Methods for Discovering Genetic Associations

02-410/710 Computational Genomics
April 17, 2017
Ben Lengerich

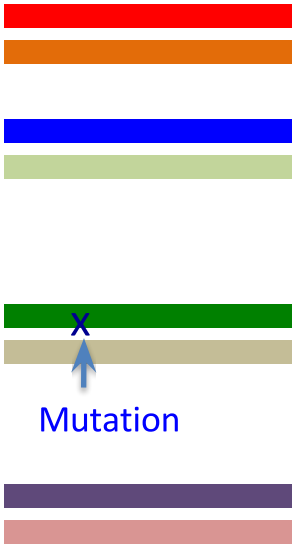
Mutations

- A natural process that changes a DNA sequence
- As a cell copies its DNA before dividing, a "typo" occurs every 10^9 basepair*year
- “Germline” mutations are inherited by the offsprings, “somatic” otherwise
- Some mutations are benign, others can be deleterious

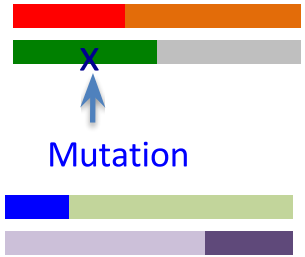
http://learn.genetics.utah.edu/content/variation/mutation/

Mutations Create Genetic Diversity

Ancestral population

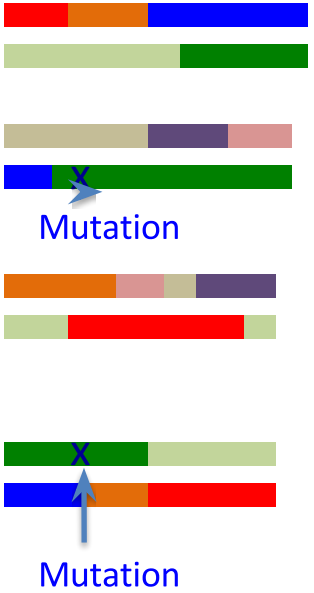


Offsprings



After many generations?

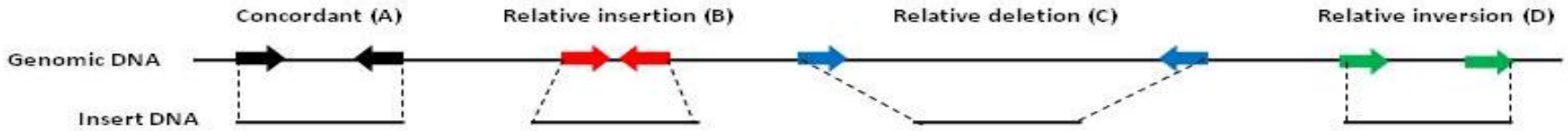
Modern chromosomes (Mosaic of ancestral chromosomes)



Genetic polymorphism

Other Types of Genetic Polymorphisms

- Structural variants
  - insertions/deletions, duplications, copy number variations



What Can We Learn from Genetic Variation?

- **Population Evolution**: the majority of human sequence variation is due to substitutions that have occurred **once** in the history of mankind at individual base pairs
  - There can be big differences between populations!
- **Markers for pinpointing a disease**: certain polymorphisms are linked to disease phenotypes
  - Association study: check for differences in SNP patterns between cases and controls
- **Forensic analysis**: the polymorphisms provide individual and familiar signatures

Single Nucleotide Polymorphisms (SNPs)

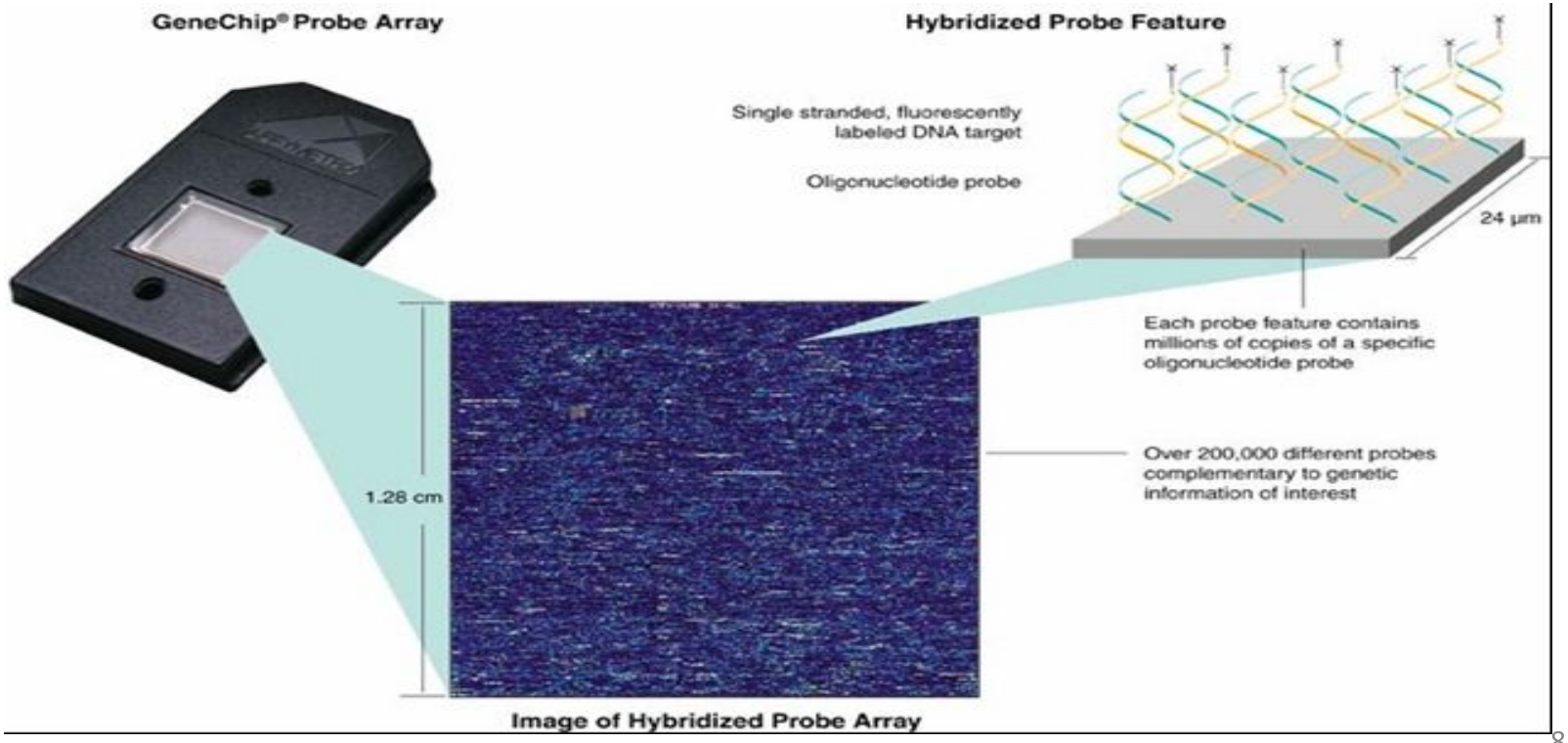
Involves a flip of a single nucleotide

```
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
```

Why SNPs?

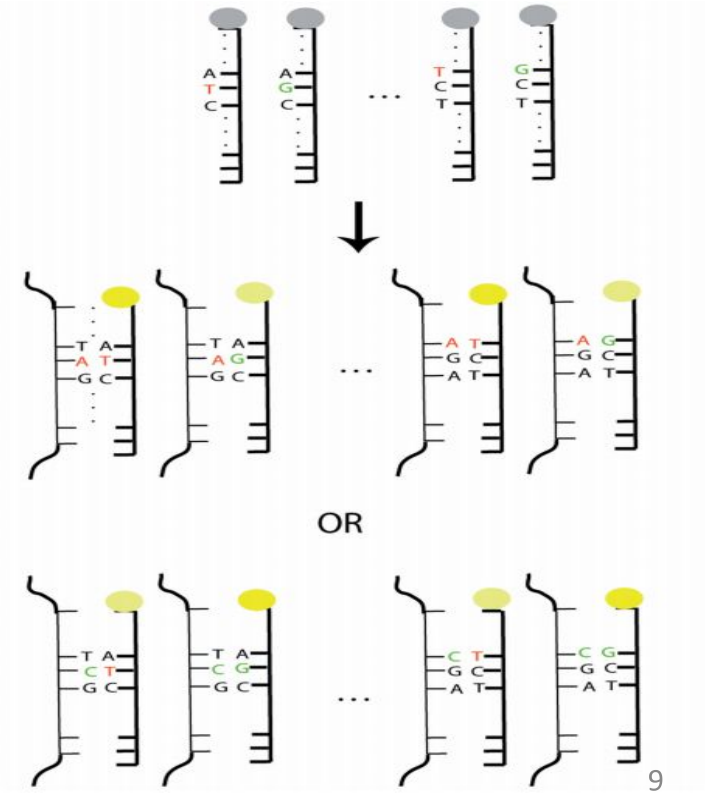
- Abundance: high frequency on the genome
  - About 40 million or more SNPs exist in human populations
- SNPs account for around 90% of human genomic variation
  - More than 5 million common SNPs each with frequency 10-50% account for the bulk of human DNA sequence difference
- Position: throughout the genome
  - Coding region, intron region, promoter site
  - It is estimated that ~60,000 SNPs occur within exons; 85% of exons are within 5 kb of the nearest SNP
- Ease of genotyping (high-throughput genotyping)

Affymetrix GeneChip Probe Array



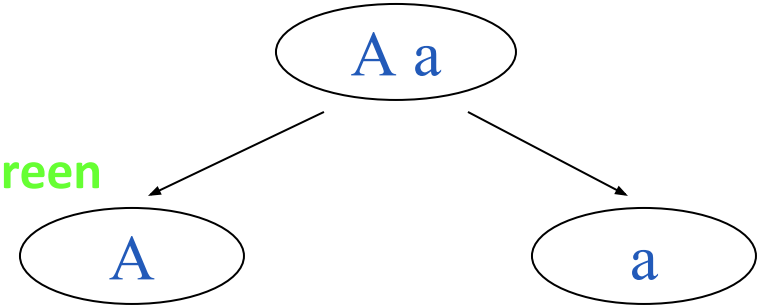
SNP Genotyping with SNP Array

- The SNP chip's basic design is similar to that of expression arrays
  - An array of 25 bp oligonucleotide sequences (features) is laid across the surface of the chip.
  - The sample's DNA is amplified, and hybridized to the array.
  - The array is scanned to quantify the relative amount of sample bound to each probe for different alleles.
- For SNPs, there is a pair of probes: one for each of the alleles.



Mendel's two laws

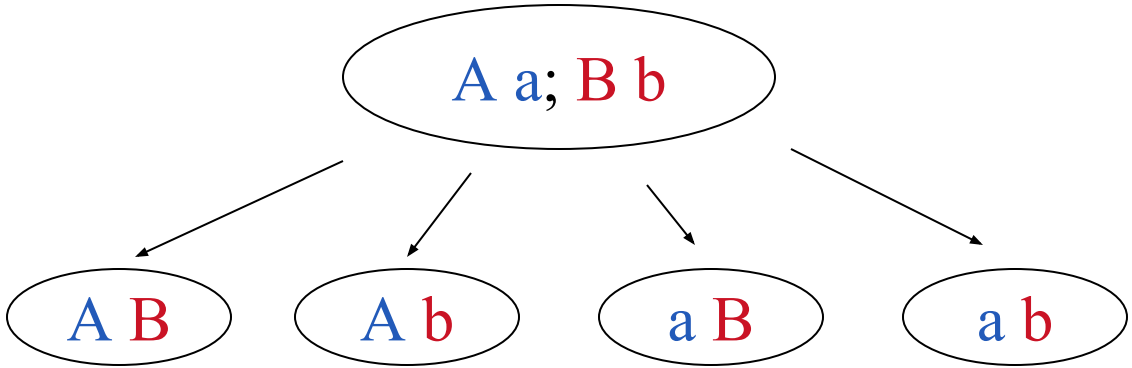
- Modern genetics began with Mendel's experiments on garden peas. He studied seven contrasting pairs of characters, including:
  - The form of ripe seeds: round, wrinkled
  - The color of the seed albumen: yellow, green
  - The length of the stem: long, short



- **Mendel's first law**: Characteristics are controlled by pairs of genes which separate during the formation of the reproductive cells (meiosis)

Mendel's two laws

- **Mendel's second law**: When two or more pairs of genes segregate simultaneously, they do so independently.



Recombination

- Inheritance of genetic material without recombination

- Inheritance of genetic material with recombination

Recombination

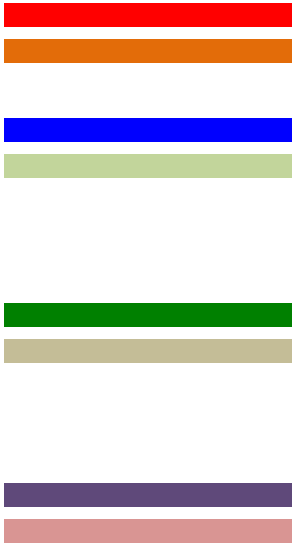
- *Parental types*: AaBb, aabb
- *Recombinants*: Aabb, aaBb
  - The proportion of recombinants between the two genes (or characters) is called the ***recombination fraction*** between these two genes.
- ***Recombination fraction*** It is usually denoted by r or θ.

If $r < 1/2$: two genes are said to be ***linked***.

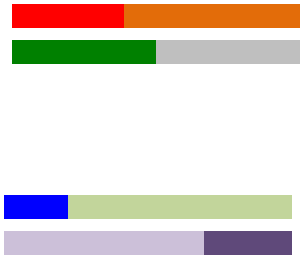
If $r = 1/2$: independent segregation (Mendel's second law).

Recombination “Breaks” Long Haplotypes Over Time!

Ancestral population

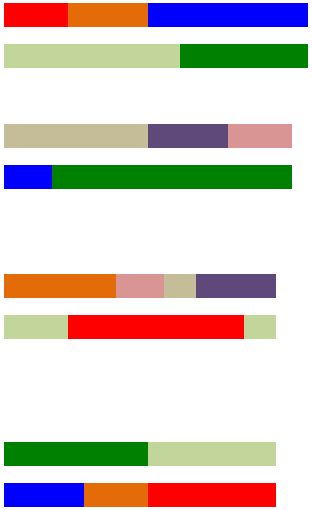


Offsprings



After many generations?

Modern chromosomes (Mosaic of ancestral chromosomes)



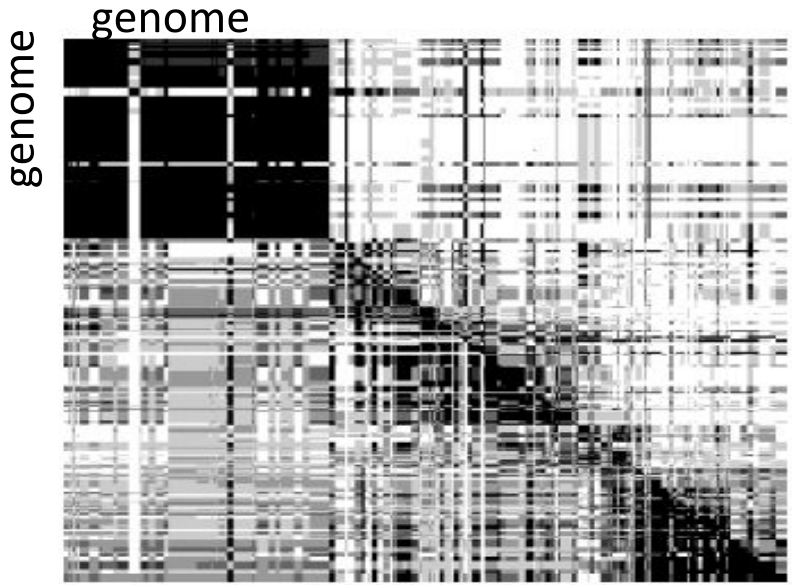
Haplotype Inference in the presence/absence of recombinations?

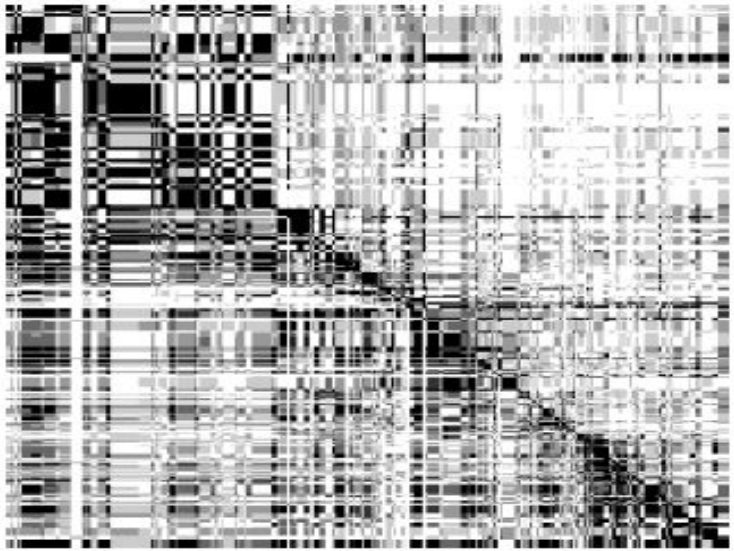
Linkage Disequilibrium (LD)

- LD reflects the relationship between alleles at different loci.
  - Linkage equilibrium: alleles at different loci are NOT linked and inherited to offsprings independently
  - Linkage disequilibrium: alleles at different loci ARE LINKED. LD is an allelic association measure

Linkage Disequilibrium in HapMap Data

- r^2 in HapMap Data

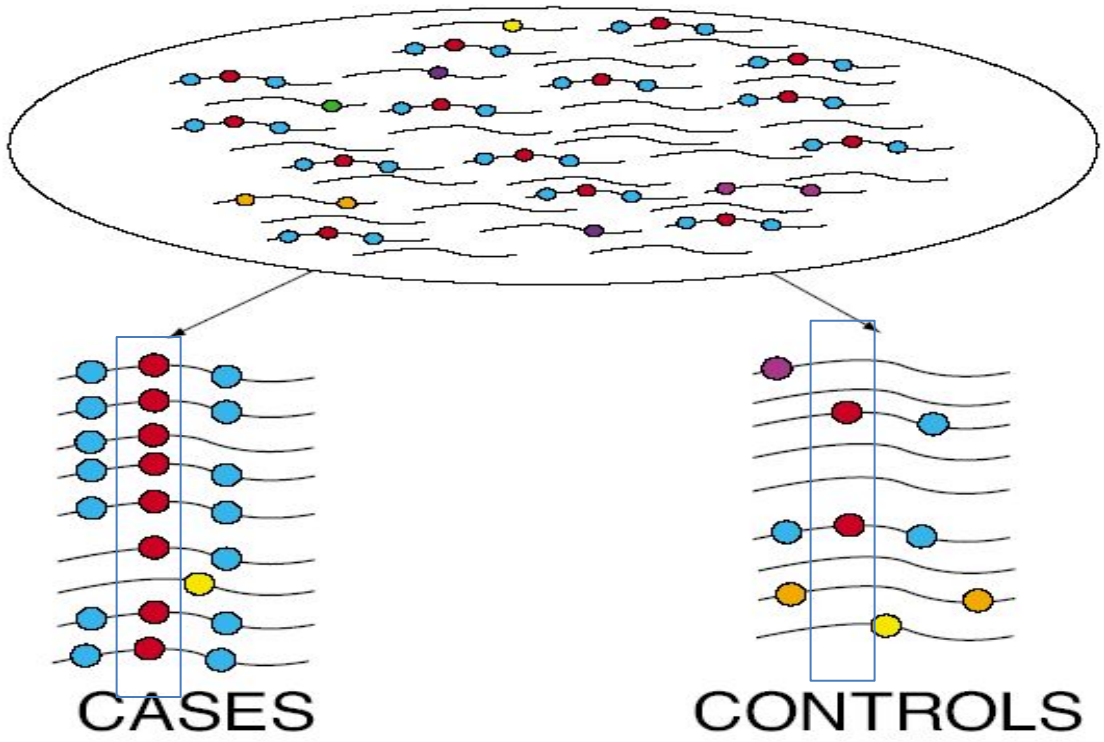




Two different populations in upper/lower diagonal

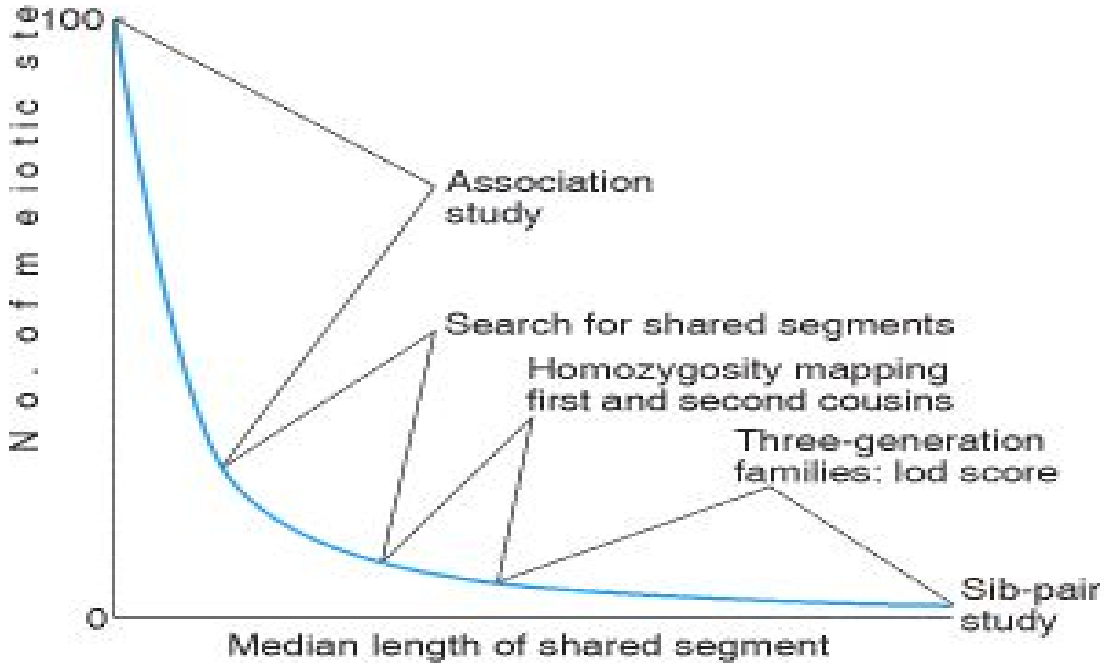
Linkage Disequilibrium in Gene Mapping

- LD is the non-random association of alleles at different loci
- Genetic recombination breaks down LD



How can we identify the genetic loci responsible for determining phenotypes?

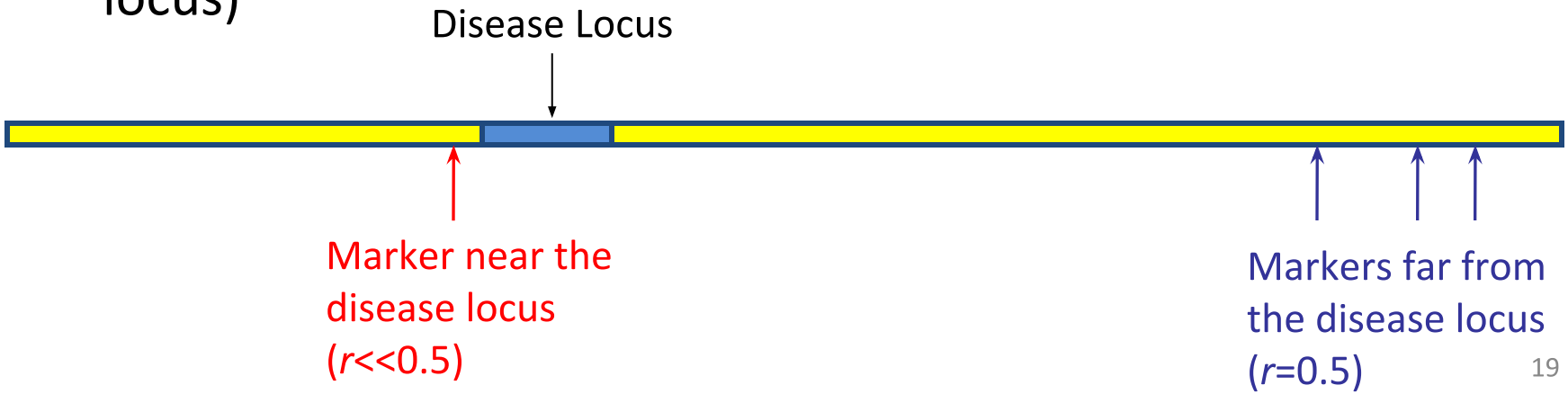
- Linkage Analysis
- Genome-wide Association Studies



Strachan & Read, Human Molecular Genetics, 2001

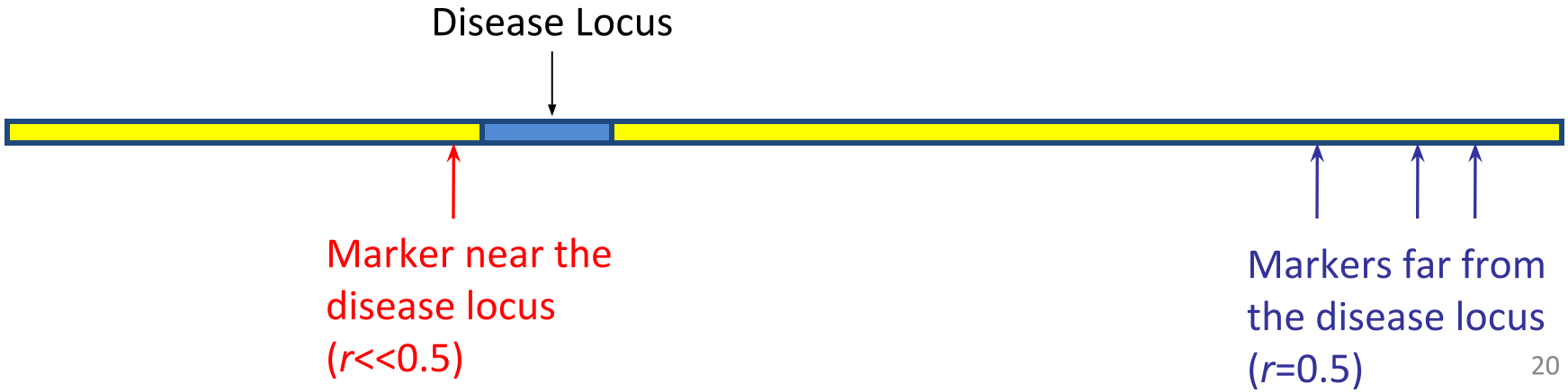
Linkage Analysis

- Goal: Identify the unknown disease locus
- Idea: Given **pedigree** data and a map of 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)



Linkage Analysis

- Data are collected for family members
  - Difficult to collect data on a large number of families
- Effective for rare diseases
- Low resolution on the genomes due to only few recombinations => Large region of linkage



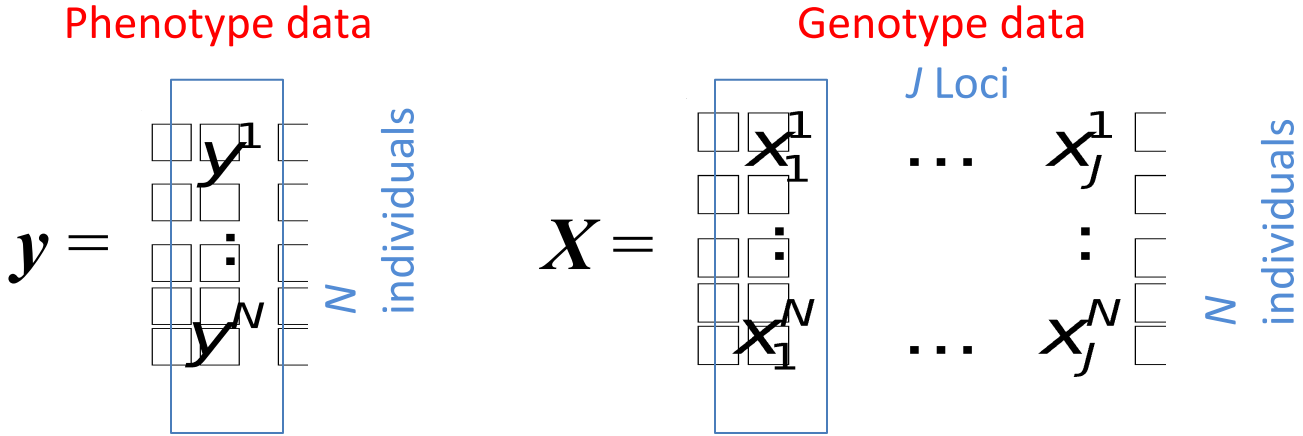
Genome-Wide Association Studies (GWAS)

- Data are collected for unrelated individuals
  - Easier to find a large number of affected individuals
- Effective for common diseases, compared to family-based method
- Relatively high resolution for pinpointing the locus linked to the phenotype

Genome-Wide Association Studies (GWAS)

- Statistical methods for testing genotype/phenotype associations
  - Discrete-valued phenotype: case/control study
  - Continuous-valued phenotype: quantitative traits
  - Sparse regression method for considering all of the SNP markers
  - Multimarker association test
- Issues arising in GWAS
  - Genotype imputation
  - From common to rare variants
  - Epistasis for multiple interacting loci
  - Correcting for population structure

Population Genotype/Phenotype Data

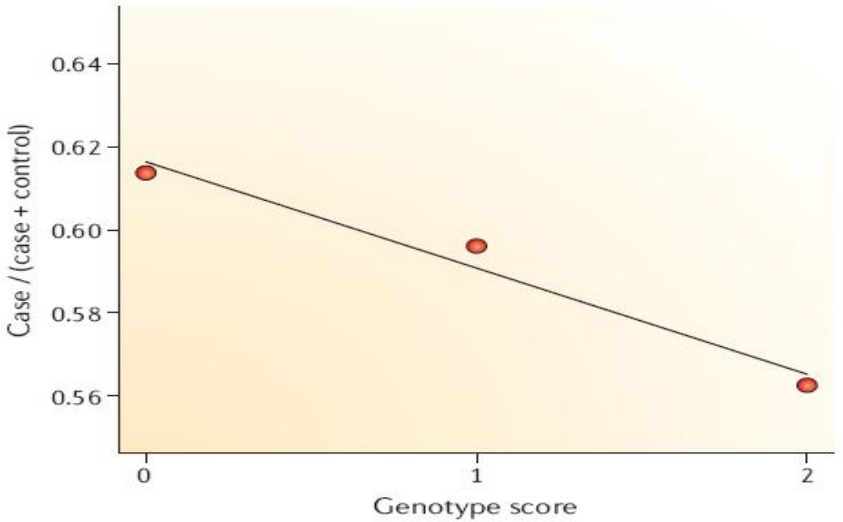


- 0 or 1 for case/control studies
  - e.g., healthy/diabetic
- Real-valued phenotypes
  - e.g., cholesterol level

Single SNP Association Test: Case/Control Study

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

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



- $\chi 2$ test with 2 df under the null hypothesis of no association

Genotype score = the number of minor alleles

Single SNP Association Analysis: Case/Control Study

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

Genotype	Case	Control
A	$G_{case,A}$	$G_{control,A}$
a	$G_{case,a}$	$G_{control,a}$
Total	$2xN_{case}$	$2xN_{control}$

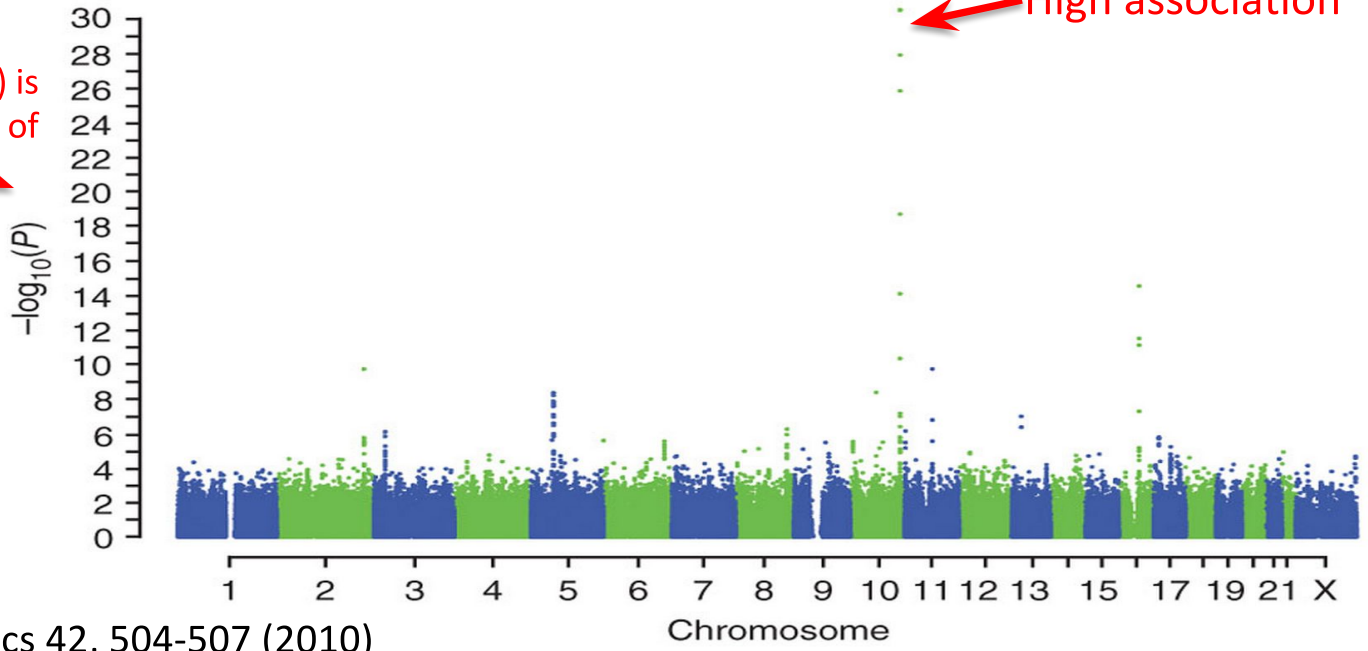
- $\chi 2$ test with 1df

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

Often $-\log_{10}$ (p-value) is plotted instead of p-value.

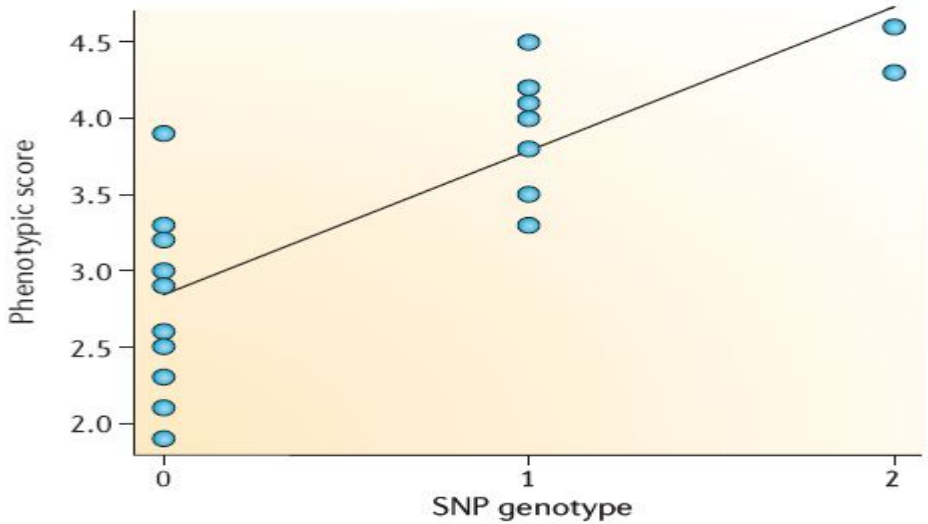
High association



Nature Genetics 42, 504-507 (2010)

Single SNP Association Test: Continuous-valued Traits

- Continuous-valued traits
  - Also called quantitative traits
  - Ex: Cholesterol level, blood pressure
- Fit a linear regression at each locus
- *t*-test with null hypothesis "No associations, i.e., $\beta = 0$"



Correcting for Multiple Testing

- What happens when we scan the genome of 1 million genetic markers for association with $\alpha = 0.05$?
  - 50,000 (=1 millionx0.05) 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
  - Permutation test

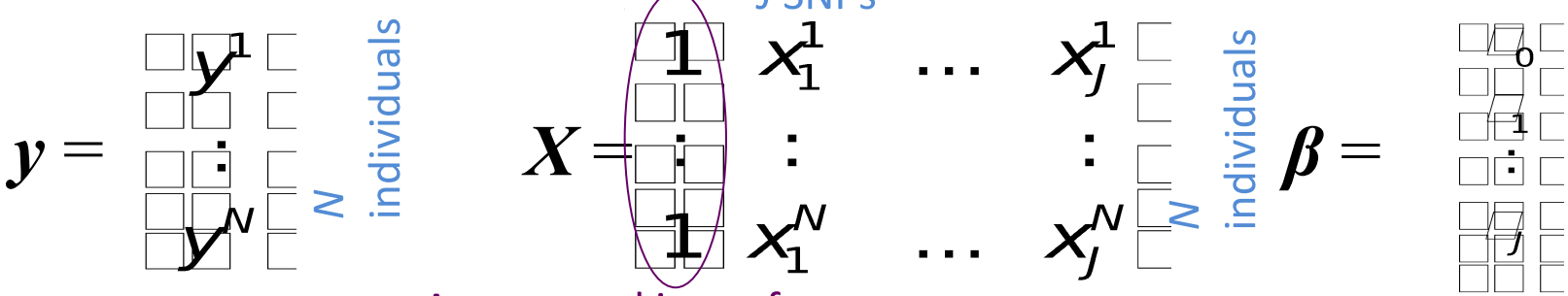
Vector/Matrix Representation

- Sparse regression method to evaluate the effect of each SNP in the context of all other SNPs

$$y = X\beta + \varepsilon$$

Phenotype data

Genotype data



- Sparsity constraint: Few SNPs are influencing the given phenotype and the rest of the SNPs have effect size 0

L1 Regularization (Lasso)

Solves the optimization problem:

$$\arg\min_{\beta}(y - X\beta)^2 + \lambda|\beta|$$

Traits (n x 1)

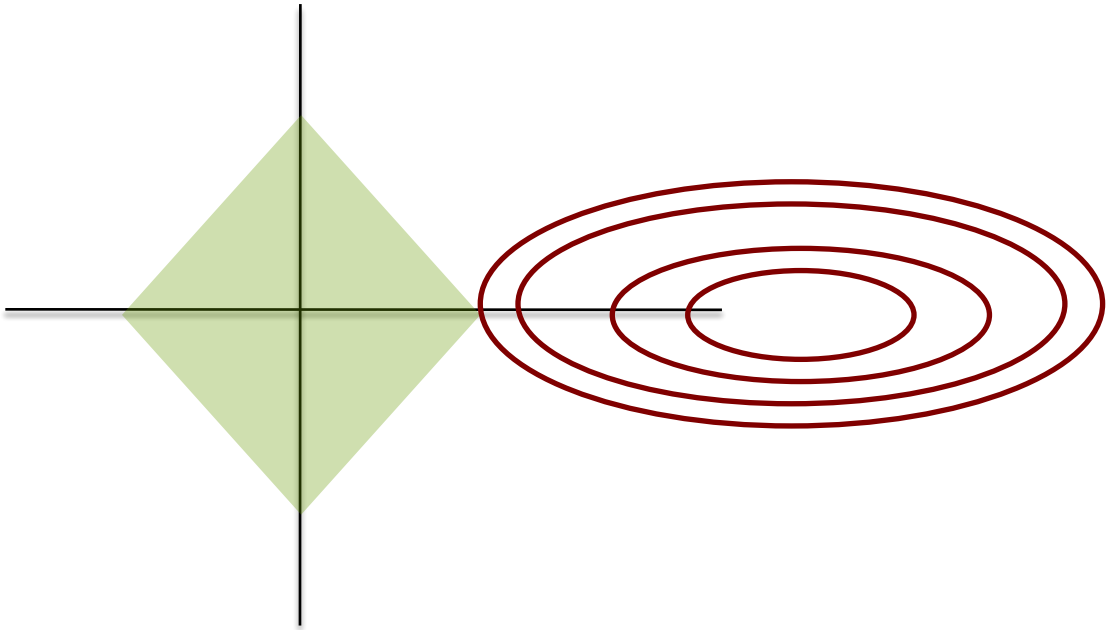
SNPs (n x p)

Effect sizes (p x 1)

Sparse effect sizes

L1 Regularization (Lasso)

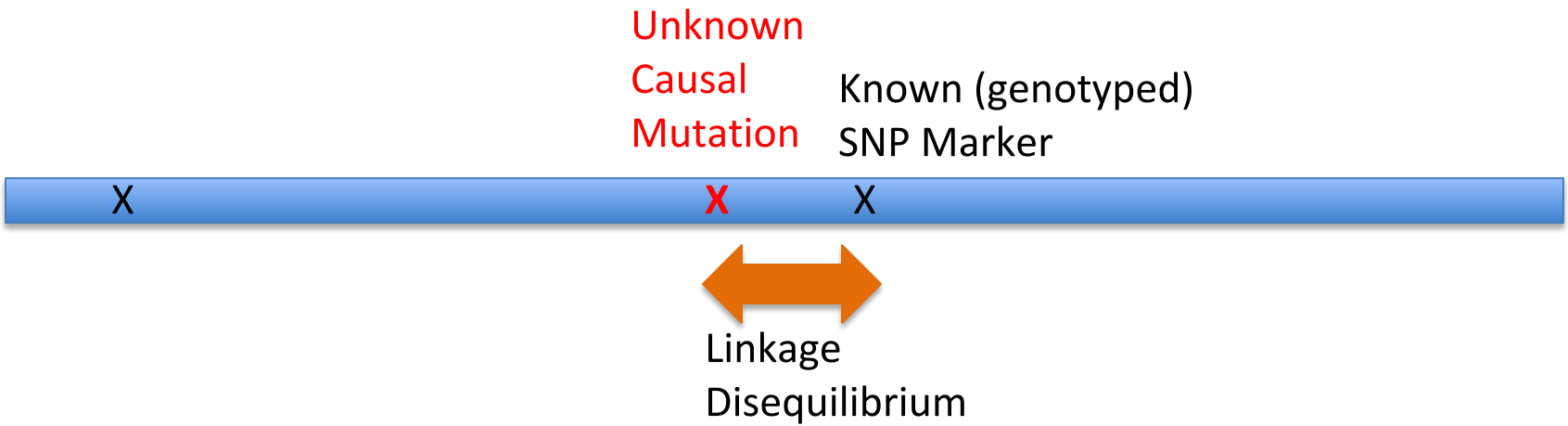
Selects variables on the corners of the polytope:



Overview

- Statistical methods for testing genotype/phenotype associations
  - Discrete-valued phenotype: case/control study
  - Continuous-valued phenotype: quantitative traits
  - Sparse regression method for considering all of the SNP markers
  - Multimarker association test
- Issues arising in GWAS
  - Genotype imputation
  - From common to rare variants
  - Epistasis for multiple interacting loci
  - Population structure

Causal Mutations and Genetic Markers



What happens when SNP density increases?

Common Variants vs. Rare Variants

- First-generation genome-wide association study (GWAS): **common variant common disease** hypothesis
- Common variants with minor allele frequency (MAF)>5%
  - dbGap: ~11 million SNPs
  - HapMap: 3.5 million SNPs
  - A successful GWAS requires a more complete catalogue of genetic variations
- Rare variants (MAF<0.5%), low-frequency variants (MAF:0.5%~5%)
  - Captured by sequencing with next-generation sequencing technology
  - Possibly significant contributors to the genetic architecture of disease
    - Causal variants are subject to negative selection

Associations to Rare Variants

- Often GWAS are underpowered for functional rare variants

Common Variant Association

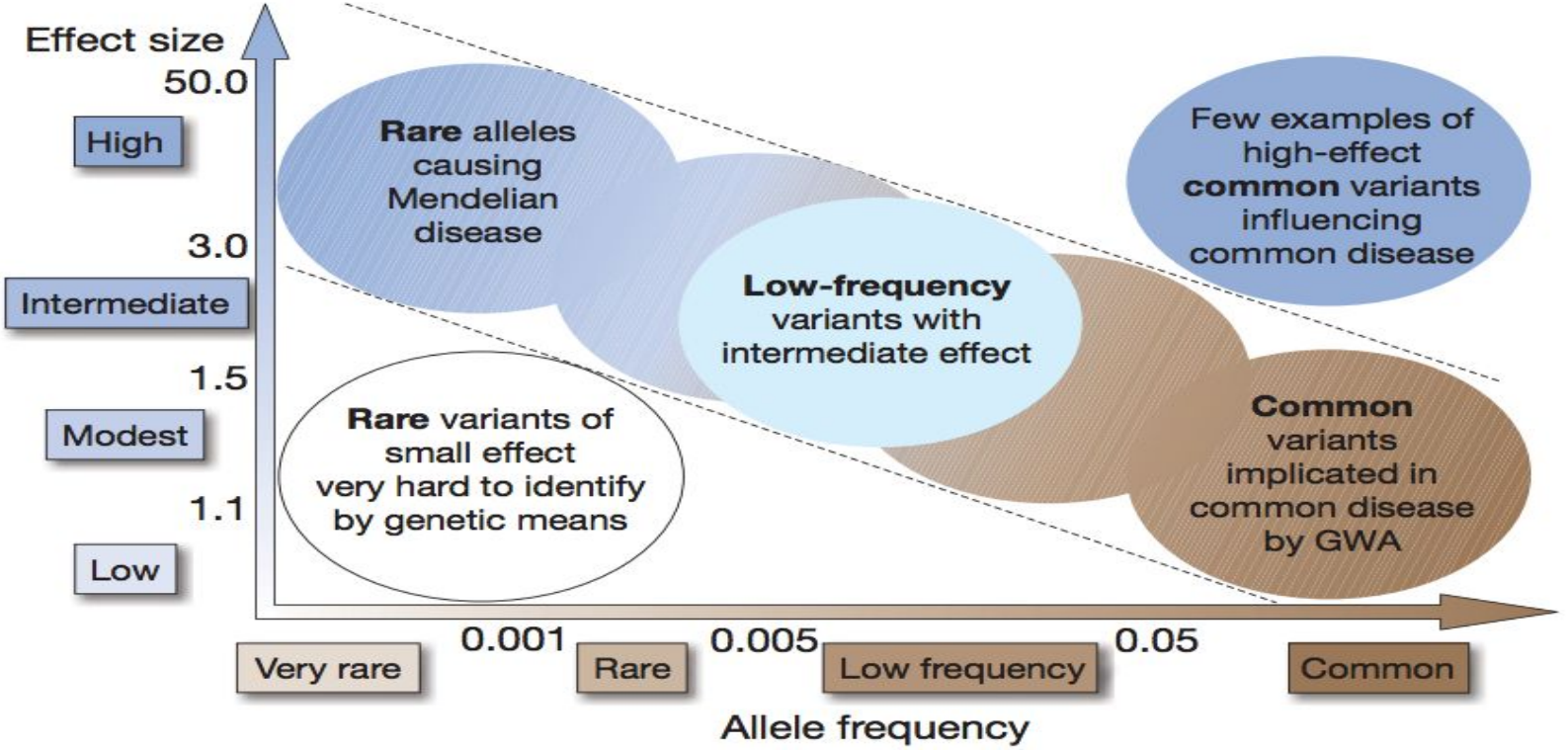
	Case	Control
Allele a	60	20
Allele A	40	80

Rare Variant Association

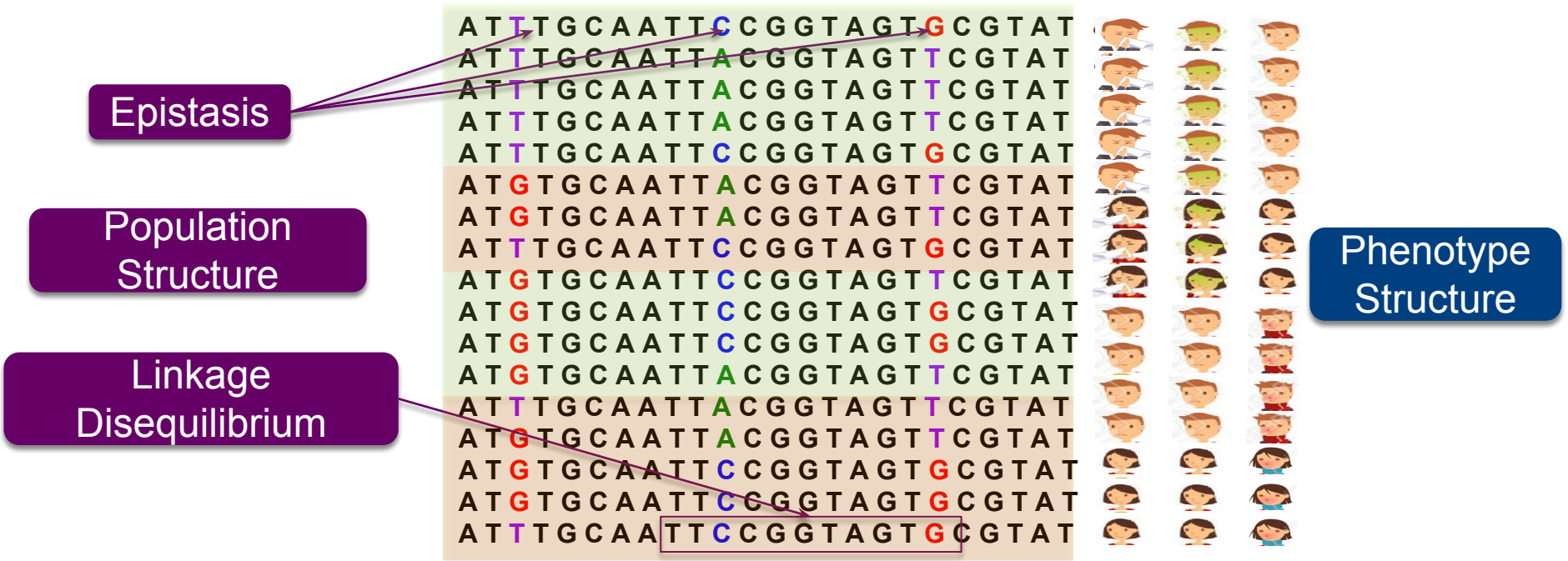
	Case	Control
Allele a	7	2
Allele A	93	98

- Common variant GWA approaches are appropriate only for common variants

Feasibility of Identifying Disease Loci



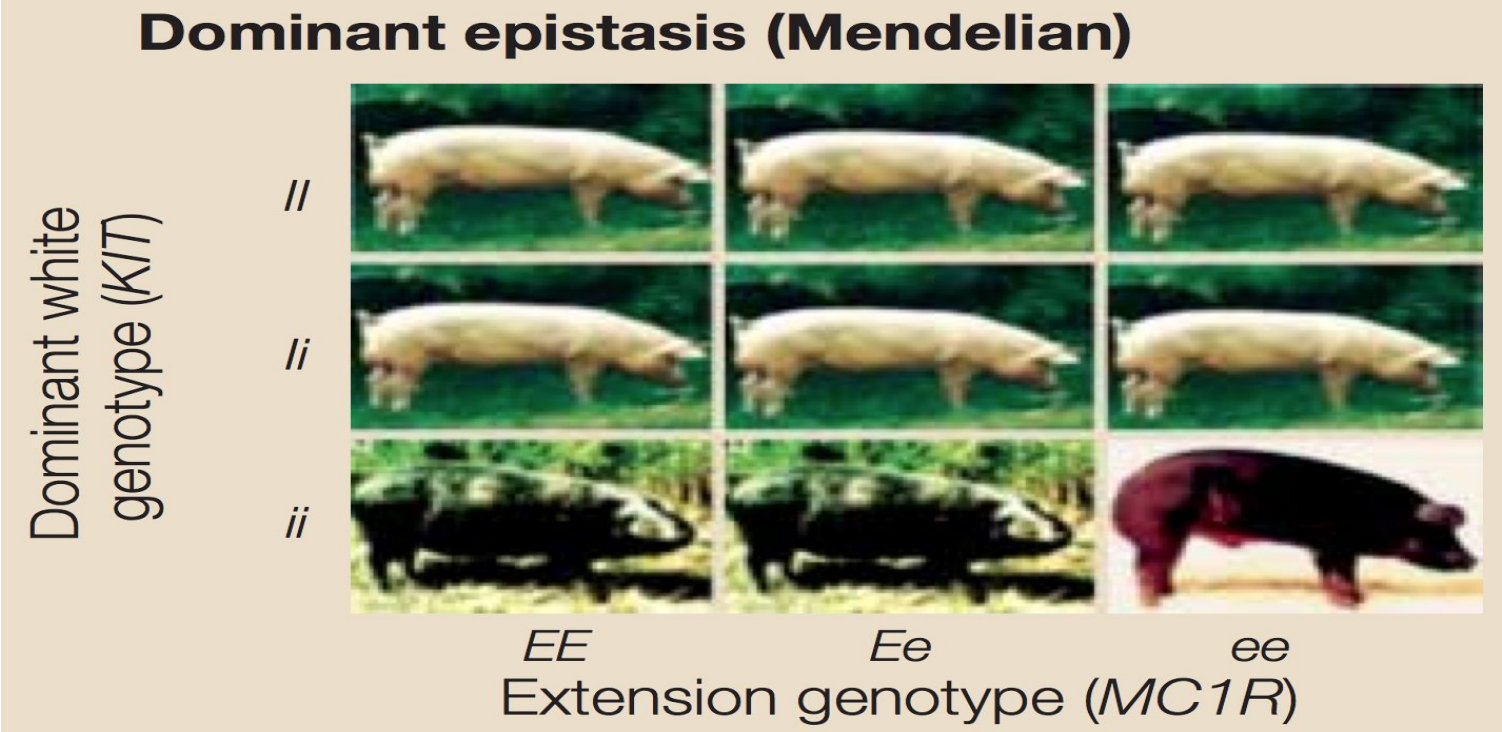
Many types of structure in genomic data



Epistasis

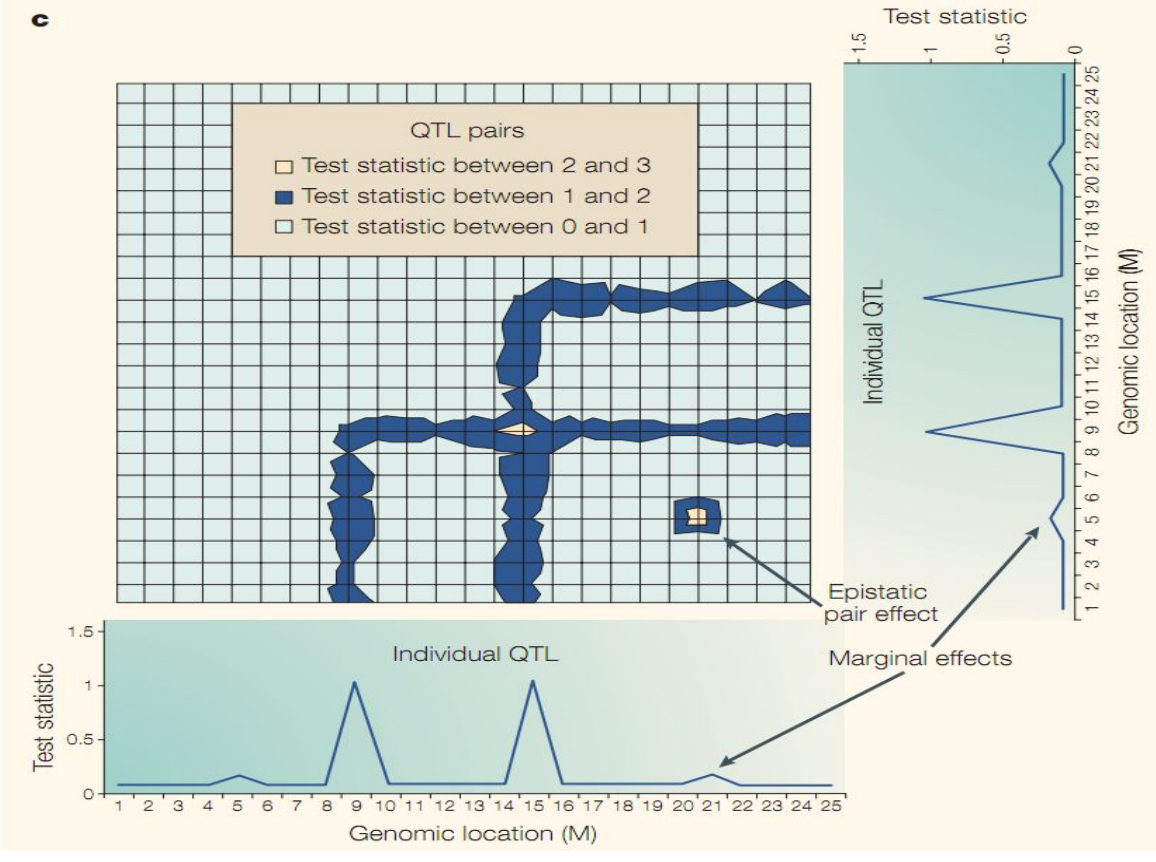
- Definition: The effect of one locus depends on the genotype of another locus
  - Epistatic effects vs. marginal effects

Epistasis for Mendelian Traits



Carlborg & Haley, Nature Reviews Genetics 2004

Epistatic and Individual QTLs

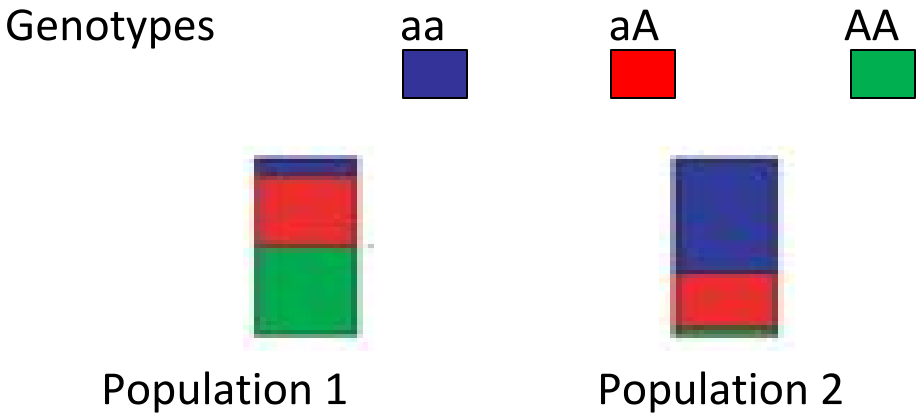


Detecting Epistasis

- Epistatic effects of SNPs can often be detected only if the interacting SNPs are considered jointly
  - The number of candidate SNP interactions is very large
    - For J SNPs, JxJ SNP pairs need to be considered for epistasis
    - In general for J SNPs and K-way interactions, there are $O(J^K)$ candidate interactions
    - Computationally expensive to consider all possible groups of interacting SNPs
    - For a reliable detection of K-way interactions, a large sample size is required
  - Multiple testing problem

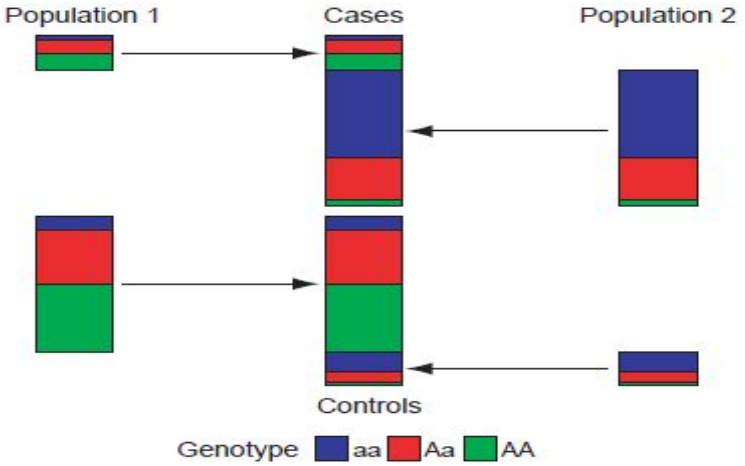
Population Structure

- A set of individuals characterized by some measure of genetic distinction
- A “population” is usually characterized by a distinct distribution over genotypes
- Example:



Population Structure and Association Analysis

- Population structure in data causes false positives
  - Samples in the case population are usually more related
  - Any SNPs more prevalent in the case population will be found significantly associated with the trait.



Accounting for Population Structure in Association Analysis

- Need to account for population structure in association mapping.
- Careful study design with each population represented in case/control groups in a balanced way.
  - Can be hard to control for population structure during data collection
  - Cryptic population structure
- Statistical Methods
  - Trace Lasso
  - Precision Lasso