








Genomes and Complex Diseases

02-223 Personalized Medicine:
Understanding Your Own Genome
Fall 2014

Genome Polymorphisms

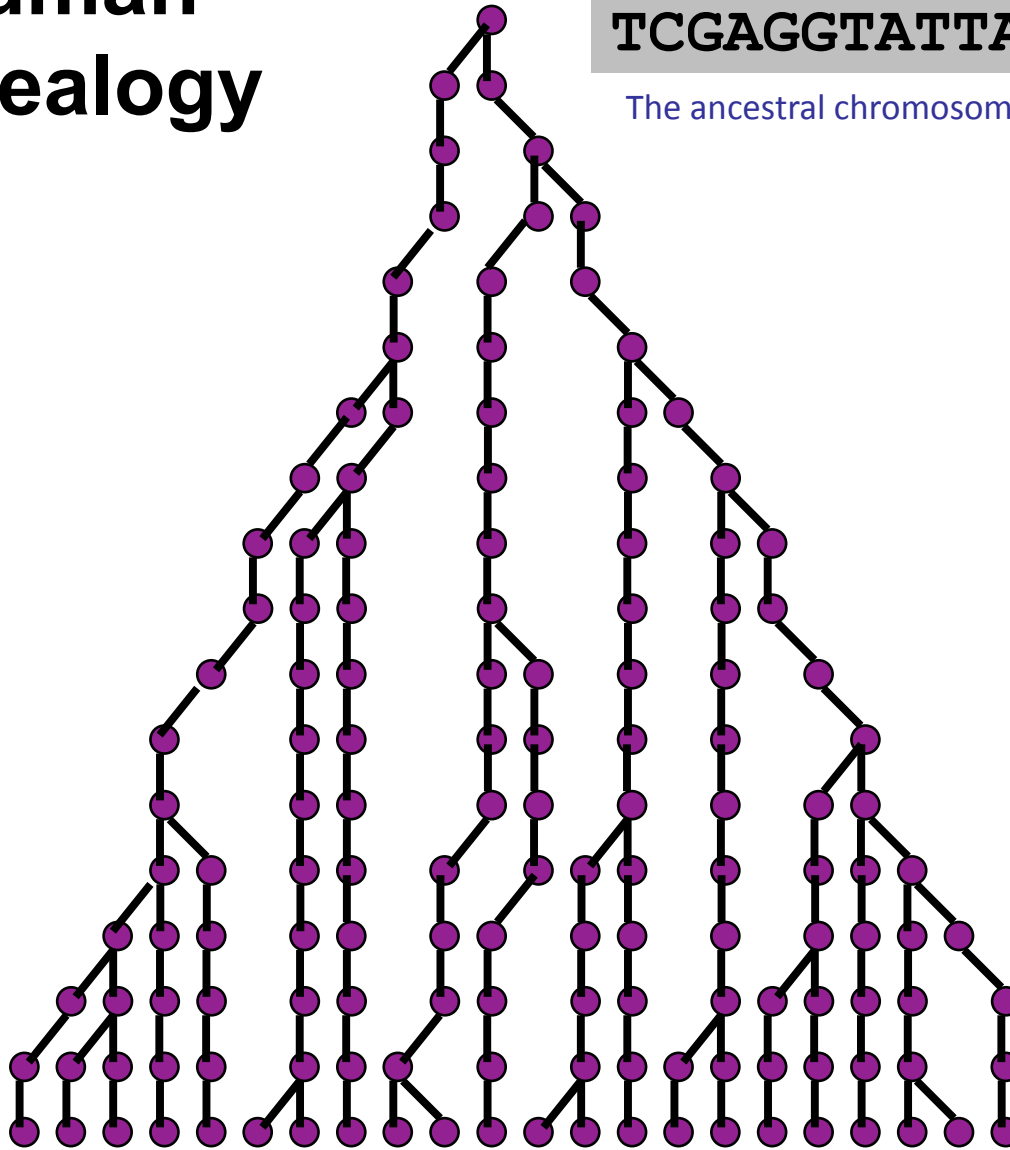


The ABO Blood System				
Blood Type (genotype)	Type A (AA, AO)	Type B (BB, BO)	Type AB (AB)	Type O (OO)
Red Blood Cell Surface Proteins (phenotype)	 A agglutinogens only	 B agglutinogens only	 A and B agglutinogens	 No agglutinogens
Plasma Antibodies (phenotype)	 b agglutinin only	 a agglutinin only	NONE No agglutinin	 a and b agglutinin

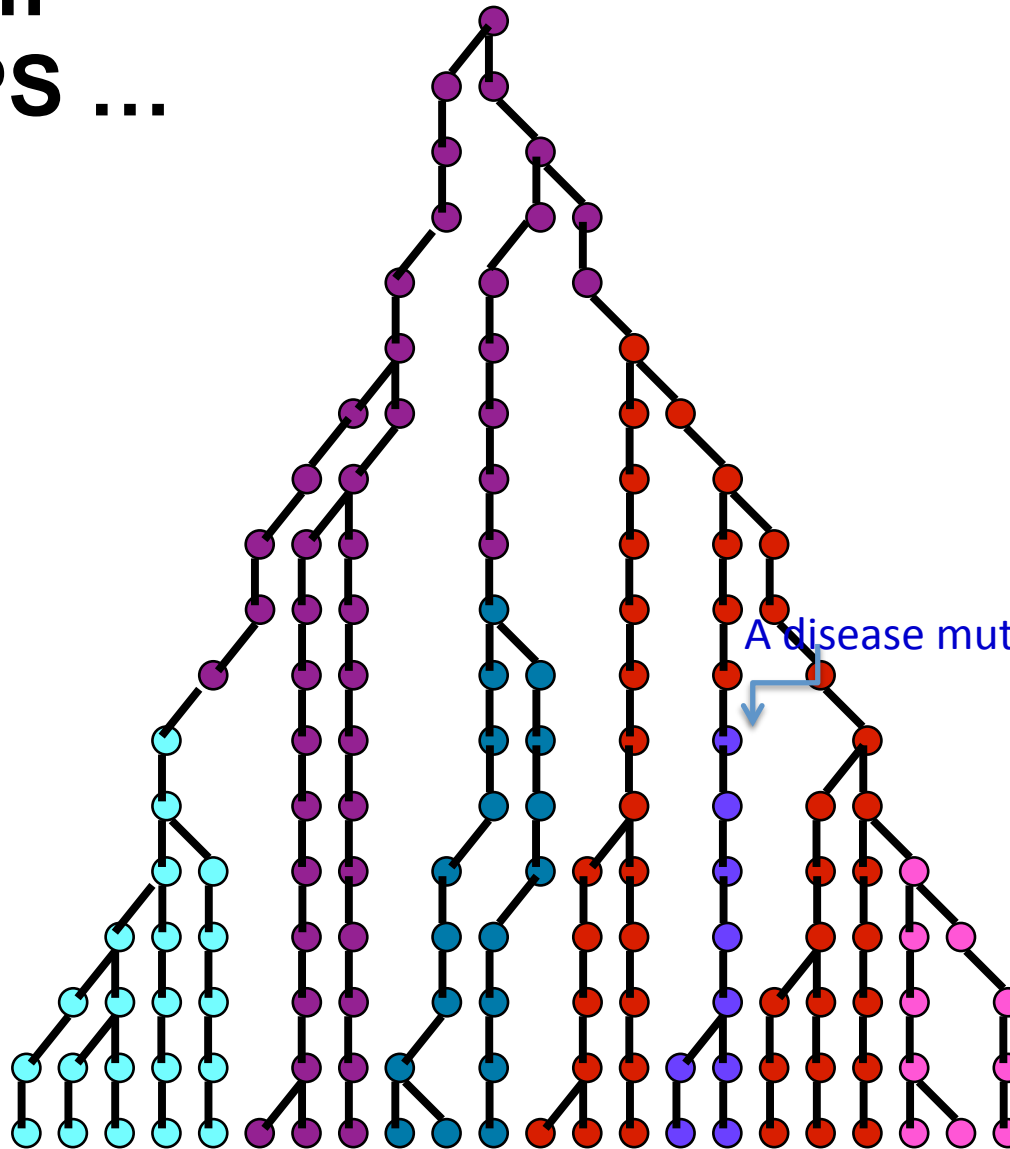
A Human Genealogy

TCGAGGTATTAAC

The ancestral chromosome



From SNPS ...



TCGAGGTATTAAAC
TCTAGGTATTAAAC
TCGAGGCATTAAAC
TCTAGGTGTTTAAAC
TCGAGGTATTAGC
TCTAGGTATCAAAC

* * * * *

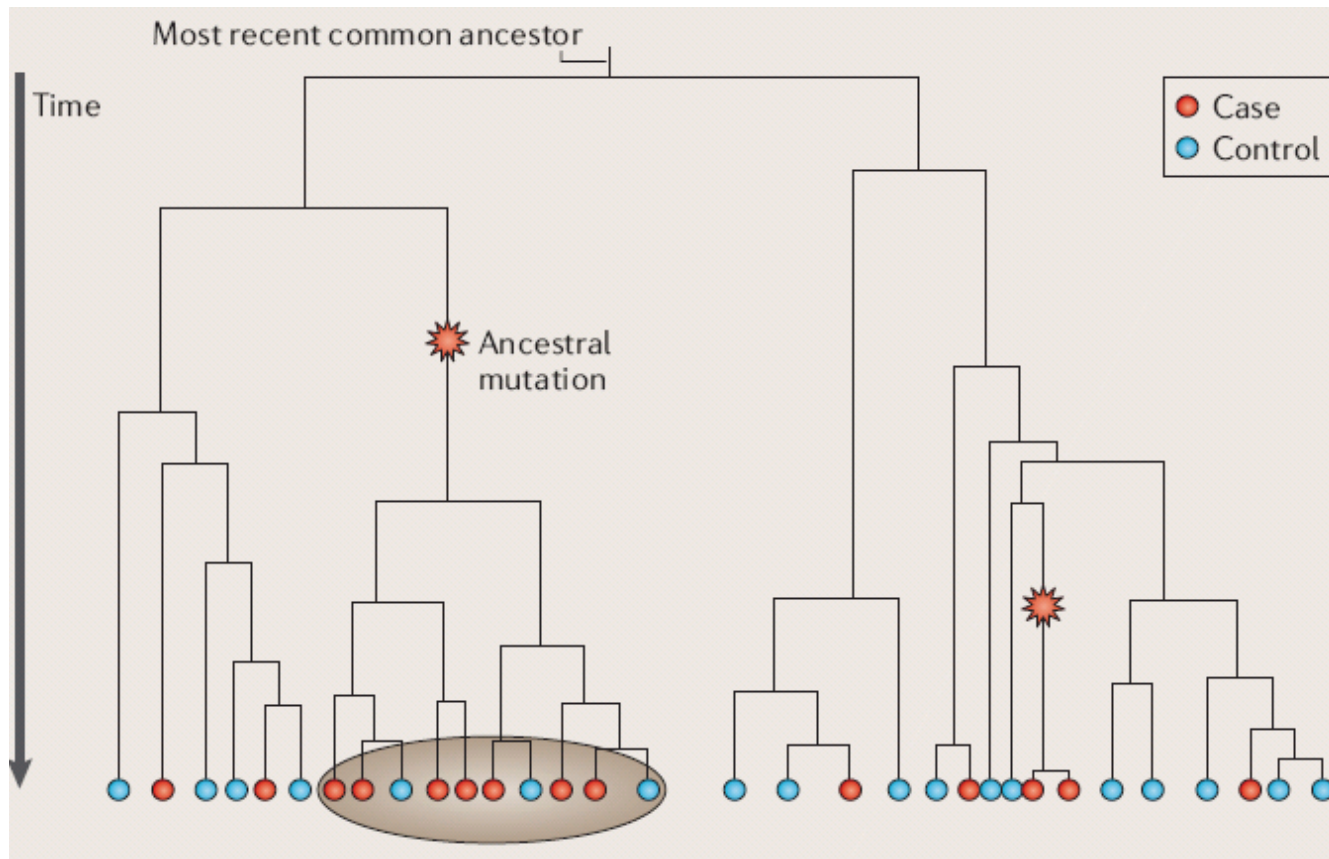
The SNPs

A disease mutation



Finding Disease Mutations

- Case/control data are collected from unrelated individuals
 - All individuals are related if we go back far enough in the ancestry



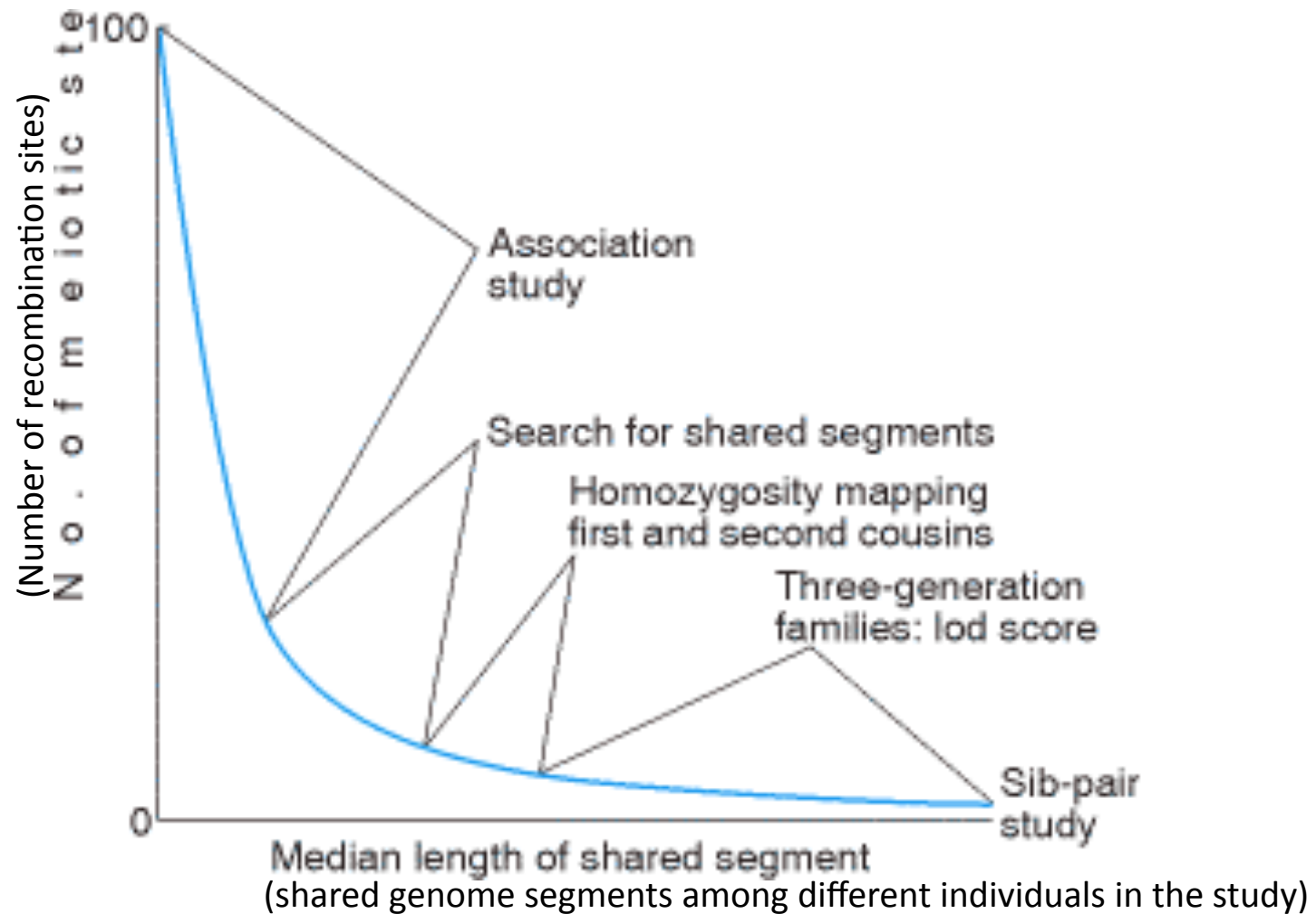
Family-Based and Population-Based Studies

- How can we identify disease-related genetic loci?
 - Linkage analysis
 - Data are collected for **family members**
 - Difficult to collect data on a large number of families
 - Effective for rare diseases
 - Genome-wide association studies (GWAS)
 - Data are collected for **unrelated individuals**
 - Easier to find a large number of affected individuals
 - Effective for common diseases, compared to family-based method

Family-Based and Population-Based Studies

- How do the LD patterns in genomes differ between family and population data?
 - Large or small linked regions?
 - Resolution for pinpointing the disease locus?

Linkage Analysis vs. Association Analysis

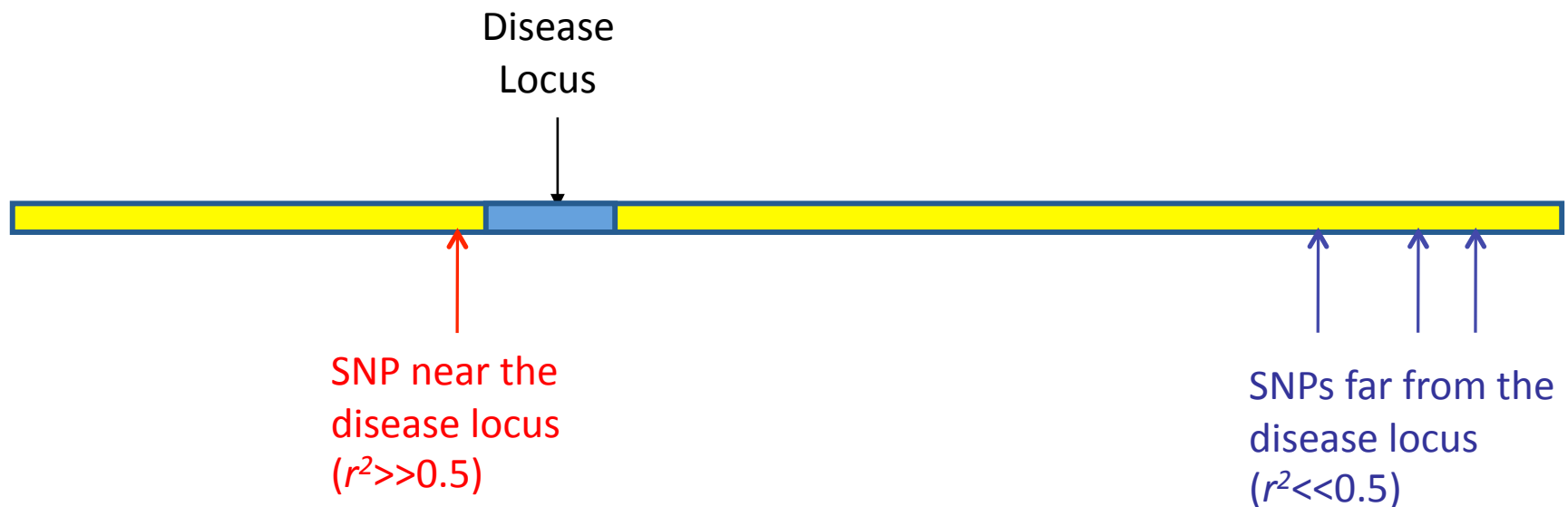


SNP Genotyping vs. Whole Genome Sequencing

- How would you identify disease mutations using whole genome sequencing technology?
- How would you identify disease mutations using SNP genotyping technology?

How Can We Identify the Unknown Disease Locus?

- Idea: Given a map of SNP genetic markers, let's look for the markers that are **linked** to the unknown disease locus (i.e. linkage between the disease locus and the marker locus)



Genome-Wide Association Study (GWAS)

- Data are collected for genotypes and phenotypes for a large number of unrelated individuals
 - Genotypes:
 - often SNP genotypes are used because of the ease of genotyping and abundance across genomes.
 - For SNPs, minor allele homozygous, heterozygous, and major allele heterozygous sites are coded as 0, 1, and 2.
 - Phenotypes:
 - Categorical data (e.g., case/control labels for individuals)
 - Continuous-valued data (e.g., height, cholesterol level, blood IgE level)

Overview

- Case control studies
 - Discrete phenotype
 - Are you a healthy normal or a patient?
- Quantitative trait studies
 - Continuous-valued phenotypes
 - Height, eye color, blood pressure, cholesterol level, body-mass index etc.

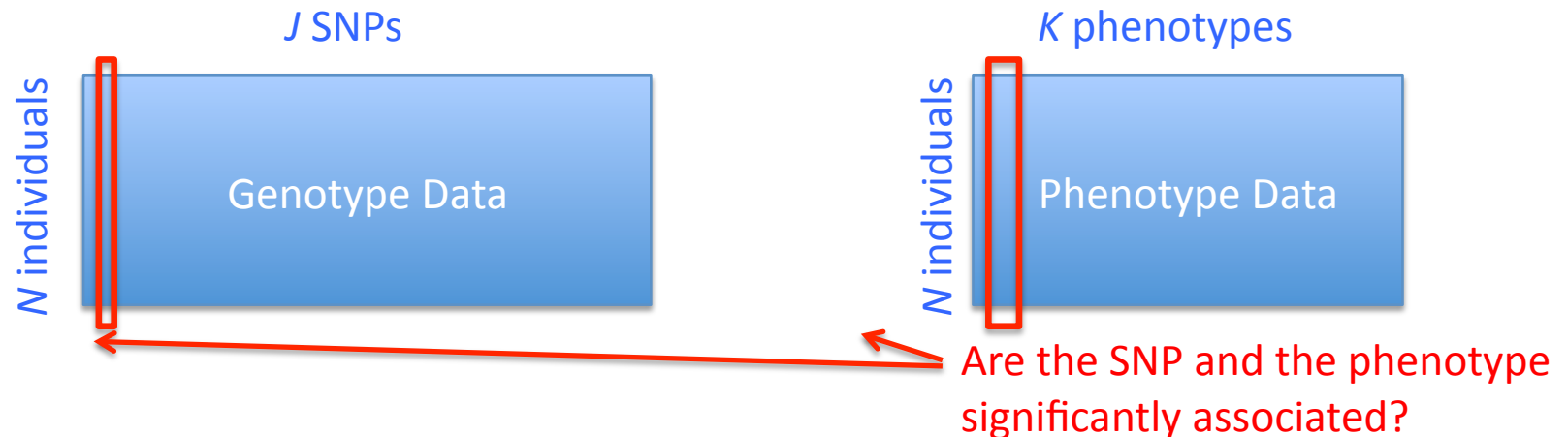
Genome-Wide Association Study (GWAS)

- Data collected for GWAS can be represented as two matrices

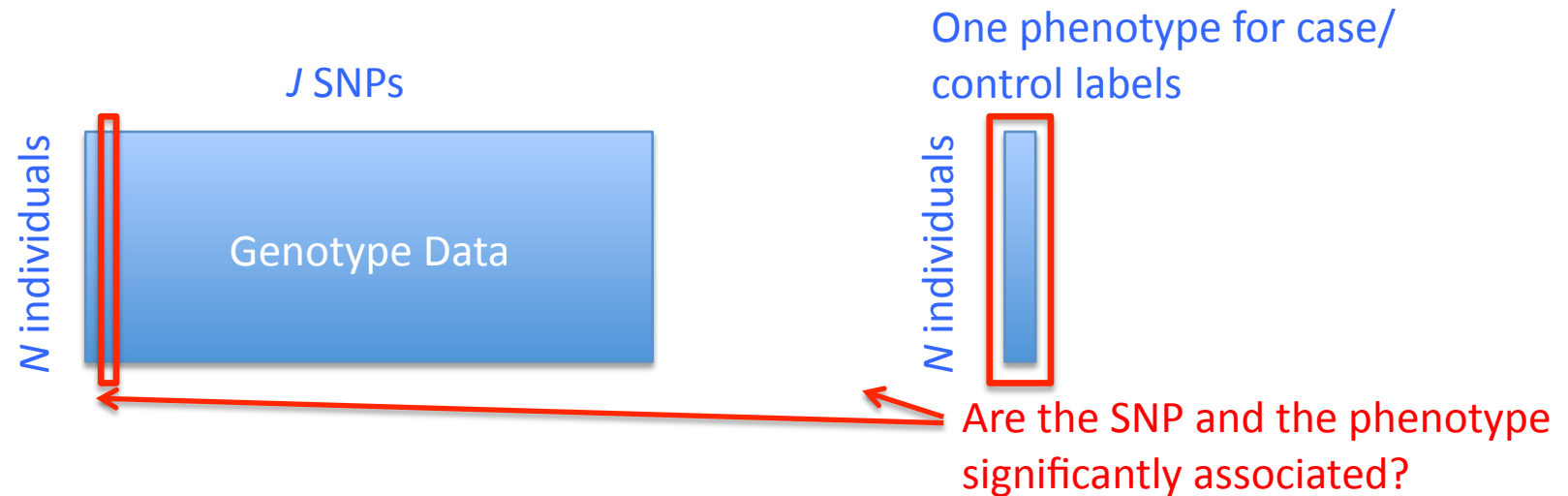


Genome-Wide Association Study (GWAS)

- For each SNP and each phenotype, perform a statistical test for “association”
- Repeat this for all (SNP, phenotype) pairs
- Identify the (SNP, phenotype) pairs with “significant” association.
 - The genome region around the SNP is likely to influence the phenotype



GWAS: Case/Control Study

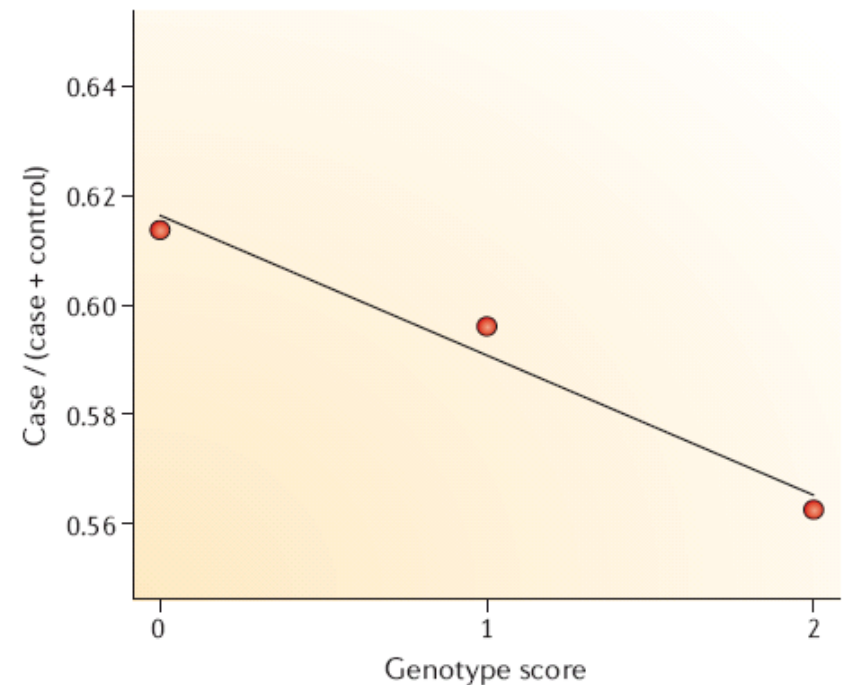


GWAS: Case/Control Study

- For each marker locus, find the 3x2 contingency table containing the counts of three genotypes

Genotype	Case	Control	Total
AA	$N_{\text{case,AA}}$	$N_{\text{control,AA}}$	N_{AA}
Aa	$N_{\text{case,Aa}}$	$N_{\text{control,Aa}}$	N_{Aa}
aa	$N_{\text{case,aa}}$	$N_{\text{control,aa}}$	N_{aa}
Total	N_{case}	N_{control}	N

- χ^2 test with 2 df (degree of freedom) under the null hypothesis of no association



Genotype score = the number of minor alleles

GWAS: Case/Control Study

- Alternatively, assume an additive model, where the heterozygote risk is approximately between the two homozygotes
- Form a 2x2 contingency table. Each individual contributes twice from each of the two chromosomes.

Allele Type	Case	Control	Total
A	$G_{\text{case,A}}$	$G_{\text{control,A}}$	G_A
a	$G_{\text{case,a}}$	$G_{\text{control,a}}$	G_a
Total	$2 \times N_{\text{case}}$	$2 \times N_{\text{control}}$	$2N$

- χ^2 test with 1df

χ^2 Test (Chi-square Test)

- Statistical test of association
- In case/control association study, the null hypothesis is
 H_0 : There is no association between the given marker and disease labels.
- P-value = probability of the observed data under the null hypothesis
- Low p-value (p-value < α , where α is a user-specified value) means the observed data are unlikely under the null hypothesis. Thus, we reject the null hypothesis (H_0) and declare there is a significant association between the SNP and disease states.
 - Often $\alpha=0.01$ or 0.05 is used.

Chi-Square Test: Null Hypothesis from Contingency Table

- We have two random variables:
 - Y: disease status (Case/Control)
 - X: allele type (A/a)
- Null hypothesis: the two variables are independent of each other (i.e., the two variables are unrelated)

Allele Type	Case	Control	Total
A	$G_{\text{case},A}$	$G_{\text{control},A}$	G_A
a	$G_{\text{case},a}$	$G_{\text{control},a}$	G_a
Total	$2 \times N_{\text{case}}$	$2 \times N_{\text{control}}$	$2N$

Chi-Square Test

Observations

Allele Type	Case	Control	Total
A	$G_{\text{case,A}}$	$G_{\text{control,A}}$	G_A
a	$G_{\text{case,a}}$	$G_{\text{control,a}}$	G_a
Total	$2xN_{\text{case}}$	$2xN_{\text{control}}$	$2N$

Expected values

Allele Type	Case	Control	Total
A	?	?	G_A
a	?	?	G_a
Total	$2xN_{\text{case}}$	$2xN_{\text{control}}$	$2N$

Chi-square statistic

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed frequency for i^{th} outcome

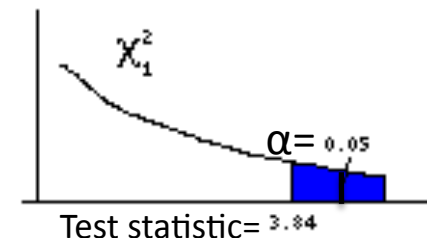
(the value can be read off of the contingency table)

E_i = expected frequency for i^{th} outcome

(the value can be obtained as described in the previous slides)

n = total number of outcomes

The probability distribution of this statistic is given by the chi-square distribution.



Using chi-square test, we can test how well observed values fit expected values computed under the independence hypothesis

Chi-Square Test: How to Compute Expected Values

- Under the null hypothesis of independence
 - $P(Y=\text{case and } X=A) = P(Y=\text{case})P(X=A)$
- Under the null hypothesis, the expected number of cases with allele A is
 - $2N \times P(Y=\text{case})P(X=A)$
 - where N is total observations and

$$P(Y=\text{case}) = (G_{\text{case},A} + G_{\text{case},a}) / (2N)$$

$$P(X=A) = (G_{\text{case},A} + G_{\text{control},A}) / (2N)$$
- Similarly
 - What is the expected number of cases with allele a?
 - What is the expected number of controls with allele A?
 - What is the expected number of controls with allele a?
- Do the probabilities sum to 1?

Observations

Allele Type	Case	Control	Total
A	$G_{\text{case},A}$	$G_{\text{control},A}$	G_A
a	$G_{\text{case},a}$	$G_{\text{control},a}$	G_a
Total	$2xN_{\text{case}}$	$2xN_{\text{control}}$	$2N$

Expected values

Allele Type	Case	Control	Total
A	?	?	G_A
a	?	?	G_a
Total	$2xN_{\text{case}}$	$2xN_{\text{control}}$	$2N$

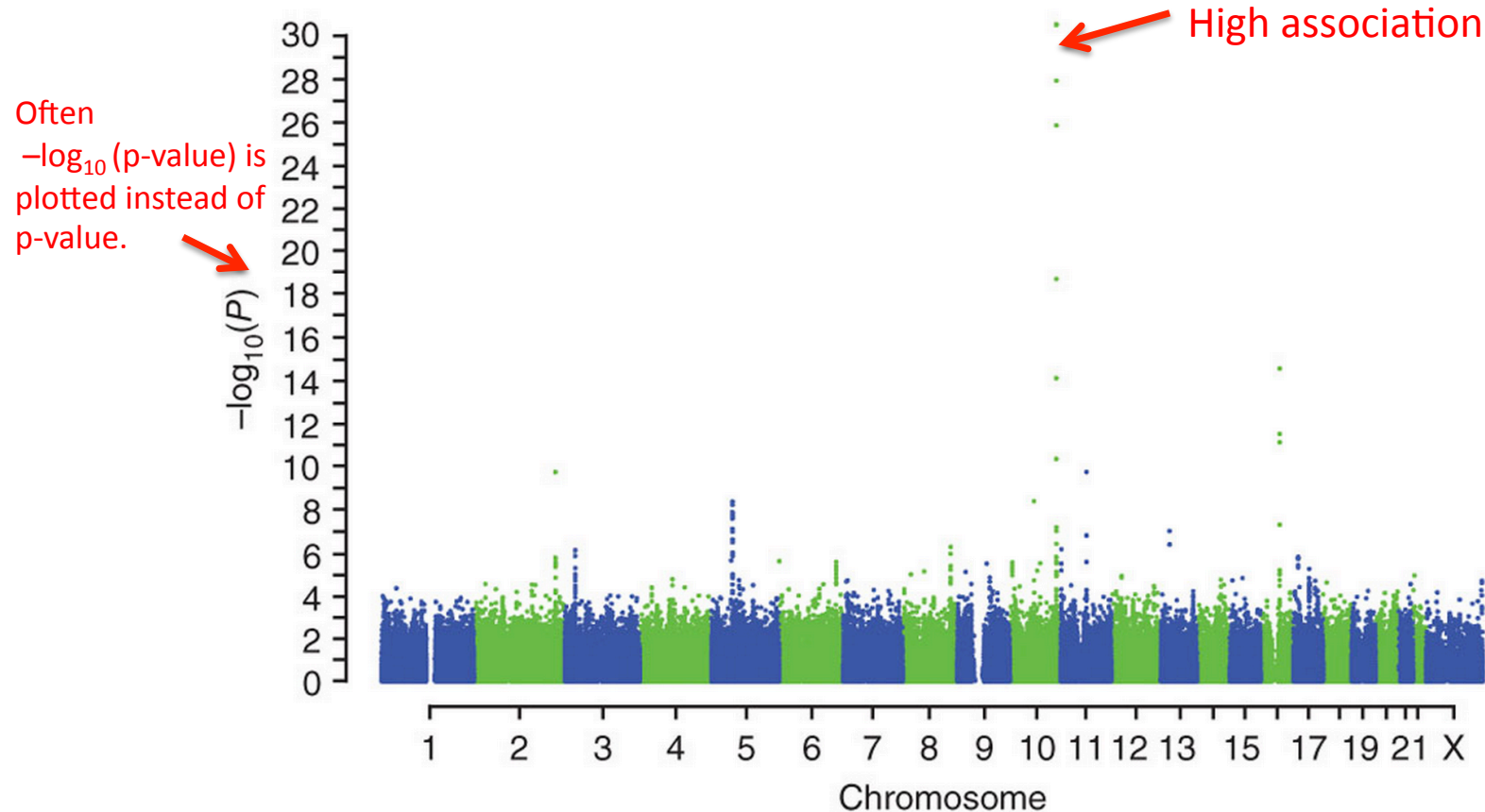
Chi-square Statistic

- Compute expected values
- Compute chi-square statistic
- Compute chi-square p-value by referring to chi-square distribution

	Case	Control
A	16	36
a	2	48

Manhattan Plot of p-values from Breast Cancer GWAS

- Analysis of 582,886 SNPs for 3,659 cases with family history and 4,897 controls



Correcting for Multiple Testing

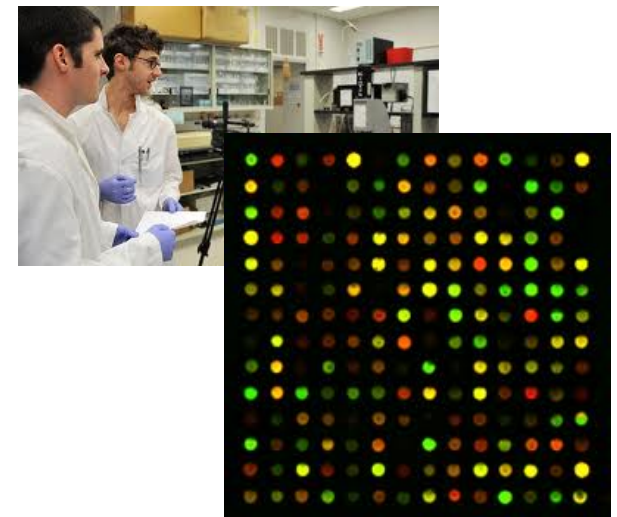
- What happens when we scan the genome of 1 million markers for association with $\alpha = 0.05$?
 - 50,000 (=1 million \times 0.05) SNPs are expected to be found significant just by chance
 - We need to be more conservative when we decide a given marker is significantly associated with the trait.
- Correction methods
 - Bonferroni correction

Overview

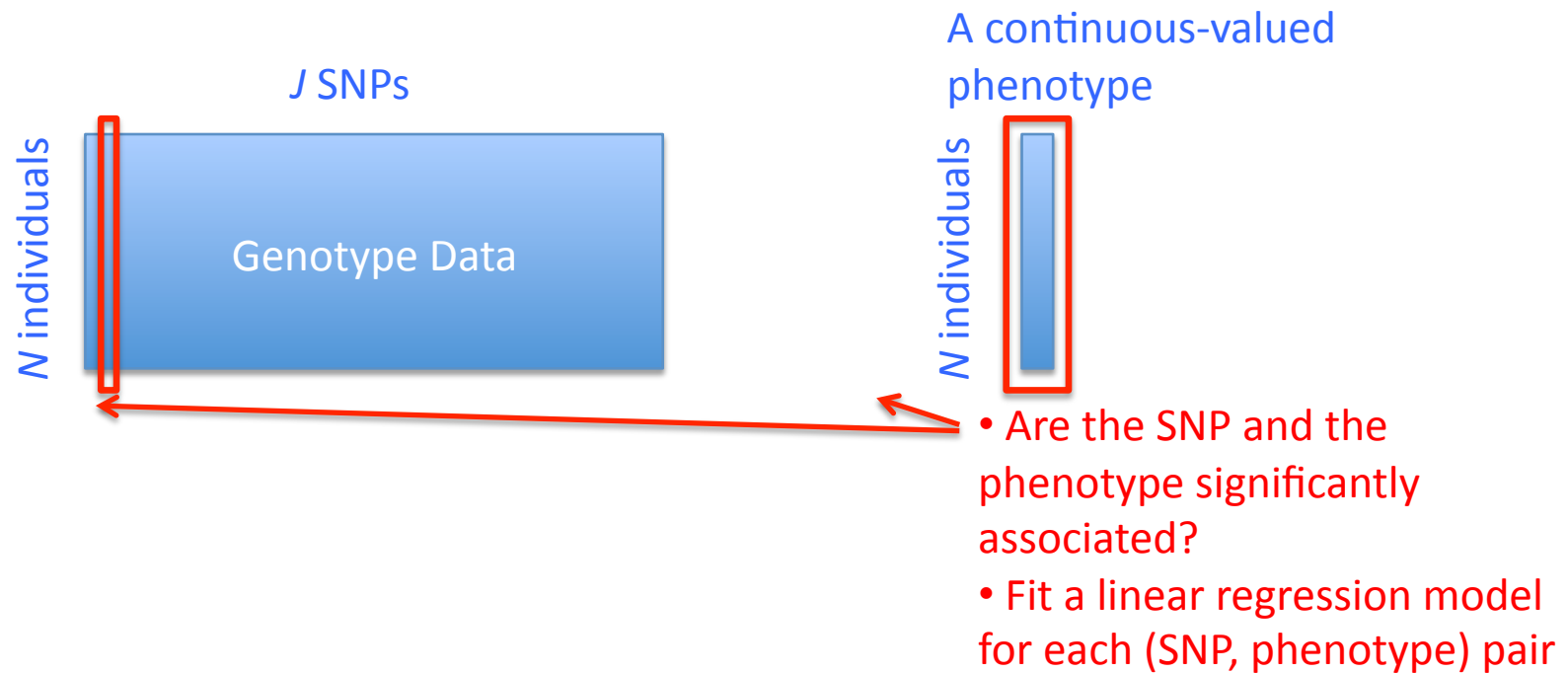
- Case control studies
 - Discrete phenotype
 - Are you a healthy normal or a patient?
- Quantitative trait studies
 - Continuous-valued phenotypes
 - Height, eye color, blood pressure, cholesterol level, body-mass index etc.

Continuous-Valued Phenotypes

- Clinical data in medicine
 - Cholesterol level, body-mass index, and weight in obesity, diabetes studies
 - Blood IgE antibody level, lung physiology measurements in asthma studies
- Gene-expression data in biology research
 - Microarray data for gene-expression measurements for tens of thousands of genes
 - Can we identify SNPs that influence the gene expression levels?
 - Often known as expression quantitative trait locus (eQTL) mapping

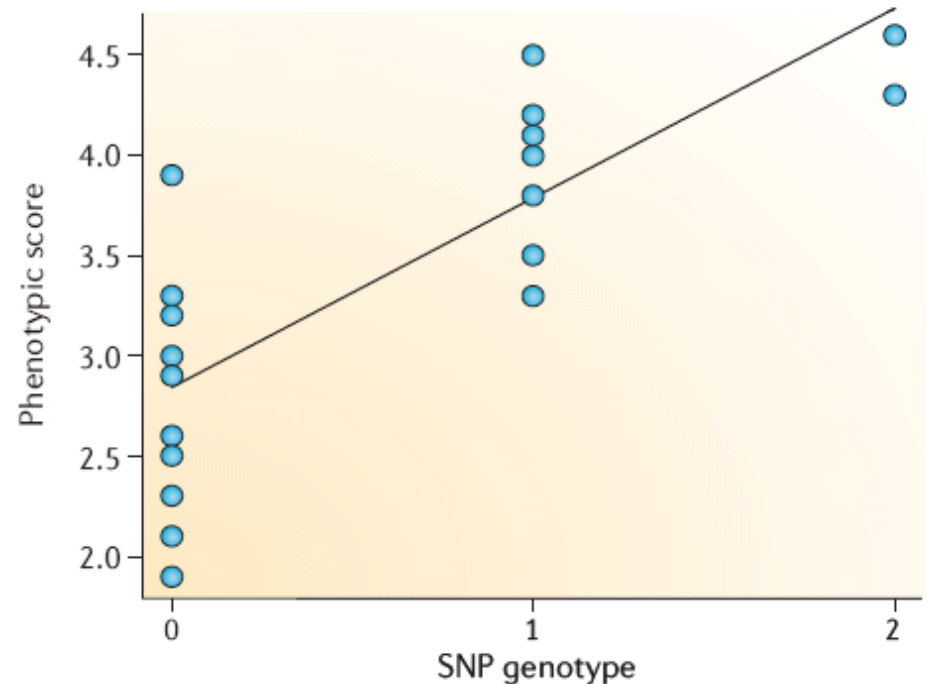


GWAS: Continuous-Valued Phenotypes



GWAS: Continuous-valued Phenotypes

- Continuous-valued traits
 - Also called quantitative traits
 - Cholesterol level, blood pressure etc.
- One cannot create a contingency matrix as in case/control studies
- For each locus, fit a linear regression using the number of minor alleles at the given locus of the individual as covariate



Linear Regression Model

- Linear regression model is defined as

$$y = x\beta_1 + \beta_0 + \varepsilon$$

- Data

- y : a continuous-valued phenotype
- x : SNP genotype at a given locus

Linear Regression Model

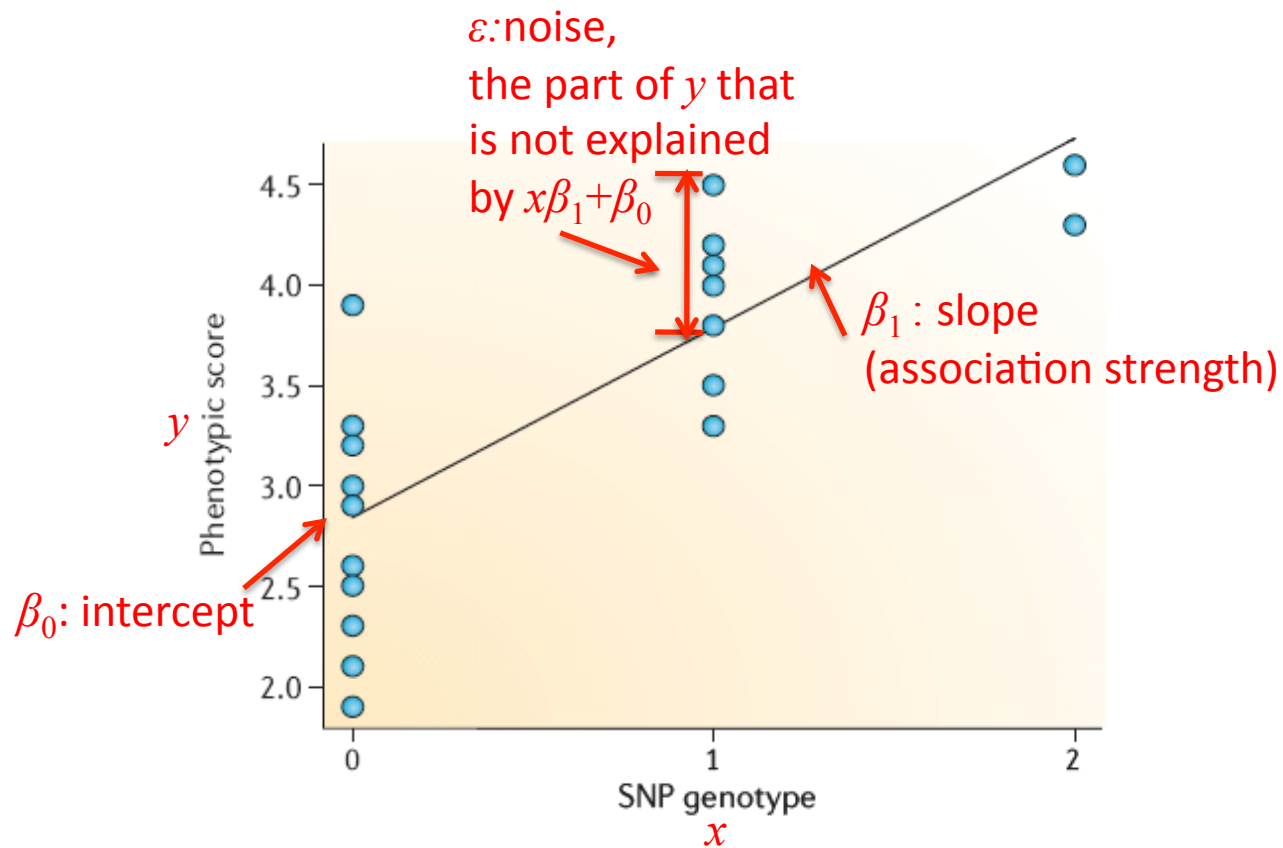
- Linear regression model is defined as

$$y = x\beta_1 + \beta_0 + \varepsilon$$

- Parameters

- β_1 : regression coefficient or the parameter that represents the strength of association between the SNP x and the phenotype y
- β_0 : intercept term
- ε : noise or the part of y that is not explained by the SNP x (e.g., environmental effect)

Linear Regression Model for GWAS



Linear Regression Model

- Linear regression model is defined as

$$y = x\beta_1 + \beta_0 + \varepsilon$$

- y and x : observed in the dataset
- β_1 and β_0 : parameters to be estimated from the data

Least Square Method for Parameter Estimation

- Popular method for parameter estimation is a least square method.

- Given data for N samples $(y_1, \dots, y_N), (x_1, \dots, x_N)$, solve the following problem

$$\operatorname{argmin} \sum_{i=1}^N (y_i - x_i\beta_1 - \beta_0)^2$$

- By solving this, we find the parameter values that minimize the squared distance between observed phenotype value (y_i 's) and predicted phenotype value $x_i\beta_1 + \beta_0$

Bonferroni Correction

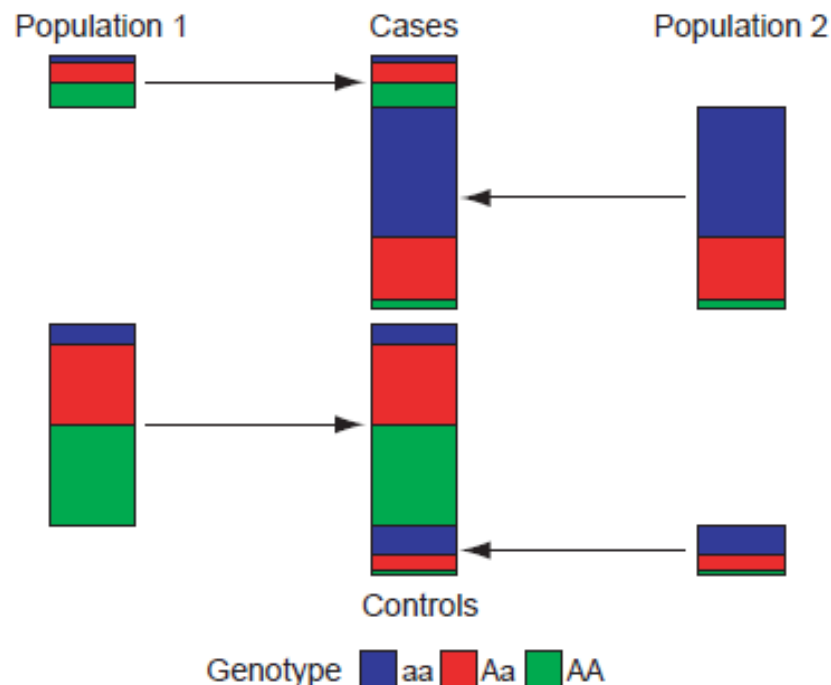
- If N markers are tested, we correct the significance level as $\alpha' = \alpha/N$
 - Assumes the N tests are independent, although this is not true because of the linkage disequilibrium.
 - Overly conservative for tightly linked markers

Population Structure and Genome-wide Association Analysis

- The mutation that gives the lactose persistence phenotype is more common in Caucasian population than in Asian population
- The allele for blonde hair color is also more common in Caucasian population than in Asian population

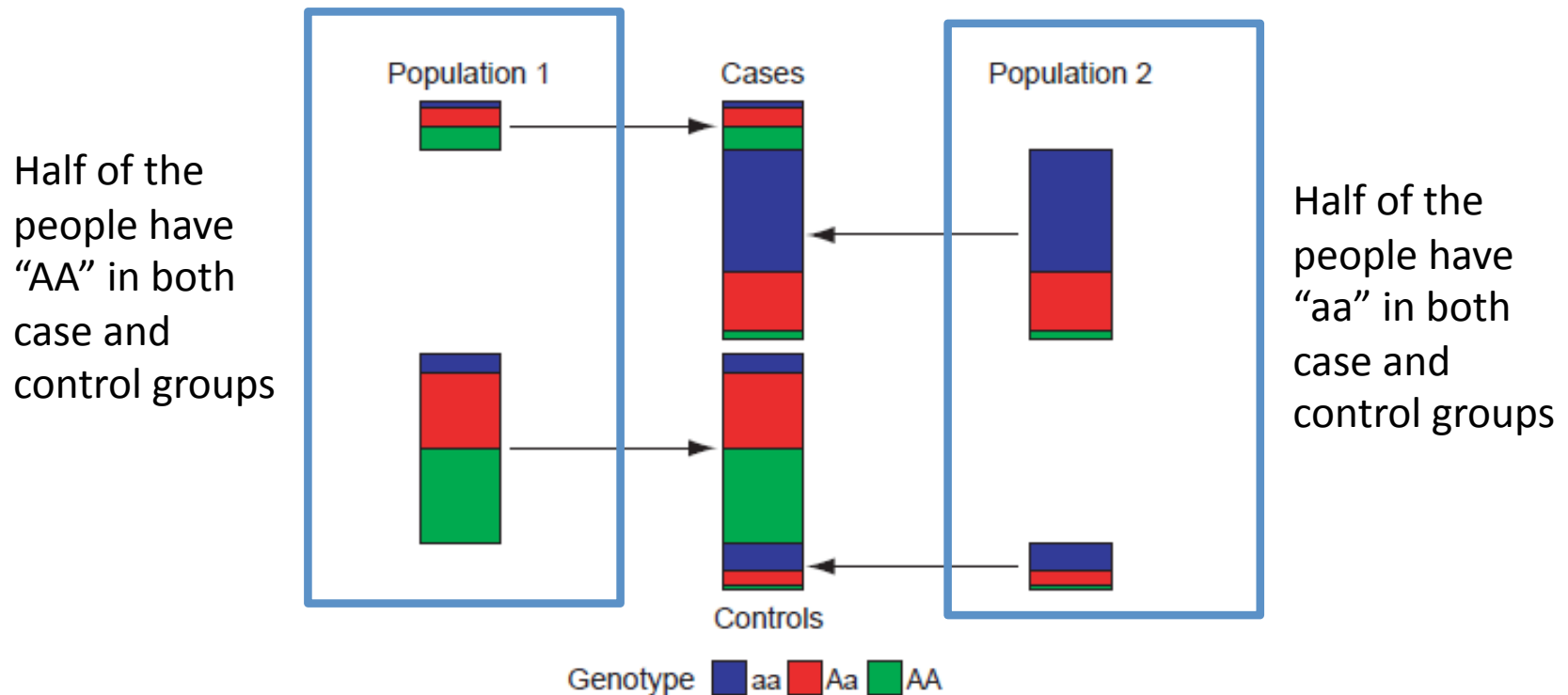
Population Structure and Genome-wide Association Analysis

- Population structure in data causes false positives in GWAS
 - If samples in the case group are more related (come from the same population group), any SNPs more prevalent in the case population will be found significantly associated with the trait.



Population Structure and Genome-wide Association Analysis

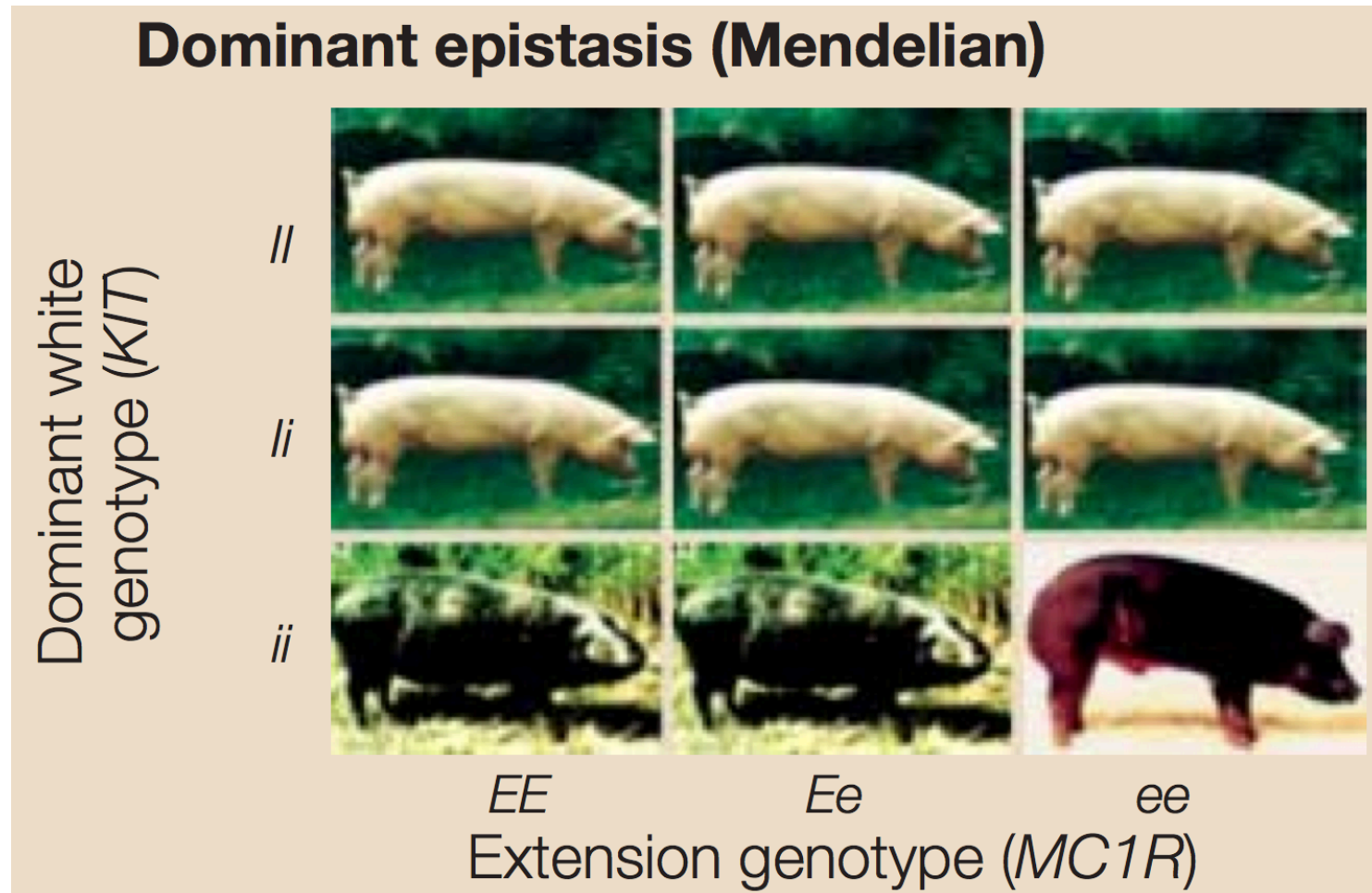
- What if we perform GWAS **within** each population groups.



Epistasis

- Epistasis: The effect of one locus depends on the genotype of another locus
 - Epistatic effects of genetic loci can be detected only if we consider the multiple loci jointly
- In contrast, marginal effects of a locus refers to the genetic effect of the locus that is independent of other loci
 - Most studies assume the phenotype can be predicted as a sum of single-locus effects
- Many studies ignore epistasis among multiple genetic loci mainly due to the high computational cost for detecting it, but epistasis is believed to be prevalent and thus important.

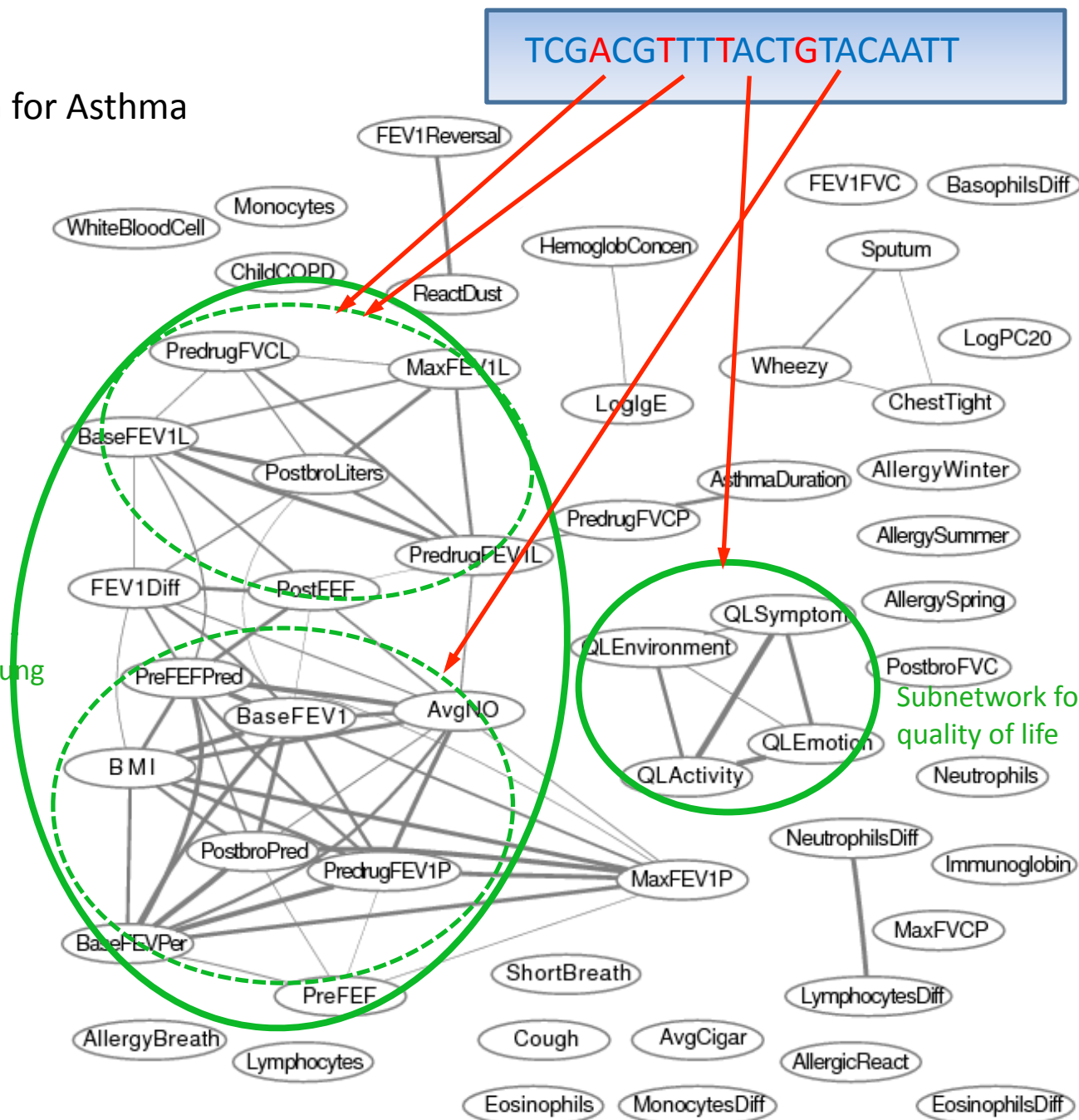
Epistasis for Mendelian Traits



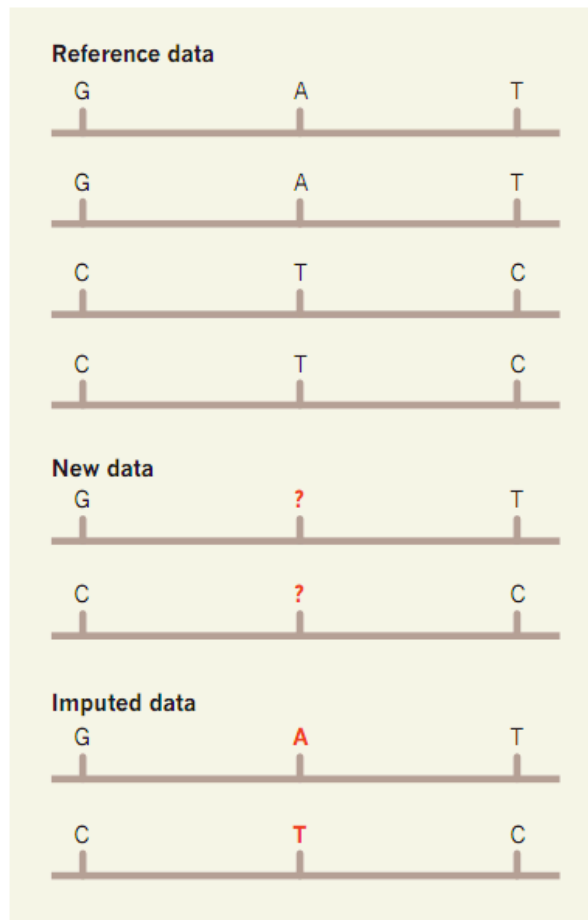
Genetic Association for Asthma Clinical Traits

Subnetworks for lung
physiology

Subnetwork for
quality of life



Tag SNPs and GWAS: Using Reference Datasets for Genotype Imputation



- Reference data: dense SNP data from HapMap III, or 1000 genome project
- New data: SNP data for individuals in a given study
- Data after imputation

Genotype Imputation

Reference set of haplotypes, for example, HapMap

0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

Genotype data with missing data at untyped SNPs (grey question marks)

1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

0	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
1	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
...
1	?	?	?	1	?	1	?	0	1	0	?	?	1	?	0
1	?	?	?	1	?	1	?	1	1	1	?	?	1	?	0
...
1	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0
0	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0

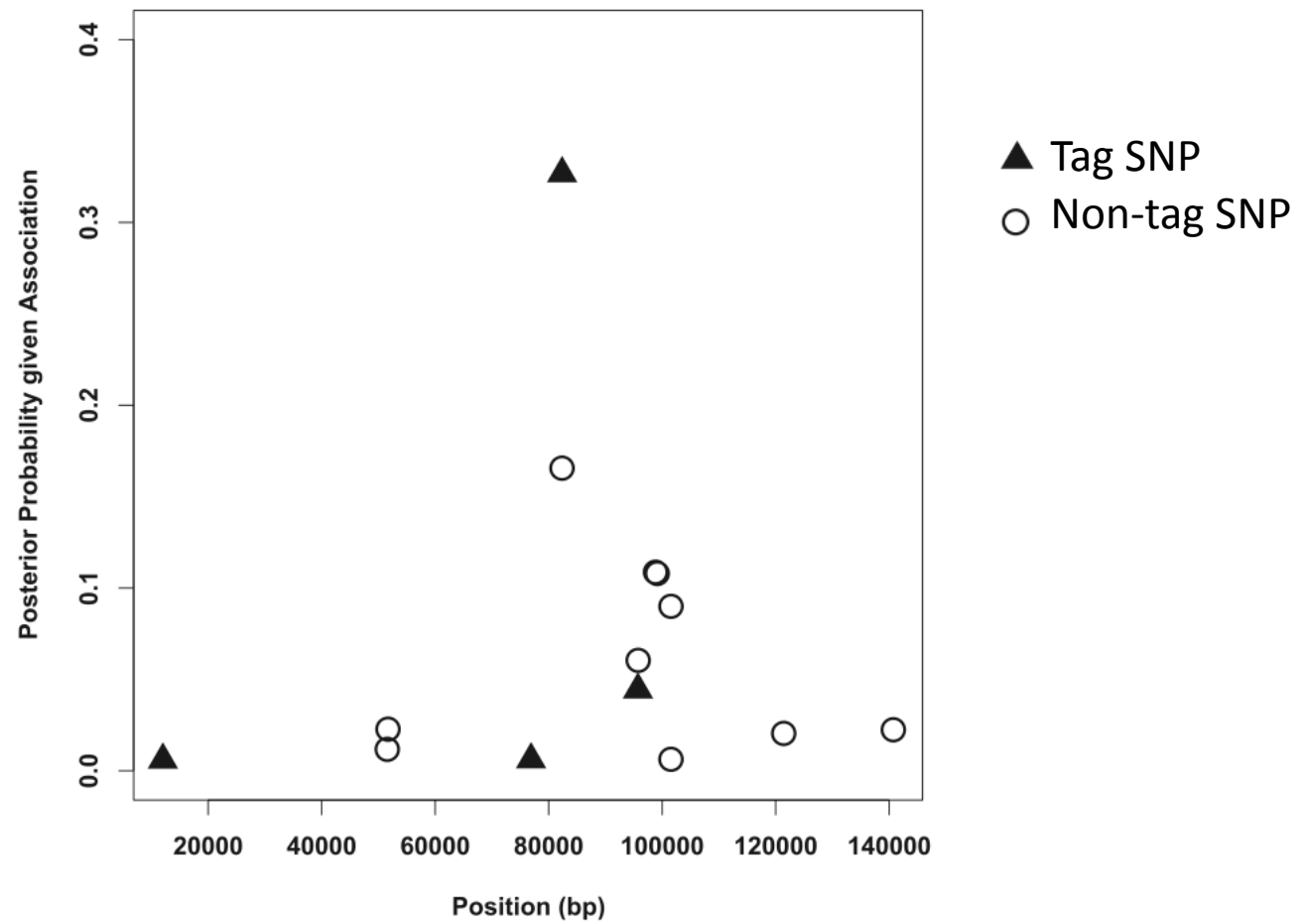
The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

PHASE can be used for imputation!

Tag SNPs and GWAS

(Servin & Stephens, 2007)



Summary

- How to identify disease-related or phenotype-related genomic loci
 - Family-based studies
 - Population-based studies for unrelated individuals
- Genome-wide association study
 - Case/control studies for discrete-valued phenotypes
 - Chi-square test based on contingency table created from genotype/phenotype data
 - Continuous-valued phenotypes (next lecture)
 - Bonferroni corrections for correcting for multiple hypothesis testing