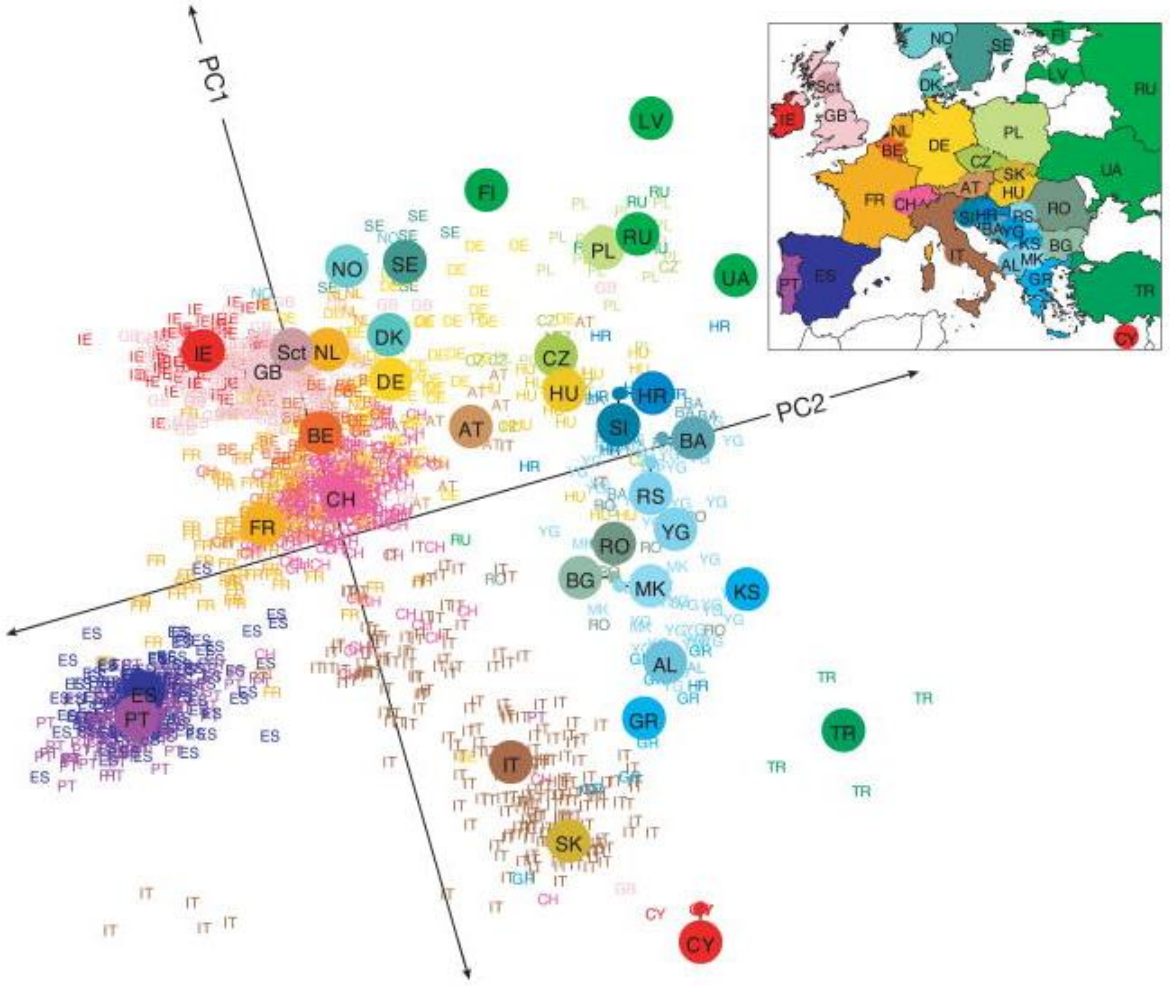
How to minimize systematic bias?

- 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.

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

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=1000

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

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

European, n=1000

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

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

Chinese + European, n = 2000

$\chi^2 = 486$, $P = 10^{-107}$

	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

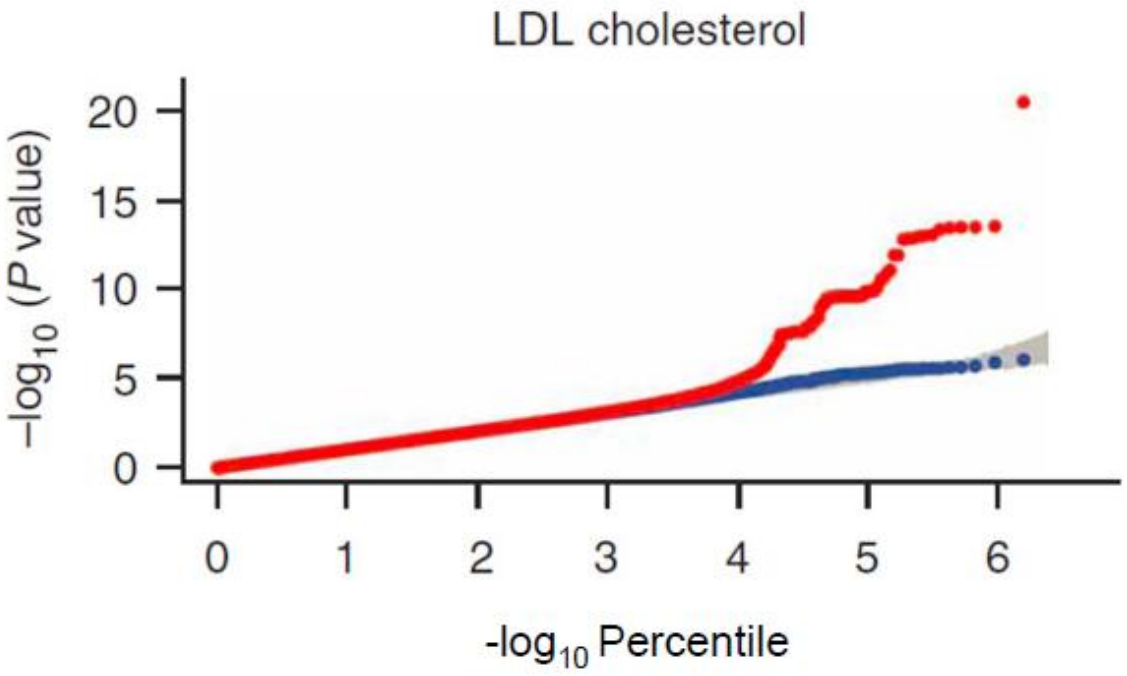
- Principal Components Analysis (PCA) and MultiDimensional Scaling (MDS) are the most commonly methods to infer population stratification

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

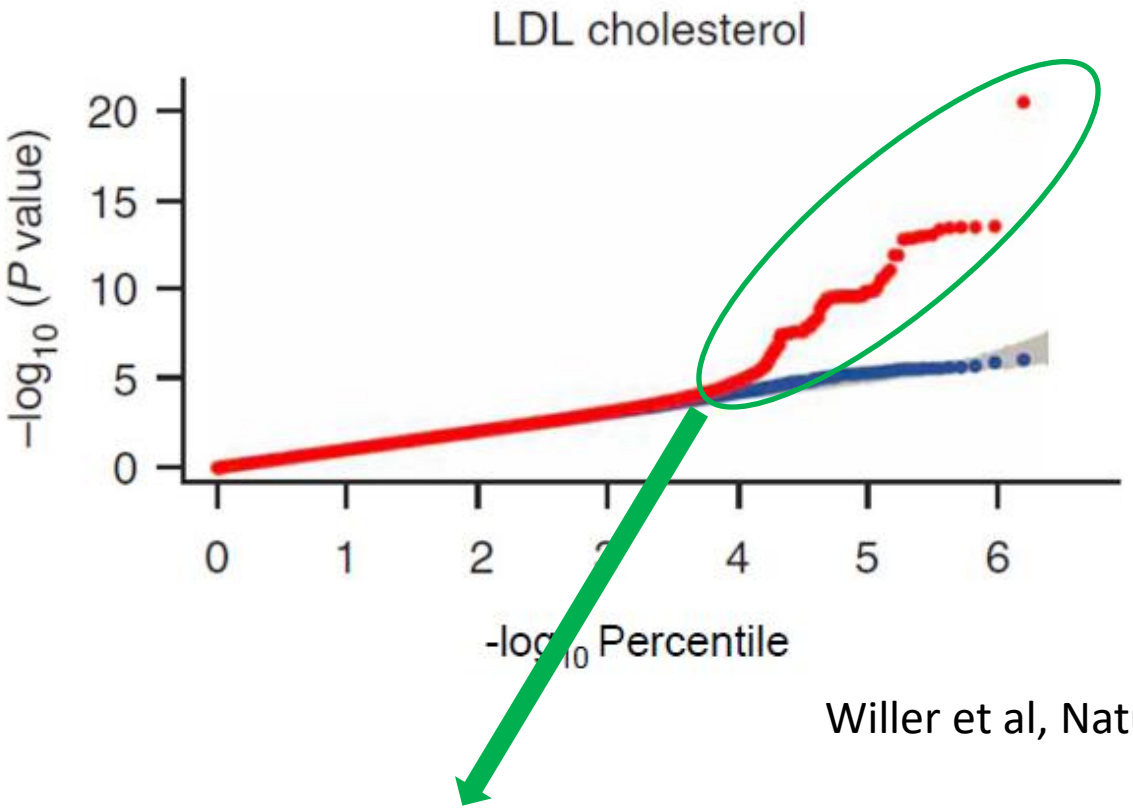
Q-Q plot : A useful diagnostic



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)

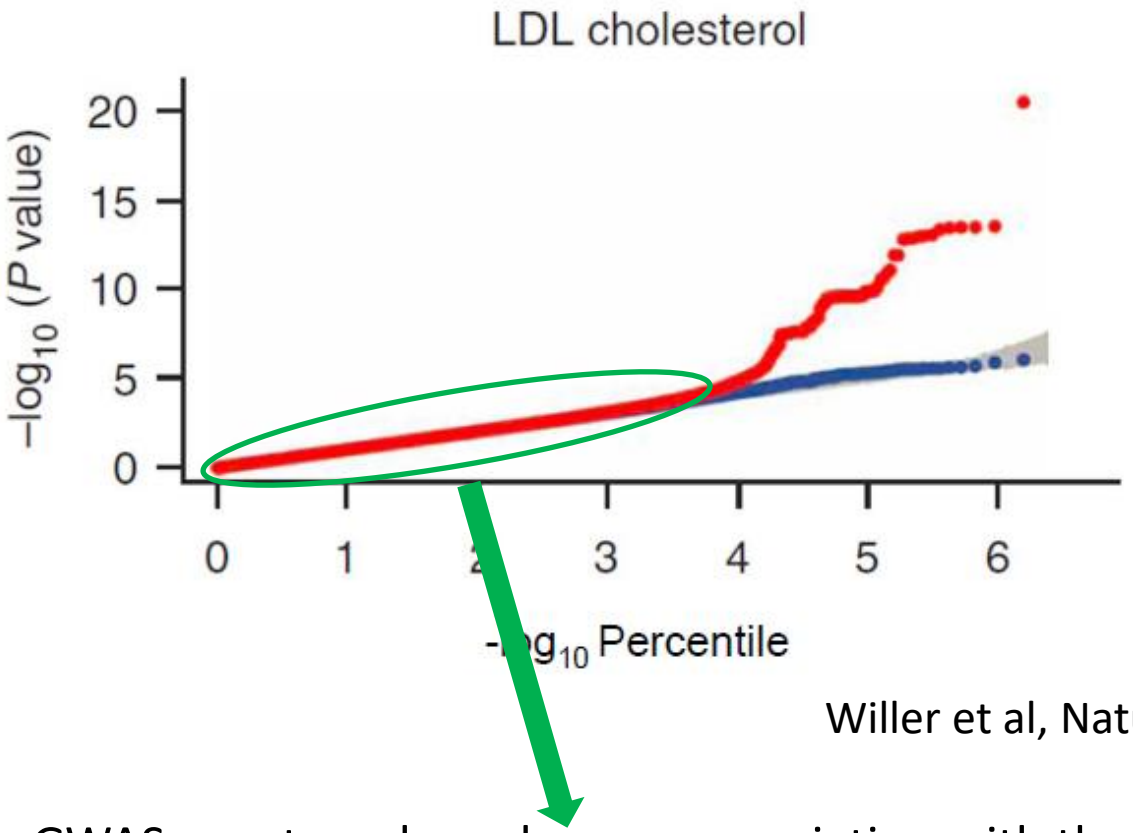
Q-Q plot : A useful diagnostic



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

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

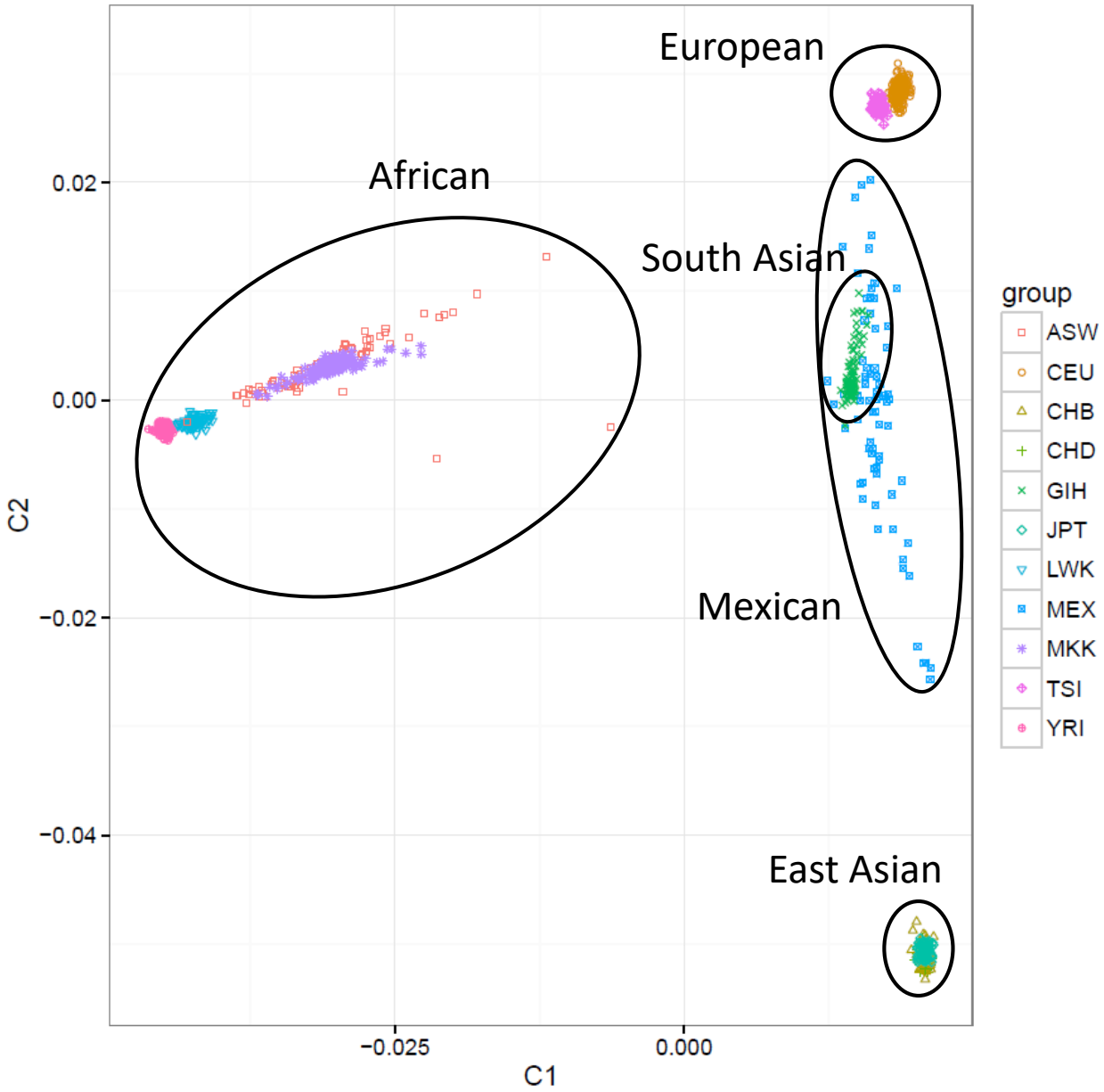
	PC1	PC2	PC3	…	PCn
Individual 1	0.019	0.002	-0.041	…	0
Individual 2	-0.033	0.015	0.037	…	0
Individual 3	-0.016	-0.003	0.019	…	0
…	…	…	…	…	…
Individual m	0.005	0.027	0.004	…	0

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

Association tests

- Binary traits (Case/Control)
  - Pearson χ^2 test
  - Fisher's exact test
  - Logistic regression
- Quantitative traits
  - Linear regression

Association between genotype and disease

Observed genotype counts

	GG	GT	TT	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Observed allele counts

	G	T	Total
Cases	$2r_0 + r_1$	$r_1 + 2r_2$	$2R$
Controls	$2s_0 + s_1$	$s_1 + 2s_2$	$2S$
Total	$2n_0 + n_1$	$n_1 + 2n_2$	$2N$

Expected allele counts

G	T
$2R(2n_0+n_1)/(2N)$	$2R(n_1+2n_2)/(2N)$
$2S(2n_0+n_1)/(2N)$	$2S(n_1+2n2)/(2N)$

Chi-square test for independence of rows and columns

$$\sum \frac{(Obs - Exp)^2}{Exp^2} \sim \chi^2$$ with 1 df

The odds ratio: a measure of effect size

- Odds of an event occurring

$$\text{Odds} = \frac{\Pr(event\ occurs)}{\Pr(event\ doesn't\ occur)} = \frac{\Pr(event\ occurs)}{1-\Pr(event\ occurs)}$$

Allele counts

	G	T
Cases	g1	t1
Controls	g2	t2

$$\text{Odds (G allele occurs in a case)} = \frac{g1}{g2}$$

$$\text{Odds (T allele occurs in a case)} = \frac{t1}{t2}$$

$$\text{Odds Ratio (OR)} = \frac{odds\ (G\ allele\ occurs\ in\ a\ case)}{odds\ (T\ allele\ occurs\ in\ a\ case)} = \frac{g1/g2}{t1/t2} = \frac{g1t2}{g2t1}$$

Odds Ratio = 1 : no association between genotype and phenotype
Odds Ratio > 1 : G allele increase risk of disease
Odds Ratio < 1 : T allele increase risk of disease

Logistic regression: more flexible

- Similar to linear regression, used for binary traits

- Y_i : phenotype for individual *i*
  - Y_i = 0 for controls
  - Y_i = 1 for cases
- X_i : genotype for individual *i*
  - X_i = 0 for TT
  - X_i = 1 for GT
  - X_i = 2 for GT
- Basic logistic regression model

  let $p_i = E(Y_i \mid X_i)$, expected value of Y_i given X_i

  Define $\text{logit}(p_i) = \lg[p_i/(1- p_i)]$

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i + \varepsilon$$

$$Odds\ Ratio = e^{\beta_1}$$

Logistic regression: more flexible

- Similar to linear regression, used for binary traits

- Y_i : phenotype for individual *i*
  - Y_i = 0 for controls
  - Y_i = 1 for cases
- X_i : genotype for individual *i*
  - X_i = 0 for TT
  - X_i = 1 for GT
  - X_i = 2 for GT
- Logistic regression model including confounding factors:
  - let $p_i = E(Y_i \mid X_i, C_i, D_i, \ldots)$, expected value of Y_i given X_i , C_i , D_i , …
  - Define $\text{logit}(p_i) = \lg[p_i/(1- p_i)]$

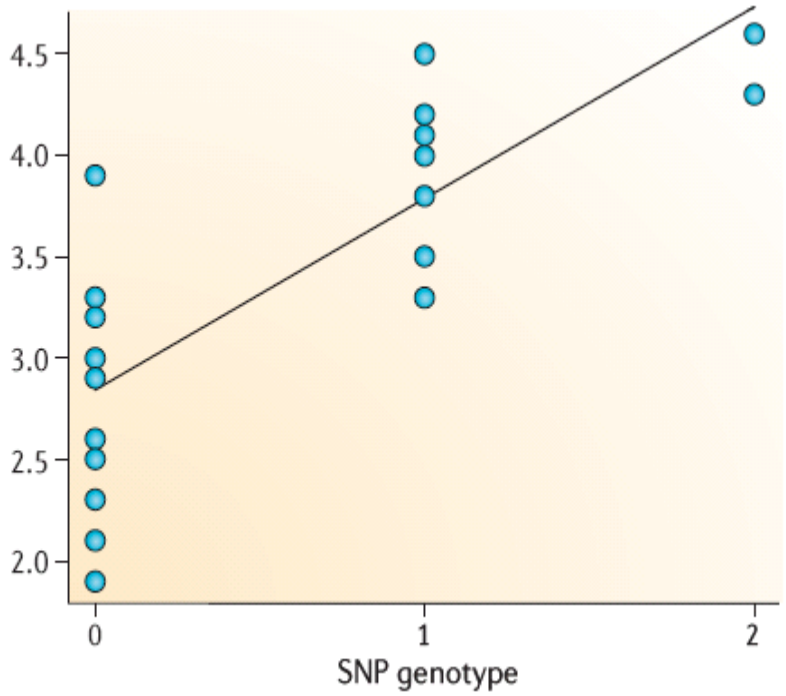
$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \cdots + \varepsilon$$

$$Odds\ Ratio = e^{\beta_1}$$

Linear regression for quantitative traits

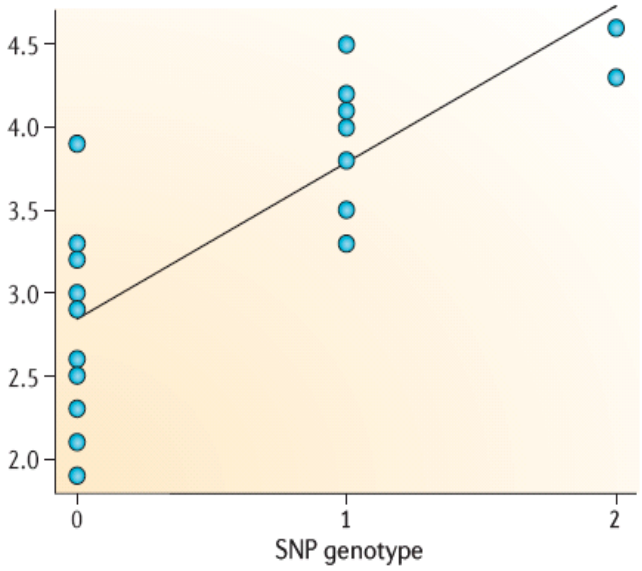
- One cannot create a contingency matrix as in case/ control studies

- For each locus, fit a linear regression using the number of minor allele (or alternative allele) of the individual as covariate



Linear regression for quantitative traits

- Y_i : (continuous) phenotype for individual *i*
  - Y_i = 0 for controls
  - Y_i = 1 for cases
- X_i : genotype for individual *i*
  - X_i = 0 for TT
  - X_i = 1 for GT
  - X_i = 2 for GT



- Basic linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

β_1 : effect of G allele

- Linear regression model with confounding factors (C_i , D_i , …)

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \cdots + \varepsilon$$

Multiple testing

- Suppose:
  - You performed association tests for 1,000,000 SNPs
  - None of the 1,000,000 SNPs is associated with the disease
  - *P* (e.g. 0.05) was used as p-value cutoff

- How many SNPs (n) are expected to have a p-value less than *P*?
  - 1,000,000 x *P*
  - n = 50,000 if *P* = 0.05

- The cutoff of p-value should be adjusted based on the tests (number of SNPs)
  - Bonferroni correction: 0.05 / 1,000,000 = 5×10^{-8}

Genome-wide significance

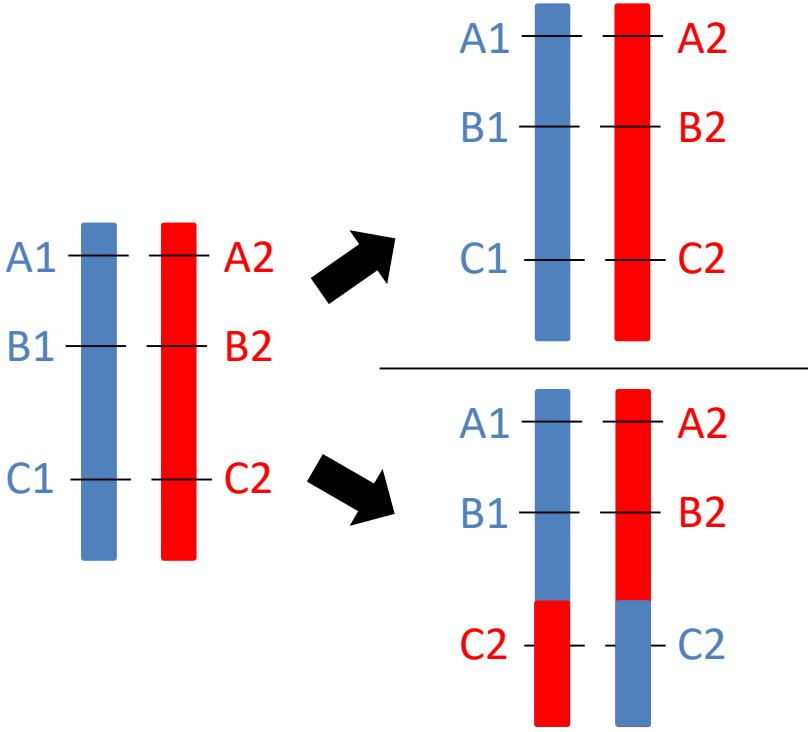
Association does not imply causation

- Association between a genetic variant and disease does not mean that variant causes the disease

- The association may be related with some other confounding factor
  - Ethnic ancestry
  - Genotyping batch, genotyping centre
  - DNA quality
  - Environmental exposures

Confounding: linkage disequilibrium

- Linkage disequilibrium (LD) is the non-independence of alleles at nearby markers in a population because of a lack of recombination between the markers



A and C	B and C	A and B
A1 C1	B1 C1	A1 B1
A1 C2	B1 C2	A2 B2
A2 C1	B2 C1	
A2 C2	B2 C2	

A1 and B1 always appear together
A2 and B2 always appear together

Databases

- NCBI dbSNP (Short Genetic Variations) and dbVAR (Genomic Structural Variations)
  - https://www.ncbi.nlm.nih.gov/projects/SNP/
  - https://www.ncbi.nlm.nih.gov/dbvar
- 1000 Genomes Project
  - 2,504 individuals from 26 populations, whole-genome sequencing
  - http://www.internationalgenome.org/
- Exome Aggregation Consortium
  - 60,706 unrelated individuals, exome sequencing
  - http://exac.broadinstitute.org/
- Online Mendelian Inheritance in Man (OMIM)
  - A comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily
  - http://omim.org
- NCBI ClinVar
  - a public archive of reports of the relationships among human variations and phenotypes
  - https://www.ncbi.nlm.nih.gov/clinvar/