

CS2220 Introduction to Computational Biology

WEEK 8: GENOME-WIDE ASSOCIATION STUDIES (GWAS)

1

Dr. Mengling FENG
Institute for Infocomm Research
Massachusetts Institute of Technology
mfeng@mit.edu

PLANS FOR WEEK 7 AND WEEK 8

- Week 7, 1st Oct 2015
 - 2 hours class: Single (Simple) Nucleotide Polymorphism
 - 1 hour briefing on project and forming of project teams
- Week 8, 7th Oct 2015
 - Definition of SNP
 - Q & A
 - 2 hours class: Genome-wide Association Study (GWAS)
 - 30 mins Q&A on the lectures and project

WEEK 8'S LEARNING OBJECTIVES

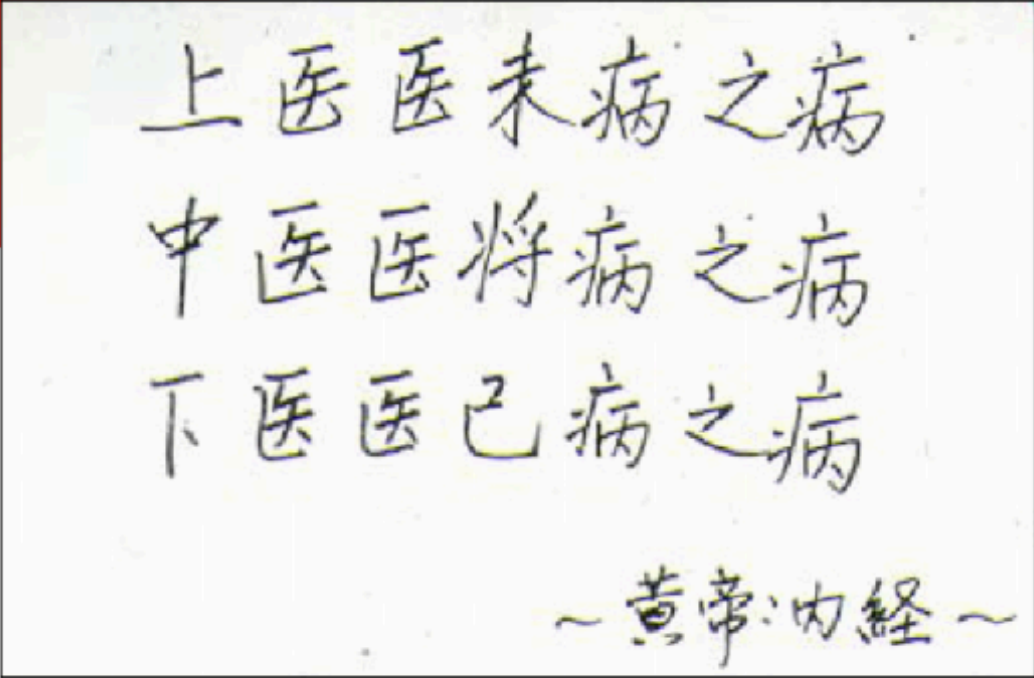
- After the class, students should be able to
 - Define Gene-disease association studies
 - Appreciate the motivations and applications of GWAS
 - Explain the differences between GWAS and Candidate Gene Studies
 - Explain the typical method and workflow for GWAS studies and, more importantly, considerations and limitations for each step
 - Understand the concepts of
 - Linkage Disequilibrium
 - Hypothesis testing
 - Multiple testing correction
 - Population stratification bias
 - Get to know the online resources

GENETIC ASSOCIATION STUDIES

- Investigate how genotypes may associate or cause particular phenotypes
- Genome-Wide Association Study (GWAS)
 - A type of genetic association study
 - Focus on common SNPs
 - Involves large amount of SNPs
- A overview
 - <https://www.youtube.com/watch?v=mblaqn4yU70>

MOTIVATIONS AND VISION

- Preventive Medication



上医医未病之病
中医医将病之病
下医医已病之病
~ 黄帝内经 ~

“Superior Doctors Prevent the Disease.
Mediocre Doctors Treat the Disease Before Evident.
Inferior Doctors Treat the Full Blown Disease.”

-Huang Dee: Nai - Ching (2600 B.C. 1st Chinese Medical Text)



THE VISION: PREVENTIVE MEDICATION

- Prevent disease from occurring
 - SNPs as bio-markers to estimate personalized disease risk
 - Inspire behavioral and environmental changes
 - Some preventive intervention



THE VISION: PREVENTIVE MEDICATION

- Prevent disease from occurring
- Identify the cause of the disease
 - Genomics identifies the cause of disease
 - “All medicine may become pediatrics” Paul Wise, Professor of Pediatrics, Stanford Medical School, 2008
 - Treat the cause of the disease rather than the symptoms
- Health care costs can be greatly reduced if
 - Invests in preventive medicine
 - One targets the cause of disease rather than symptoms
- Challenges and limitations:
 - Penetrance and environmental factors

PENETRANCE AND ENVIRONMENTAL FACTORS

○ Penetrance

- Is the proportion of individuals carrying a particular variant of a gene (allele or genotype) that also expresses an associated trait (phenotype).

○ Highly penetrant Mendelian single gene diseases

- Huntington's Disease caused by excess CAG repeats in huntingtin's protein gene
- Autosomal dominant, 100% penetrant, invariably lethal

○ Reduced penetrance, some genes lead to a predisposition to a disease

- BRCA1 & BRCA2 genes can lead to a familial breast or ovarian cancer
- Disease alleles lead to 80% overall lifetime chance of a cancer, but 20% of patients with the rare defective genes show no cancers

○ Complex diseases requiring alleles in multiple genes

- Many cancers (solid tumors) require somatic mutations that induce cell proliferation, mutations that inhibit apoptosis, mutations that induce angiogenesis, and mutations that cause metastasis
- Cancers are also influenced by environment (smoking, carcinogens, exposure to UV)

○ Some complex diseases have multiple causes

- Genetic vs. spontaneous vs. environment vs. behavior
- Some complex diseases can be caused by multiple pathways
- Type 2 Diabetes can be caused by reduced beta-cells in pancreas, reduced production of insulin, reduced sensitivity to insulin (insulin resistance) as well as environmental conditions (obesity, sedentary lifestyle, smoking etc.).

CANDIDATE GENE STUDIES VS GWAS



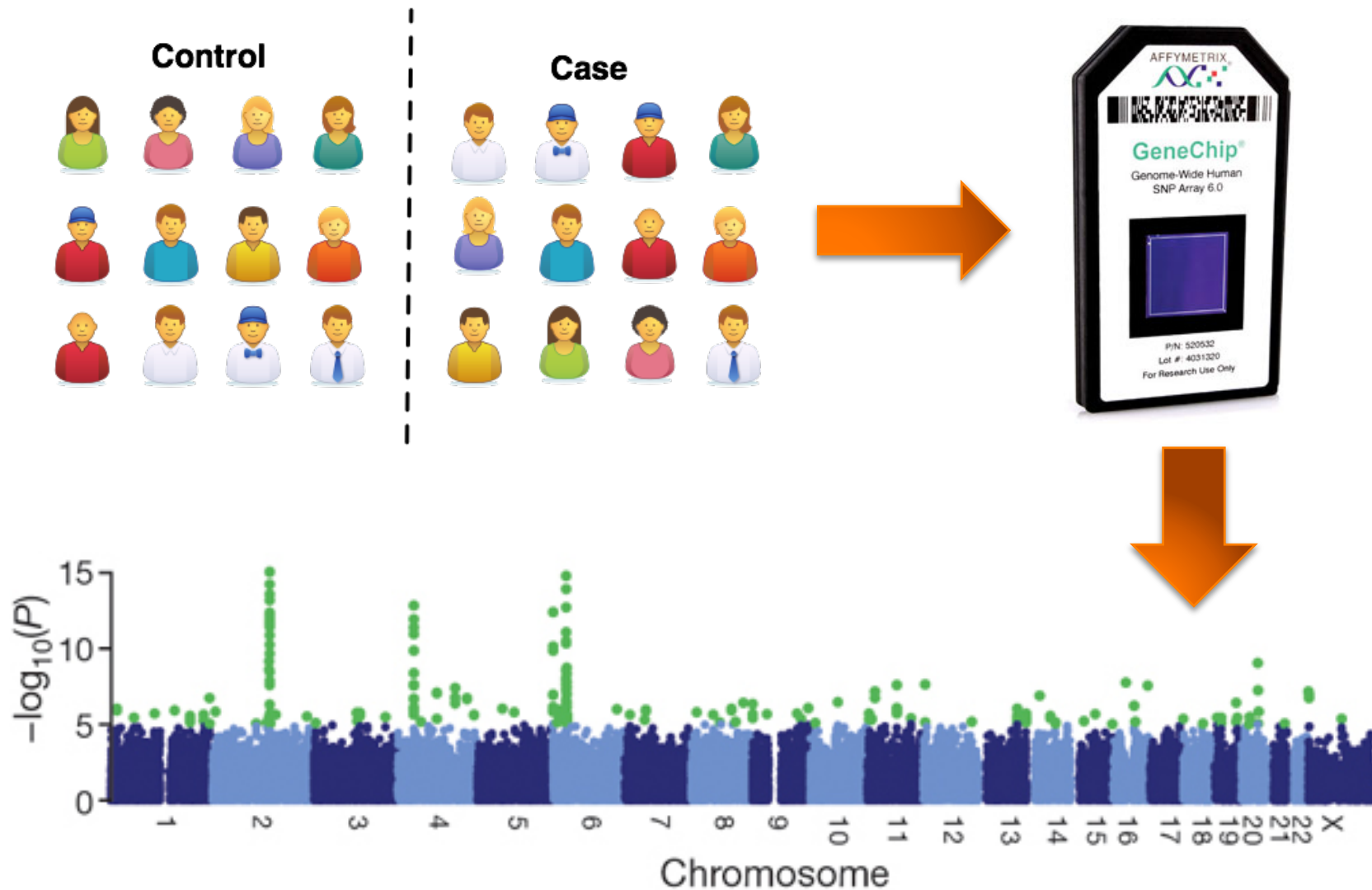
CANDIDATE GENE STUDIES VS GWAS



© Francis Collins, 2008

TYPES OF GWAS

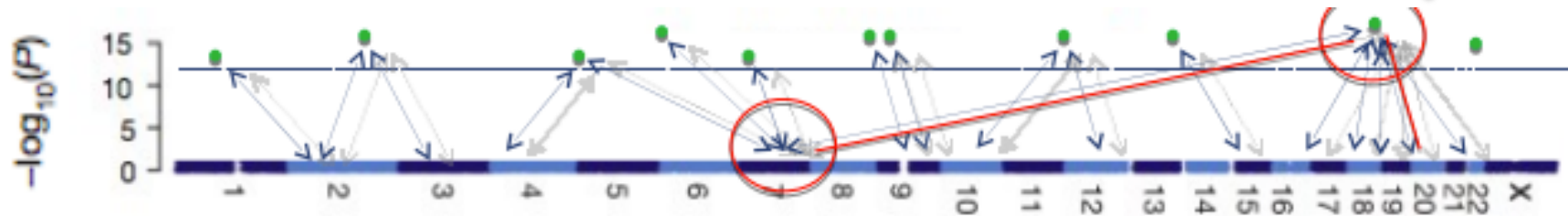
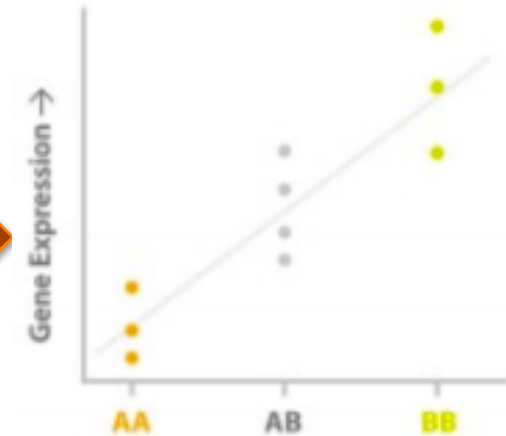
DISEASE ASSOCIATIONS (BINARY)



TYPES OF GWAS

QUANTITATIVE TRAITS

Cohort Population



TYPICAL STEPS OF GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)

SAMPLING (CASE & CONTROL)

- Primer on Causal Inference
 - Definition of Causation



DEFINITION OF CAUSATION



DEFINITION OF CAUSATION

	$Y^{\text{red}=0}$	$Y^{\text{red}=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	1
Hestia	0	0
Poseidon	1	1
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaestus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	0	0
Dionysus	0	0

$$P[Y^{\text{red}=0} = 1] = 8/20 = 0.4$$

$$P[Y^{\text{red}=1} = 1] = 12/20 = 0.6$$

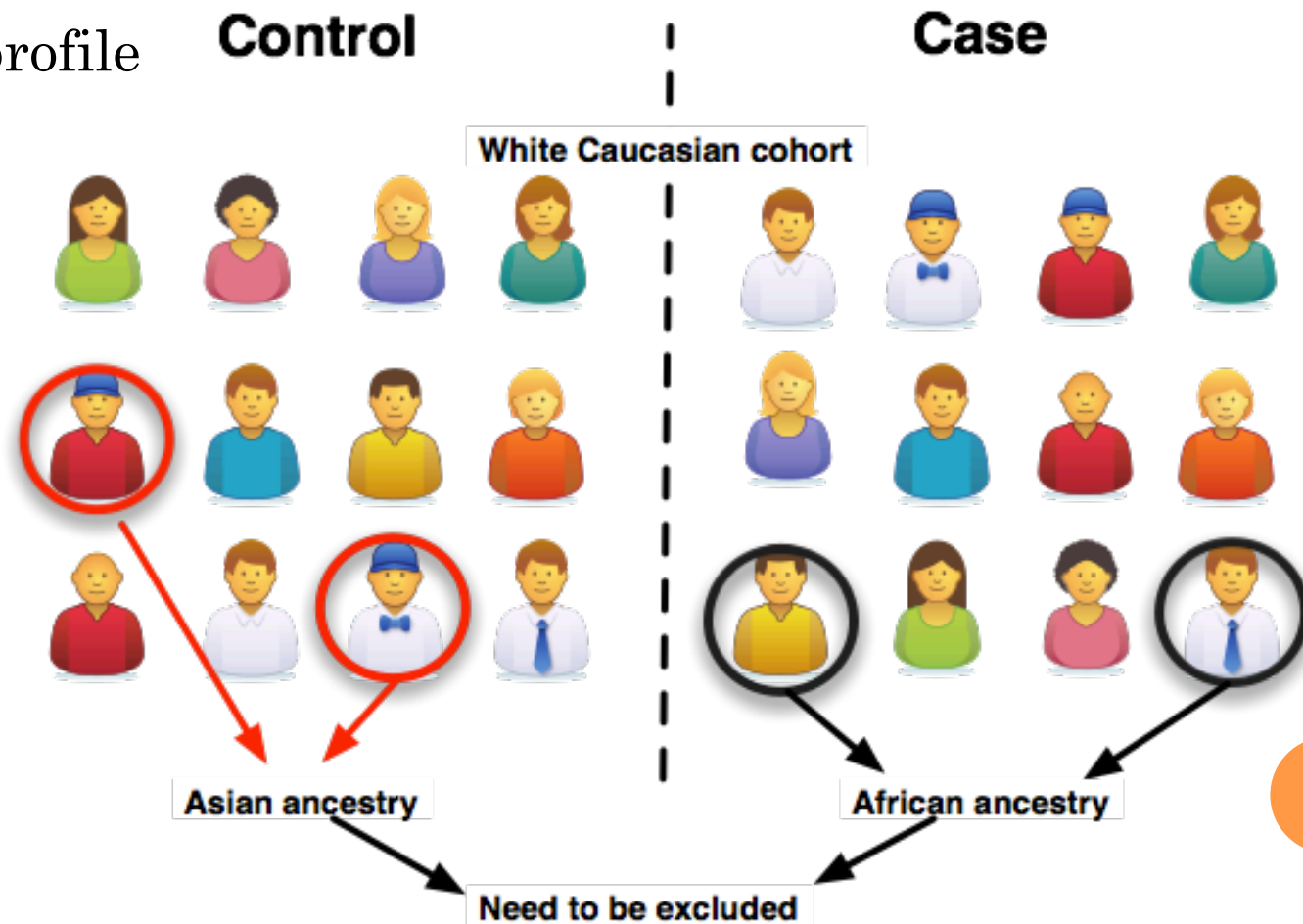
$$P[Y^{\text{red}=0} = 1] \neq P[Y^{\text{red}=1} = 1]$$

indicates sign of causation

Red Pill => Cure

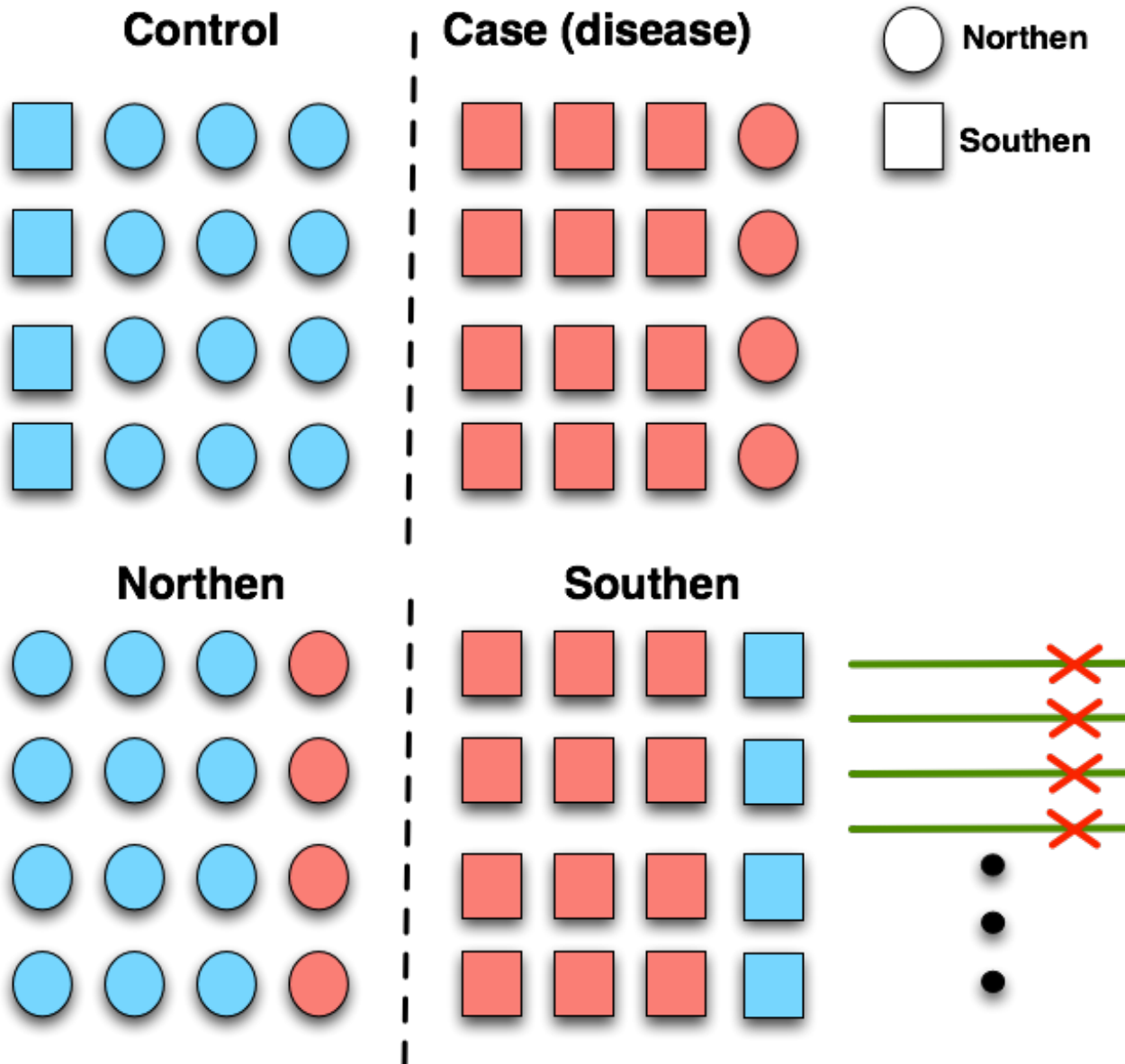
SAMPLING (CASE & CONTROL)

- A matched cohort
 - Age
 - Gender
 - Other demographics
 - Ancestry profile



POTENTIAL BIASES

POPULATION STRATIFICATION



DETECTION OF POPULATION STRATIFICATION

GENOMIC CONTROL

Observed Armitage Trend Statistics

$$Y^2 = \frac{N(N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{R(N - R)(N(n_1 + 4n_2) - (n_1 + 2n_2)^2)}$$

Alleles	aa	Aa	AA	total
North	r0	r1	r2	R
South	s0	s1	s2	S
total	n0	n1	n2	N

Expected Chi2 Statistics

$$\chi^2 \sim X_A^2 = \frac{2N(2N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{4R(N - R)(2N(n_1 + 2n_2) - (n_1 + 2n_2)^2)}$$

Measure of population stratification λ

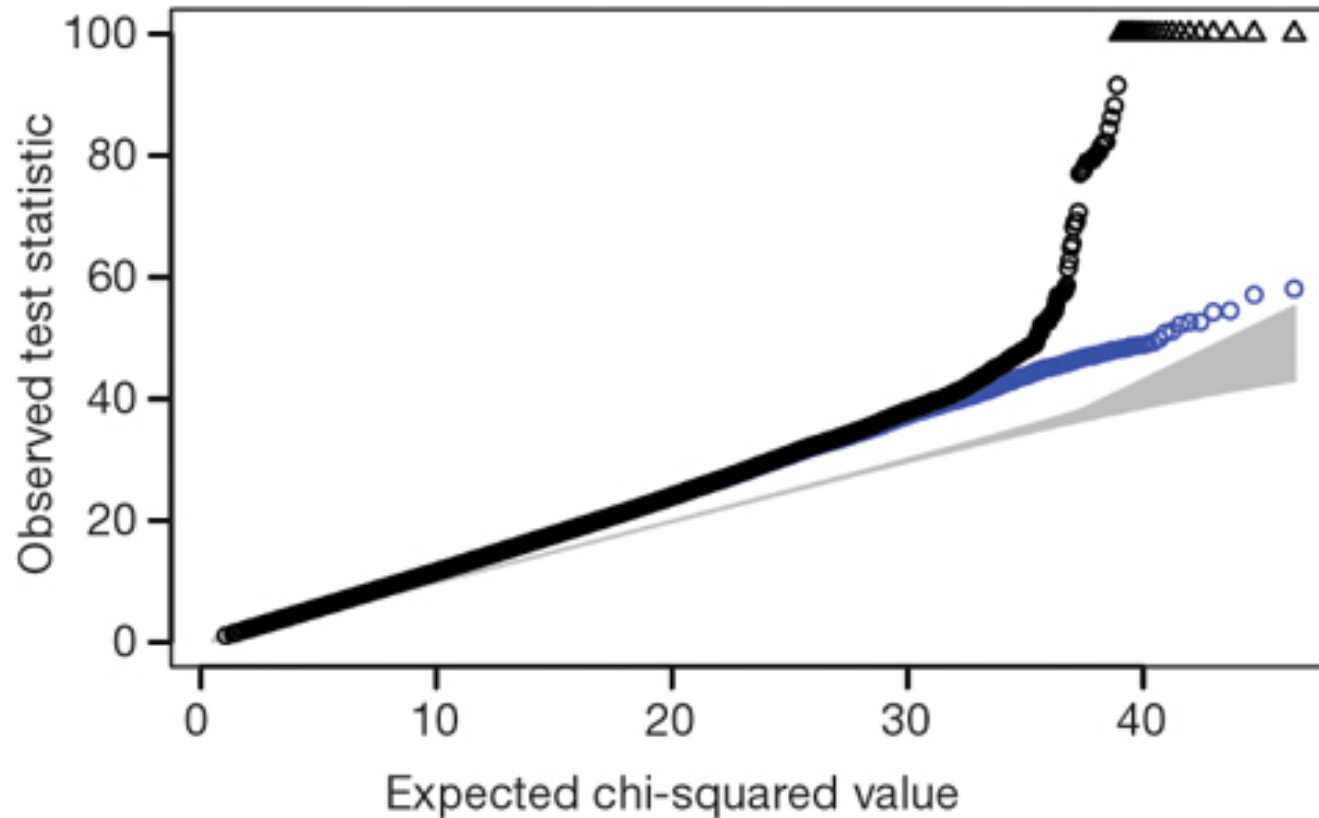
$$Y^2 \sim \lambda \chi_1^2$$

$\lambda = 1$ → No population stratification

$\lambda > 1.05$ → Significant population stratification

POPULATION STRATIFICATION

QQ PLOT & CORRECTIONS



○ Solutions

- Remove the deviating SNPs
- Conduct separate studies for different subpopulations

TYPICAL STEPS OF GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)

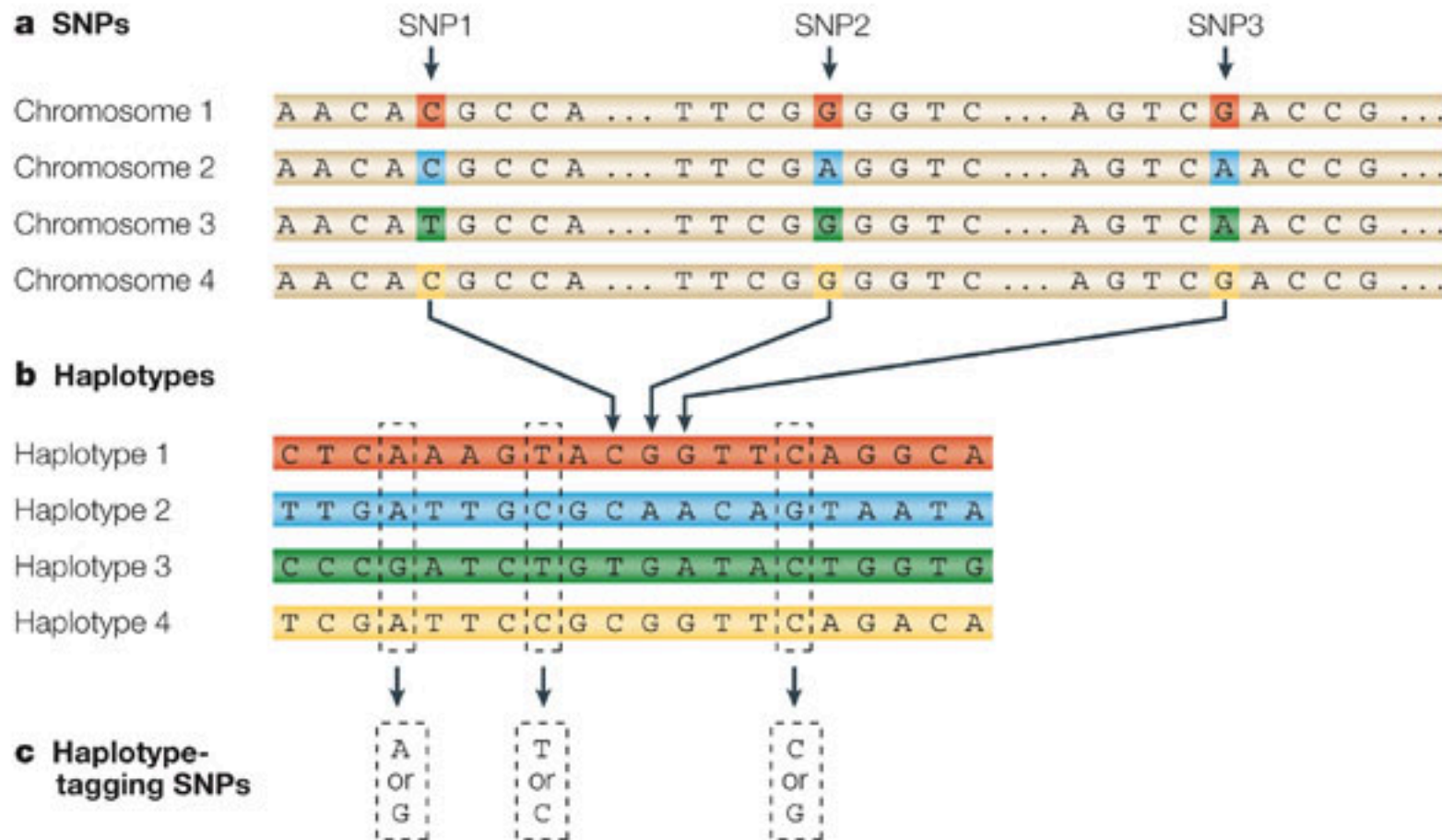
GENOTYPING NAÏVE APPROACH

- Identify all 10 million common SNPs
- Collect 1,000 cases and 1,000 controls
- Genotype all DNAs for all SNPs
- That adds up to 20 billion genotypes
- This won't work in practice:
 - Cost:
 - In 2002, this approach cost 50 cents a genotype.
 - That was \$10 billion for each disease – completely out of the question
 - Nowadays, 50 cents/2000 genotypes => \$500K per disease
 - Statistical:
 - Multiple test correction => lead to lower power => high rate of false negative

SOLUTION: SUB-SAMPLING HAPLOTYPE, LINKAGE DISEQUILIBRIUM & TAGSNPs

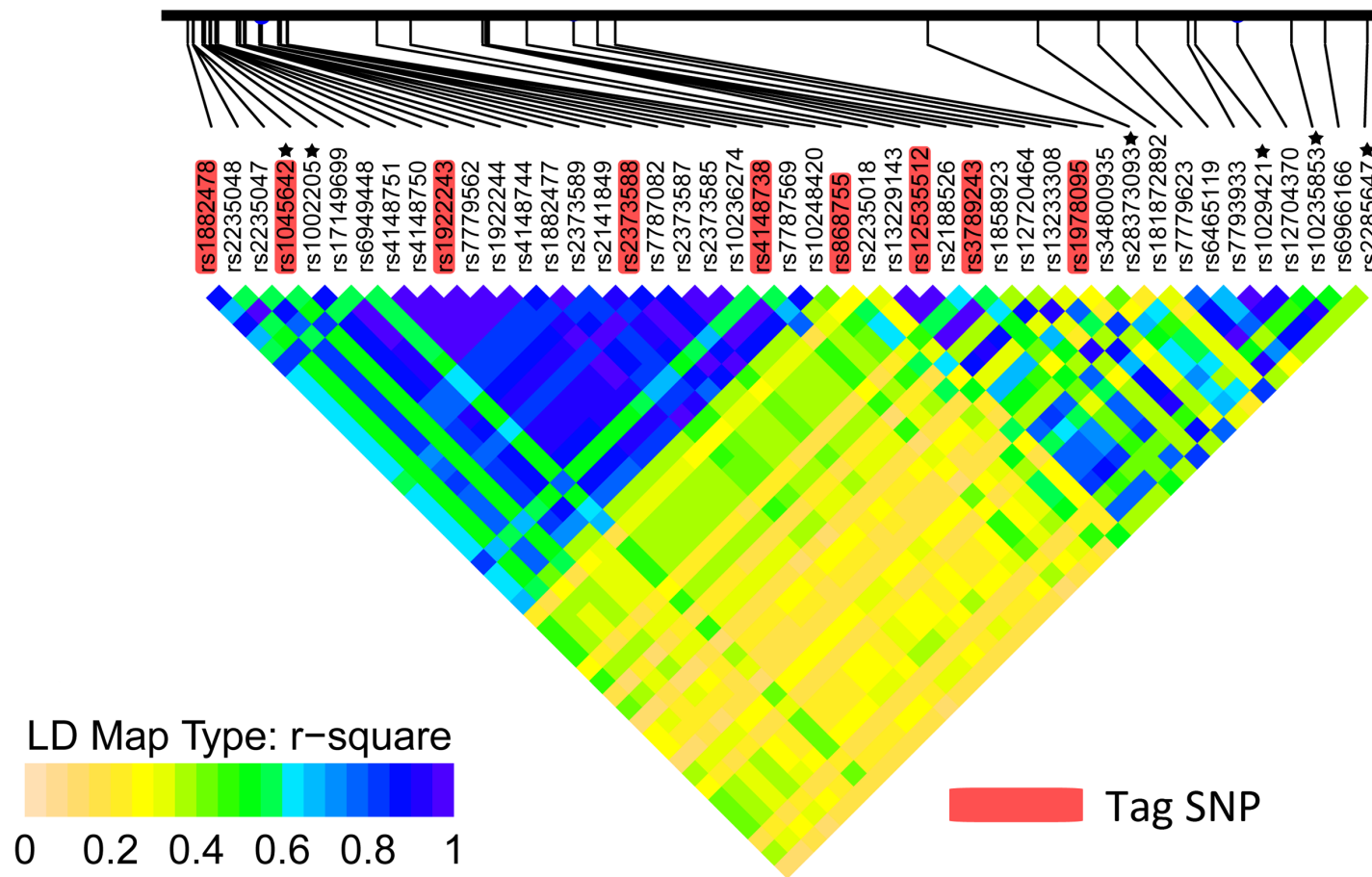
○ Haplotype: Genetic Home Reference

- A set of SNPs (mutations) on the same chromosome that tend to be inherited together
- SNPs can be highly correlated => sub-sampling is possible



DETECTION OF HAPLOTYPE: LINKAGE DISEQUILIBRIUM

- Linkage disequilibrium:
 - Nonrandom association of alleles at two or more loci



LINKAGE DISEQUILIBRIUM (LD)

At Equilibrium (independence)

$$P_{AB} = P_A P_B$$

At Disequilibrium (dependence)

$$P_{AB} \neq P_A P_B$$

Linkage Disequilibrium Coefficient D

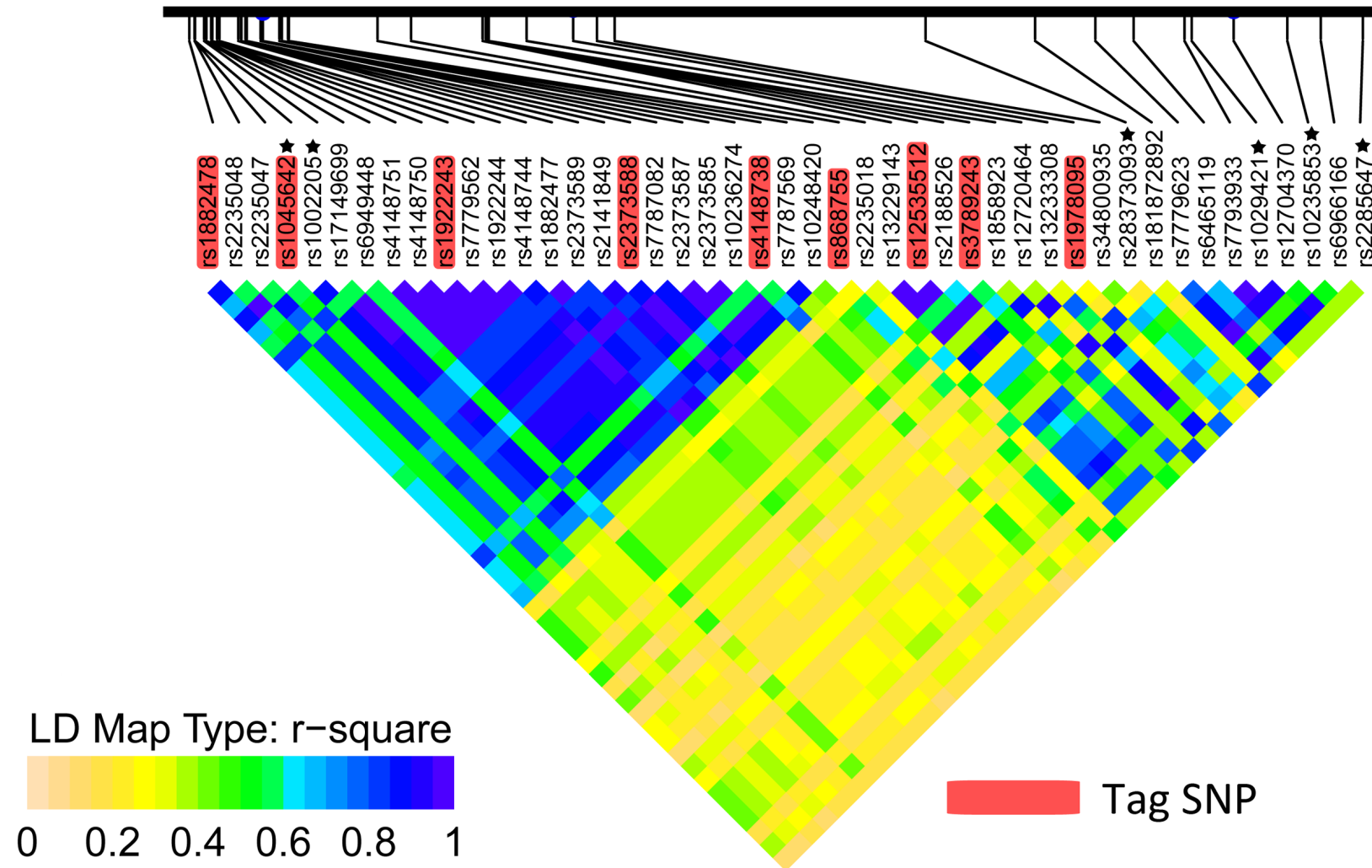
$$D = P_{AB} - P_A P_B$$

Linkage Disequilibrium r^2

$$r^2 = \frac{D^2}{p_A p_a p_B p_b} = \frac{D^2}{p_A (1 - p_A) p_B (1 - p_B)}$$

		<u>Locus B</u>		Totals
		<i>B</i>	<i>b</i>	
<u>Locus A</u>	<i>A</i>	p_{AB}	p_{Ab}	p_A
	<i>a</i>	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0

LINKAGE DISEQUILIBRIUM



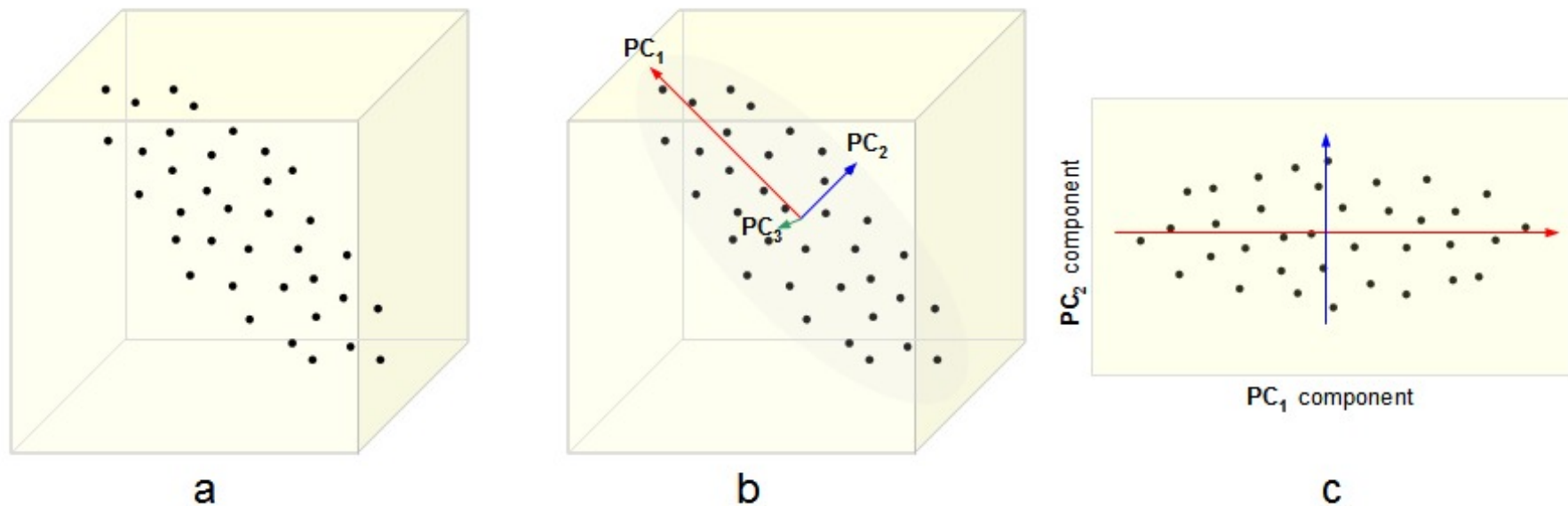
TAGSNPS (TAGGING SNPs)

- A SNP that can represent a group of SNPs
- Typical steps to identify tagSNPs:
 - Identify the search region
 - Define the metric for assessing the tagging
 - How well the tagSNP/tagSNPs predict their neighbors
 - Select an algorithm
 - Validate the performance of the learned tagSNPS

SELECTION OF TAGSNPs:

PRINCIPAL COMPONENT ANALYSIS (PCA)

- A orthogonal transformation to concise represent a set of data

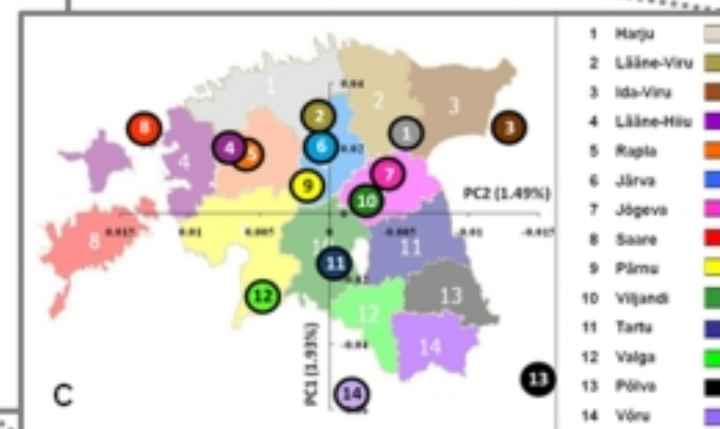
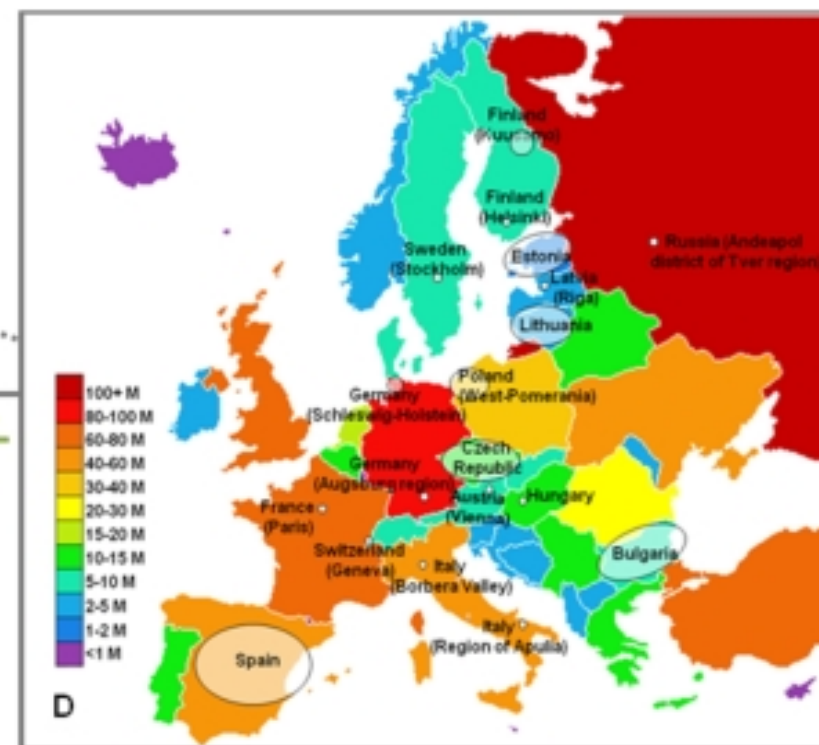
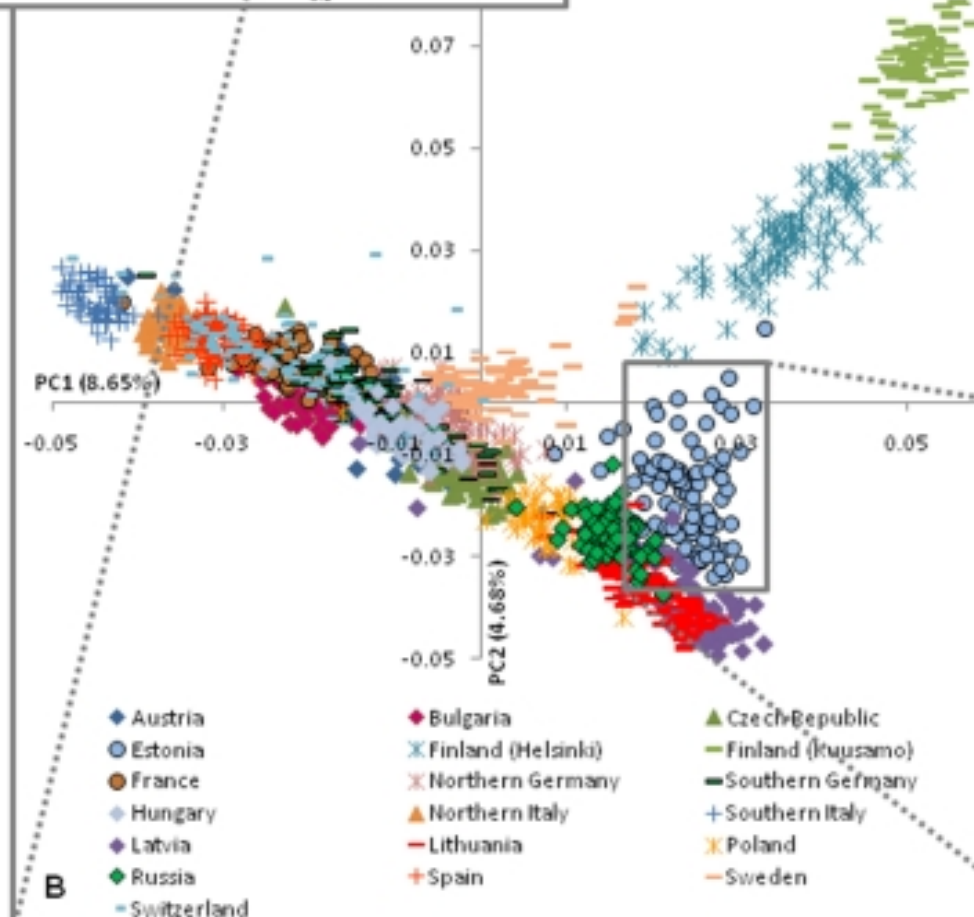
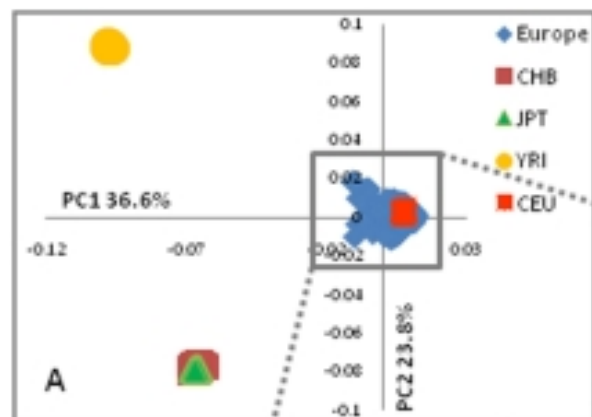


- Principal components are the eigenvectors of the covariance matrix $X^T X$

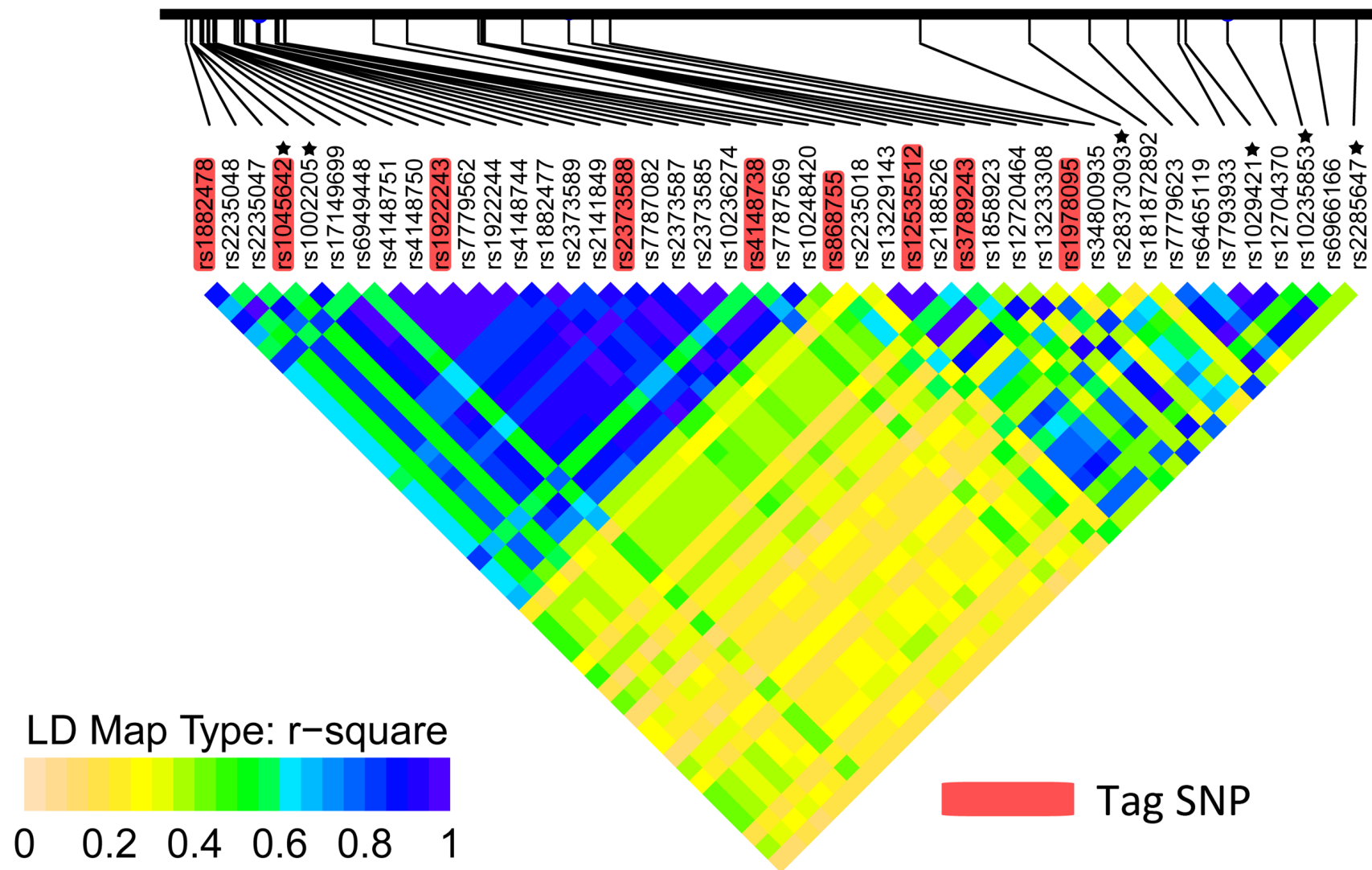
STEPS OF PRINCIPAL COMPONENT ANALYSIS

Give data X

- Step 1: Data normalization
- Step 2: Calculate eigenvalues and eigenvectors of $X^T X$
- Step 3: Sort the eigenvalues
- Step 4: Pick the top k eigenvalues and the corresponding eigenvectors (the principal components); one may adopt a threshold cut off selection strategy
→ W
- Step 5: Project your data onto the principal components
 - $T = XW$
- Alternative Approach



PCA FOR TAGSNPs



GENOTYPING WITH TAGSNPS

- Identify all 300K (instead of 10 million) tagSNPs
- Collect 1,000 cases and 1,000 controls
- Genotype all DNAs for all SNPs
- That adds up to 600 million genotype
 - Reduction in three magnitudes
 - Genotype costs can be reduced to just thousands of dollars

THE HAPMAP PROJECT

<http://hapmap.ncbi.nlm.nih.gov/index.html.en>



International HapMap Project

[Home](#) | [About the Project](#) | [Data](#) | [Publications](#) | [Tutorial](#)

中文 | [English](#) | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States. The project is a resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "[About the International HapMap Project](#)"

Project Information

[About the Project](#)
[HapMap Publications](#)
[HapMap Tutorial](#)
[HapMap Mailing List](#)
[HapMap Project Participants](#)

Project Data

HapMap Genome Browser release #28 (Phases 1, 2 & 3 - merged genotypes & frequencies)
[HapMap3 Genome Browser release #3 \(Phase 3 - genotypes & frequencies\)](#)

News

- 2013-06-14: **HapMap data conversion tool**

There are several inquiries for a conversion tool to convert HapMap data into the VCF format. Please see [Analysis Toolkit](#) (by Broad Institute).

- 2012-12-06: **Downtime for hardware maintenance**

From December 15 - 16, Hapmap site will be taken offline for an internal hardware maintenance. See [HapMap Downtime](#).

- 2011-06-13: **HapMap help desk announcement**

There was a problem with the HapMap help desk system. In the past several weeks, emails sent to the help desk were not reaching the help desk, and thus user requests were not addressed. Please resend your email requests to the HapMap help desk in the past several weeks. Sorry for the inconvenience.

TYPICAL STEPS OF GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)

STATISTICAL TESTING

- Effect size: Odds ratio
 - Ratio of odds

	Diseased	Healthy
Exposed	D_E	H_E
Not exposed	D_{NE}	H_{NE}

$$OR = \frac{D_E / H_E}{D_{NE} / H_{NE}}$$

- Statistical significance
 - Chi-2 test to obtain p-value
- For single hypothesis testing:
 - To control Type-I error (false positive) to be below 5%
 - p-value cut off at 0.05
- **BUT**, we are not testing only 1 SNP but 300K of them

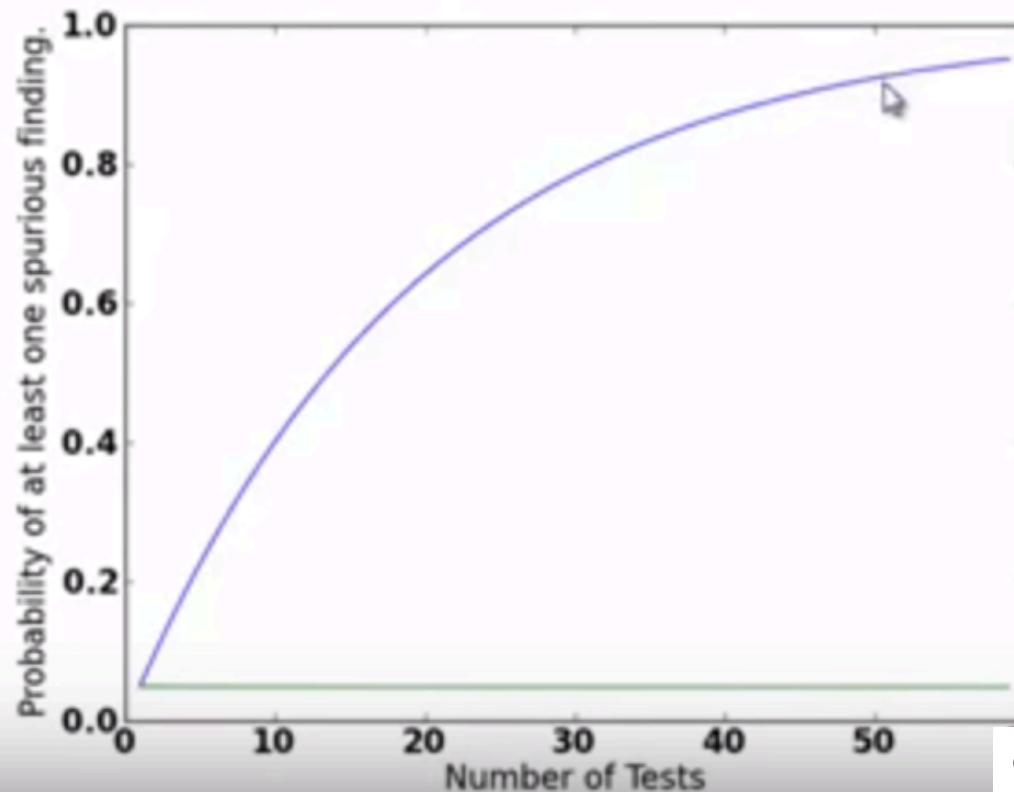
MULTIPLE TEST EFFECT

$P(\text{detecting an effect when there is none}) = \alpha = 0.05$

$P(\text{not detecting an effect when there is none}) = 1 - \alpha$

$P(\text{not detecting an effect when there is none in every experiment}) = (1 - \alpha)^k$

$P(\text{detecting an effect when there is none on at least one experiment}) = 1 - (1 - \alpha)^k$



$\alpha = 0.05$

"Familywise Error Rate"

MULTIPLE TEST CORRECTION

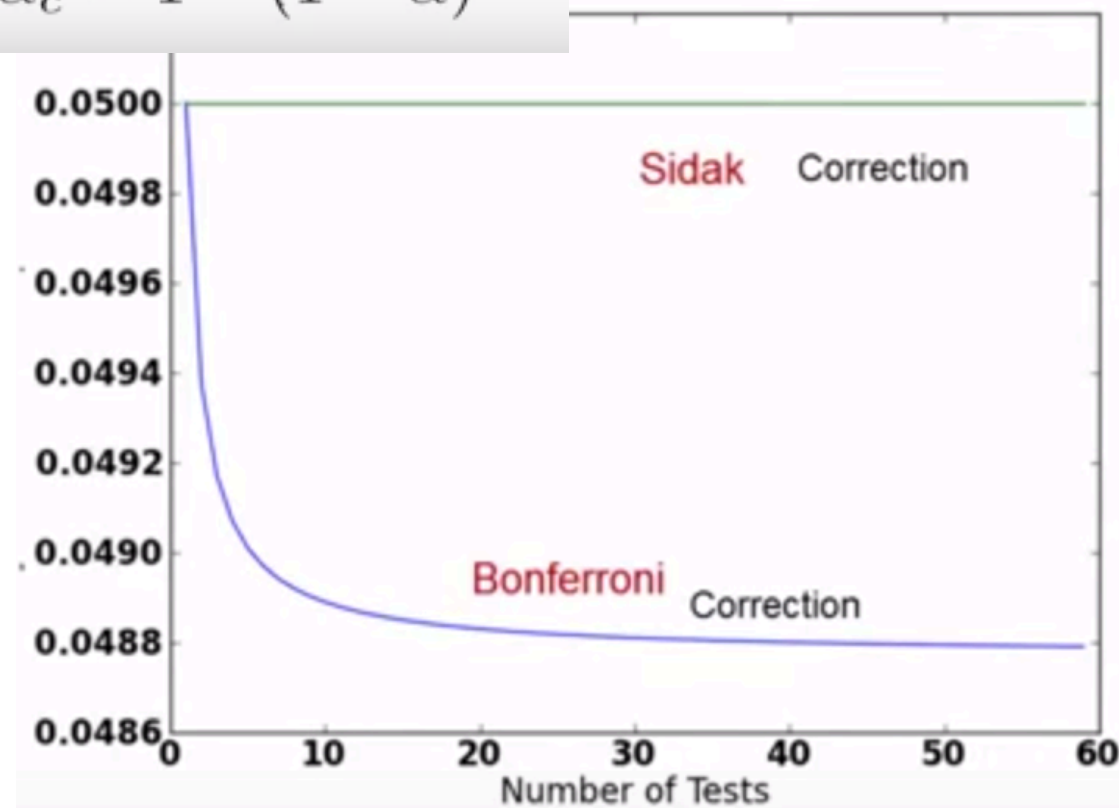
- Bonferroni Correction
 - Just divide by the number of hypotheses

$$\alpha_c = \frac{\alpha}{k}$$

- Šidák Correction
 - Asserts independence

$$\alpha = 1 - (1 - \alpha_c)^k$$

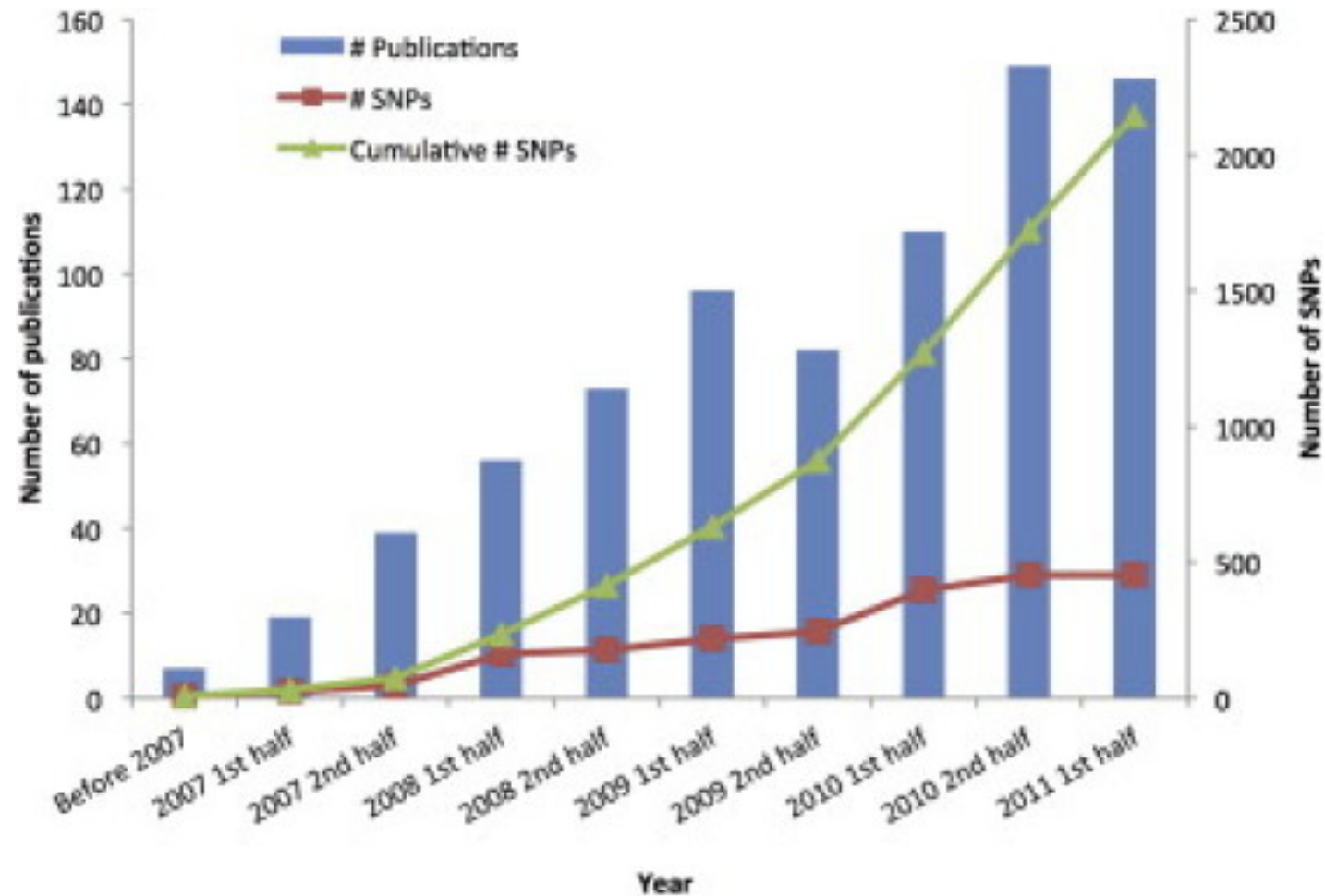
$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$



TYPICAL STEPS OF GWAS

- **Sampling** (Case-Control method)
- **Genotyping** (Data generation & collection)
- **Quality Control** (Data pre-processing)
- **Statistical Testing** (Data analysis)
- **Replication** (Verification)

GOLDEN AGE OF GWAS



THE VISION: PREVENTIVE MEDICATION

- Prevent disease from occurring
- Identify the cause of the disease
 - Genomics identifies the cause of disease
 - “All medicine may become pediatrics” Paul Wise, Professor of Pediatrics, Stanford Medical School, 2008
 - Treat the cause of the disease rather than the symptoms
- Health care costs can be greatly reduced if
 - Invests in preventive medicine
 - One targets the cause of disease rather than symptoms
- Challenges and limitations:
 - Penetrance and environmental factors

WEEK 8'S LEARNING OBJECTIVES

- After the class, students should be able to
 - Define Gene-disease association studies
 - Appreciate the motivations and applications of GWAS
 - Explain the differences between GWAS and Candidate Gene Studies
 - Explain the typical method and workflow for GWAS studies and, more importantly, considerations and limitations for each step
 - Understand the concepts of
 - Linkage Disequilibrium
 - Hypothesis testing
 - Multiple testing correction
 - Population stratification bias
 - Get to know the online resources