BST227
Introduction to Statistical Genetics

Lecture 4:

Introduction to linkage and association analysis

Housekeeping

- Homework #1 due today
- Homework #2 posted (due Monday)
- Lab at 5:30PM today (FXB G13)

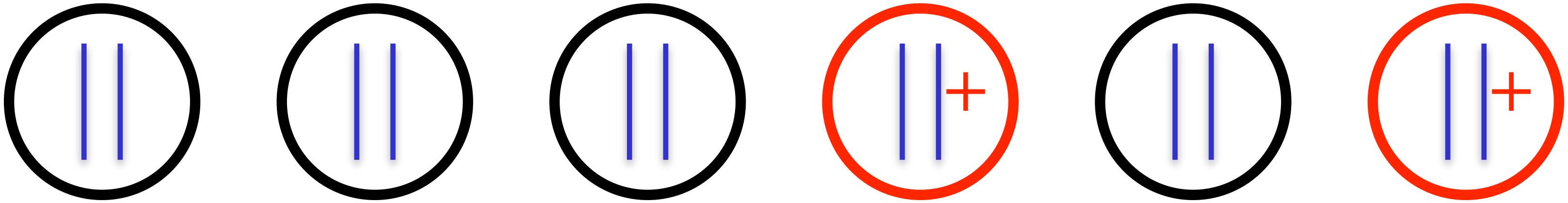
Reading

The Fundamentals of Modern Statistical Genetics
by Nan Laird and Christoph Lange

Lecture title	**Reading**
1. Background	Chapter 1
2. Mendel’s Laws, genetic models for disease	Chapter 2
3. Hardy Weinberg Equilibrium and Recurrence risk Ratios	Chapters 3, 4.1-2
4. An overview of linkage and association	Chapter 5

- Last time: Relationships between allele and genotype frequencies: Hardy Weinberg Equilibrium

- Today: Relationships between different loci



Y = 0 0 0 1 0 1

+ denotes the presence of a disease-causative variant at the DSL

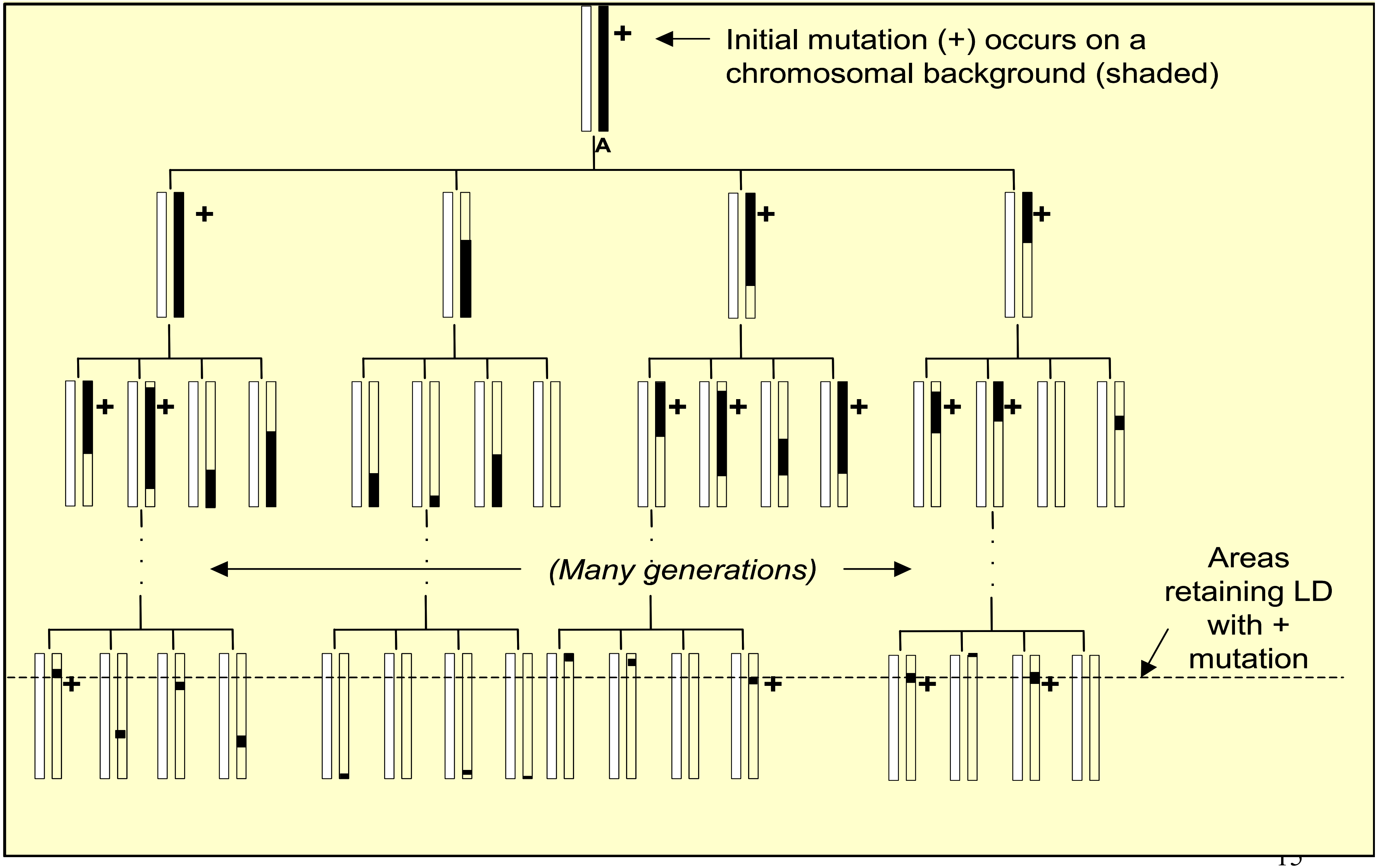
Goal: Locate the genetic variant(s) (i.e. the DSL) presumed to be biologically causal for a disease.

Two main statistical methods for finding disease susceptibility loci

DSL

- Linkage Analysis: Based on recombination
- Association Analysis: Based on linkage disequilibrium (LD)

We take advantage of failure of Mendel’s second law (Independent assortment)



Initial mutation (+) occurs on a chromosomal background (shaded)

(Many generations)

Areas retaining LD with + mutation

Mapping Strategies

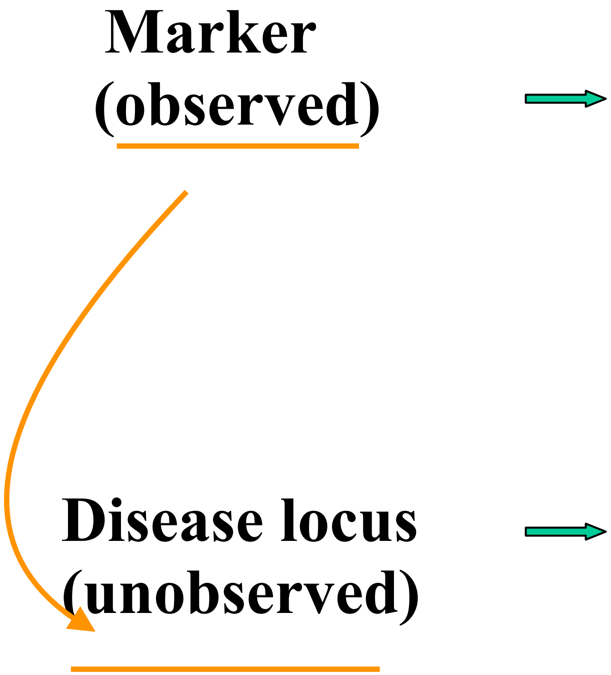
Linkage analysis

- Family data
- Few (<1000 genome-wide) markers required
- Can find potential disease loci located “far” from marker
- Low resolution (finds big candidate regions)

Association analysis

- Case/Control data
- Relatively dense markers required (~1M genome-wide)
- Marker needs to be very close to DSL
- Higher resolution

Linkage and Association between a Genetic Marker and the DSL



Marker (observed)

Disease locus (unobserved)

LINKAGE: Based on physical concept of distance. Two linked loci are on the same chromosome and close enough that they are not inherited independently.

ASSOCIATION: Statistical concept. A particular allele at a marker is associated with the disease variant (DSL) at the causal locus in the population. Population concept, AKA Linkage Disequilibrium

Linkage vs Association: Statistical tests

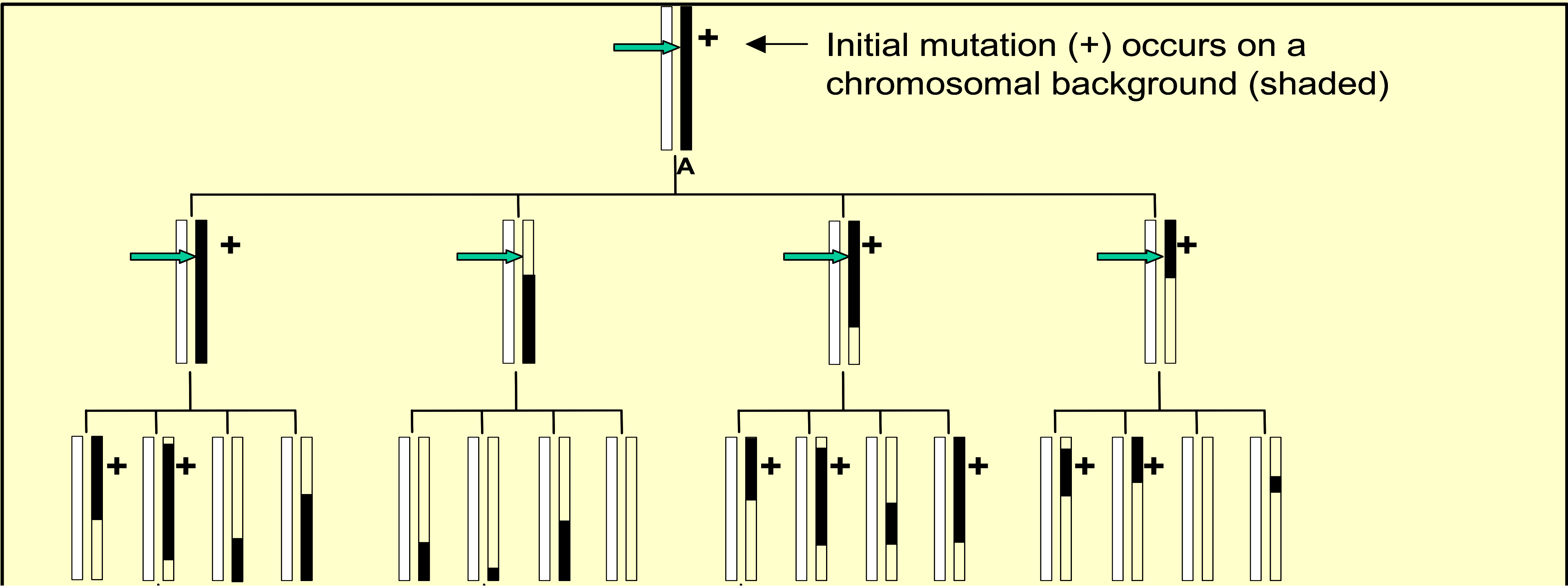
Linkage

Two loci are linked if the recombination fraction is less than ½, i.e. the loci are NOT inherited independently. Linkage analysis is based on testing if the recombination fraction between a marker and the DSL is = ½.

Association

Two loci are associated if the alleles at one locus are not independent of the alleles at another locus (allelic association). Association analysis is based on testing statistical independence between disease and a marker

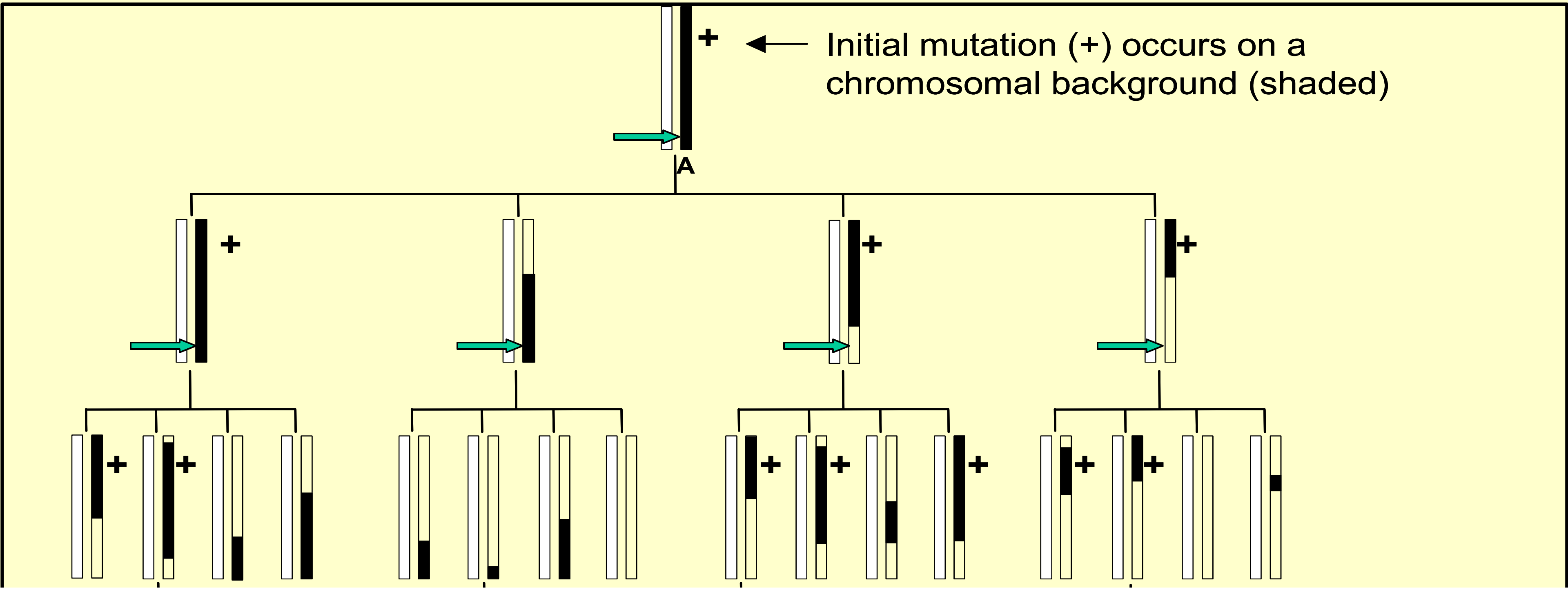
Linkage



+
Initial mutation (+) occurs on a chromosomal background (shaded)
A
+
+
+
+
+
+
+
+
+
+
+

Marker (green arrow) is "close" to DSL (+)

No Linkage


+
Initial mutation (+) occurs on a chromosomal background (shaded)
A
+
+
+
+
+
+
+
+
+
+
+

Marker (green arrow) is “far” from DSL (+)

Basic Idea of Linkage analysis

Based on recombination:

$$\theta = P(\text{recombination occurs between two loci})$$

If two loci are on top of each other, $\theta = 0$

If two loci are far apart (or on different chromosomes), $\theta = ½$

To test for linkage:

Count number of recombinations we observe; estimate θ as proportion of recombinations

Null hypothesis H_0: $\theta = ½$

Alternative hypothesis H_1: $\theta < ½$

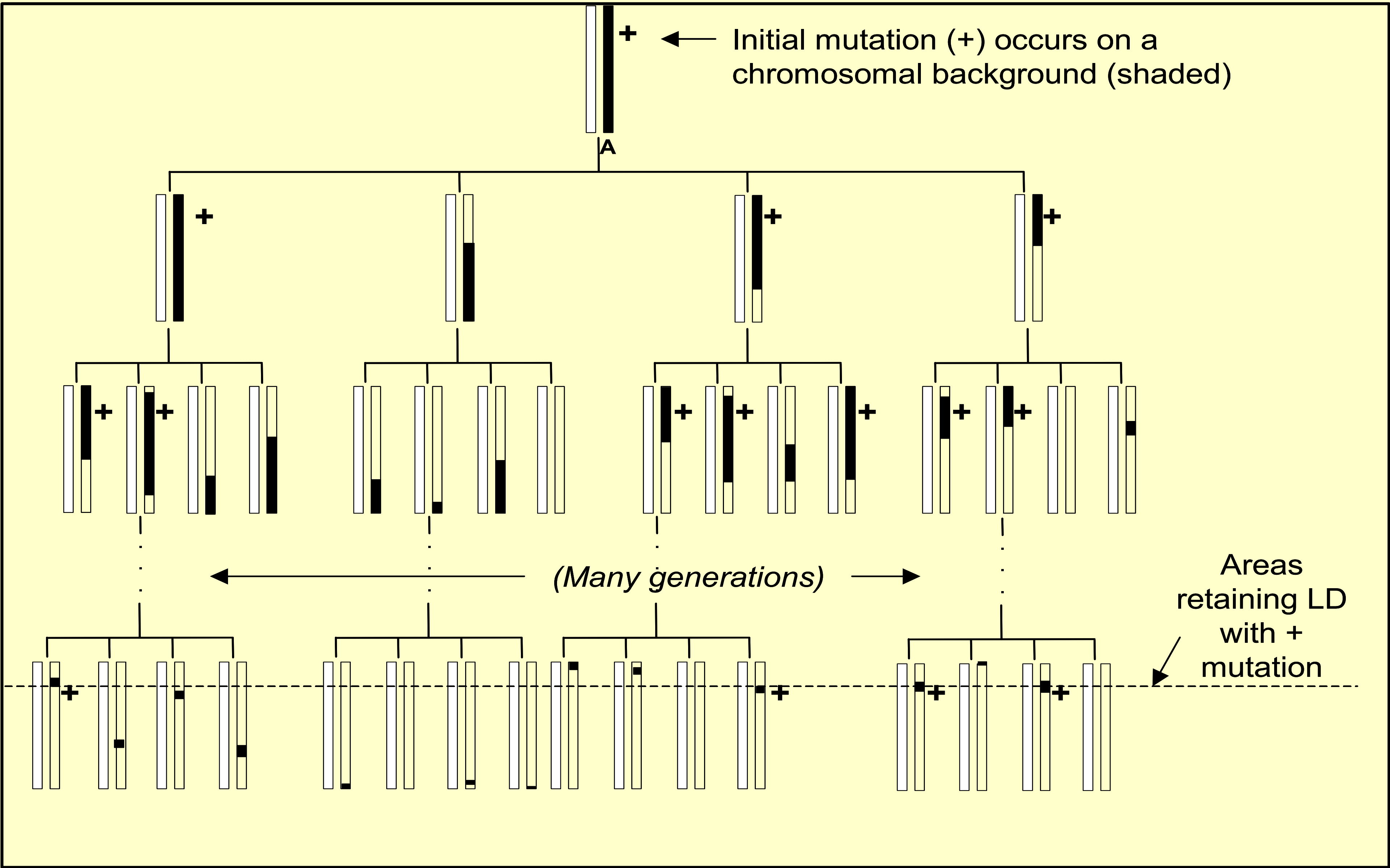
Rejection of null implies there is linkage.

Features of Linkage Analysis

- Must have family data with multiple affected individuals
- Requires being able to infer relationship between variant at DSL and disease trait. *Easiest with Mendelian 0/1 penetrance functions and mode of inheritance known.*
- Uses relatively few markers (<1000) for whole genome linkage analysis
- Rejecting null hypothesis of no linkage may implicate thousands of genes
- Very successful for Mendelian disorders, less so for complex

Association / Linkage Disequilibrium

Figure 1. Example of Linkage Disequilibrium through generations


+
Initial mutation (+) occurs on a chromosomal background (shaded)
A
+
+
+
+
+
+
+
+
+
+
+
(Many generations)
Areas retaining LD with + mutation
+
+
+
+

Haplotype

- Haplotype: Set of alleles at multiple loci on a particular chromosome transmitted from parent to child
- Red for haplotype from Mom, blue from Dad

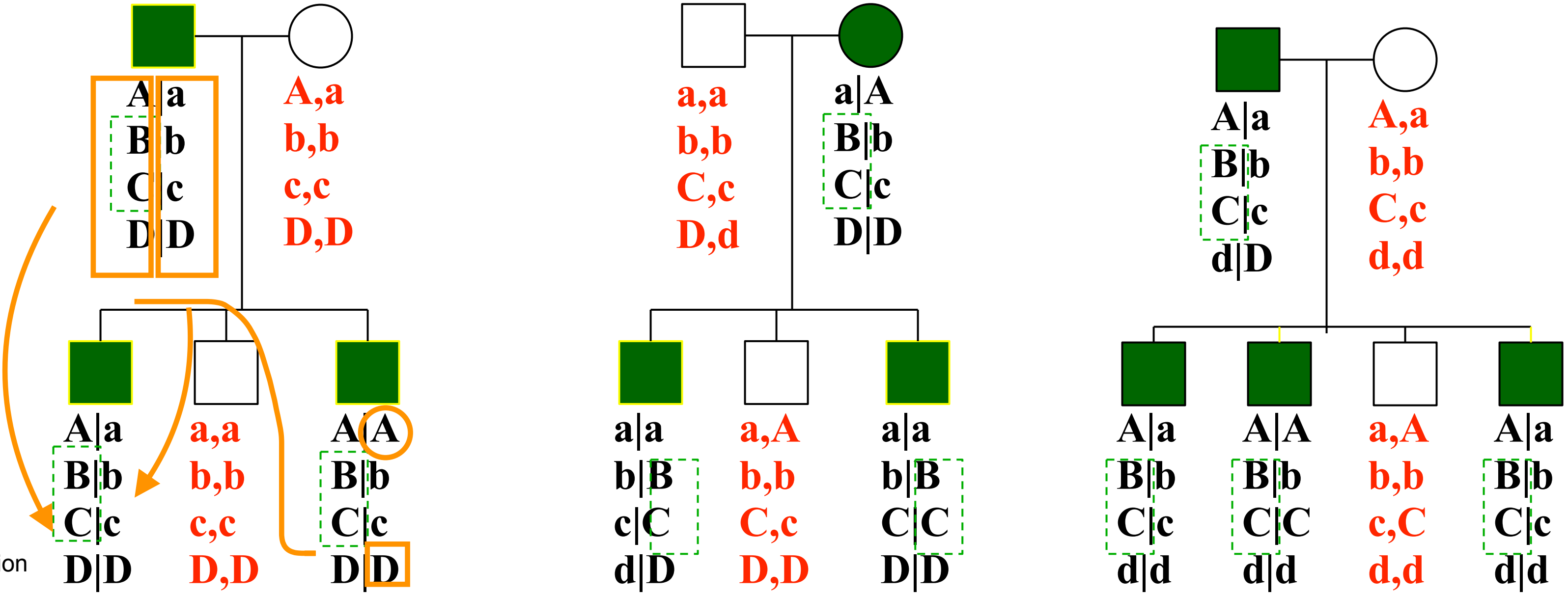
Genotype	Haplotypes	
A/A	**A**	**A**
C/T	**T**	**C**
C/C	**C**	**C**
T/T	**T**	**T**
G/G	**G**	**G**
A/A	**A**	**A**

Phase –

- Knowledge of the origin of alleles (either from mom or dad).

Linkage
vs.
Linkage Disequilibrium

B is the DSL



NO recombination

Which loci are linked to B?

all 3 loci are linked to B

no reconbinantion
A はA
BはBとして保存さ
れているから

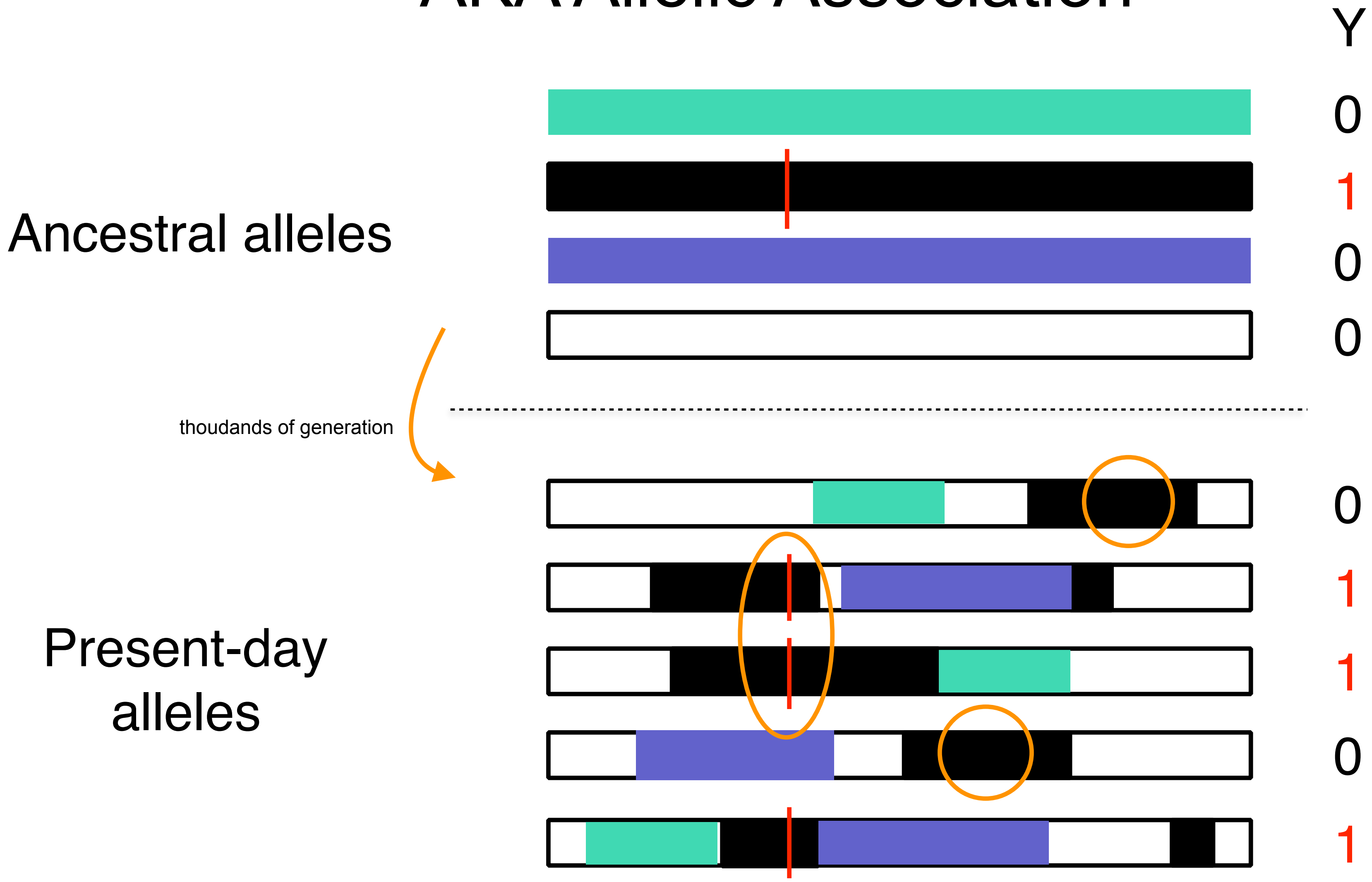
Which loci are in LD with B?

only B and C are in LD associated at
teh populational levle

Linkage Disequilibrium (LD) AKA Allelic Association

- LD usually exists when two loci are 'close' (about 50kb – maybe up to 300kb)
- When a mutation arises in a population, there is high LD between the mutation and other variants on the same chromosome.
- Over generations, LD dissipates between mutation and loci far away via recombination.

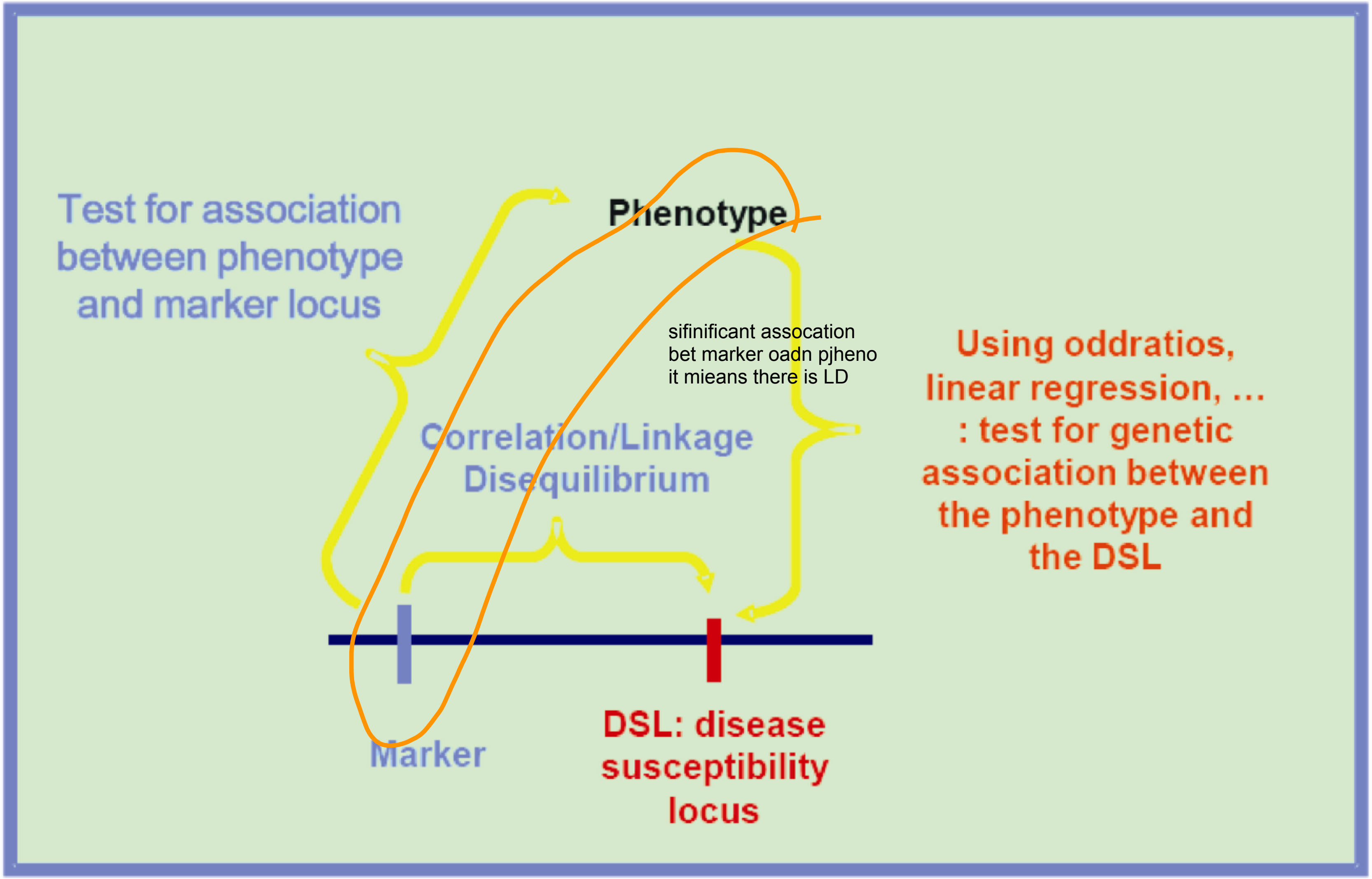
Linkage Disequilibrium (LD) AKA Allelic Association



red line is one snp or one tandem reoeat
50-300kb gene around red line is LD

Based on Abecasis

Using LD for Mapping: Association



Test for association between phenotype and marker locus

Phenotype

sifinificant assocation
bet marker oadn pjheno
it mieans there is LD

Using oddratios, linear regression, … : test for genetic association between the phenotype and the DSL

Correlation/Linkage Disequilibrium

Marker

DSL: disease susceptibility locus

Formal Definiton of LD

LD refers to association between alleles at two markers on same haplotype (chromosome)

	B	**b**	
A	p_{AB}	p_{Ab}	p_A
a	p_{aB}	p_{ab}	p_a
	p_B	p_b	**1**

p_{AB} **= proportion of chromosomes with an A and a B at the two loci**

Linkage Equilibrium
(absence of linkage)

Linkage Equilibrium occurs when alleles at two loci are independent of each other:

	B	**b**	
A	$p_{AB} = p_A p_B$		p_A
a			p_a
	p_B	p_b	**1**

Independence ⟺ Linkage Equilibrium

Linkage Disequilibrium (LD)

Association between alleles A and B
Coefficient of linkage disequilibrium (“Disequilibrium coefficient”):

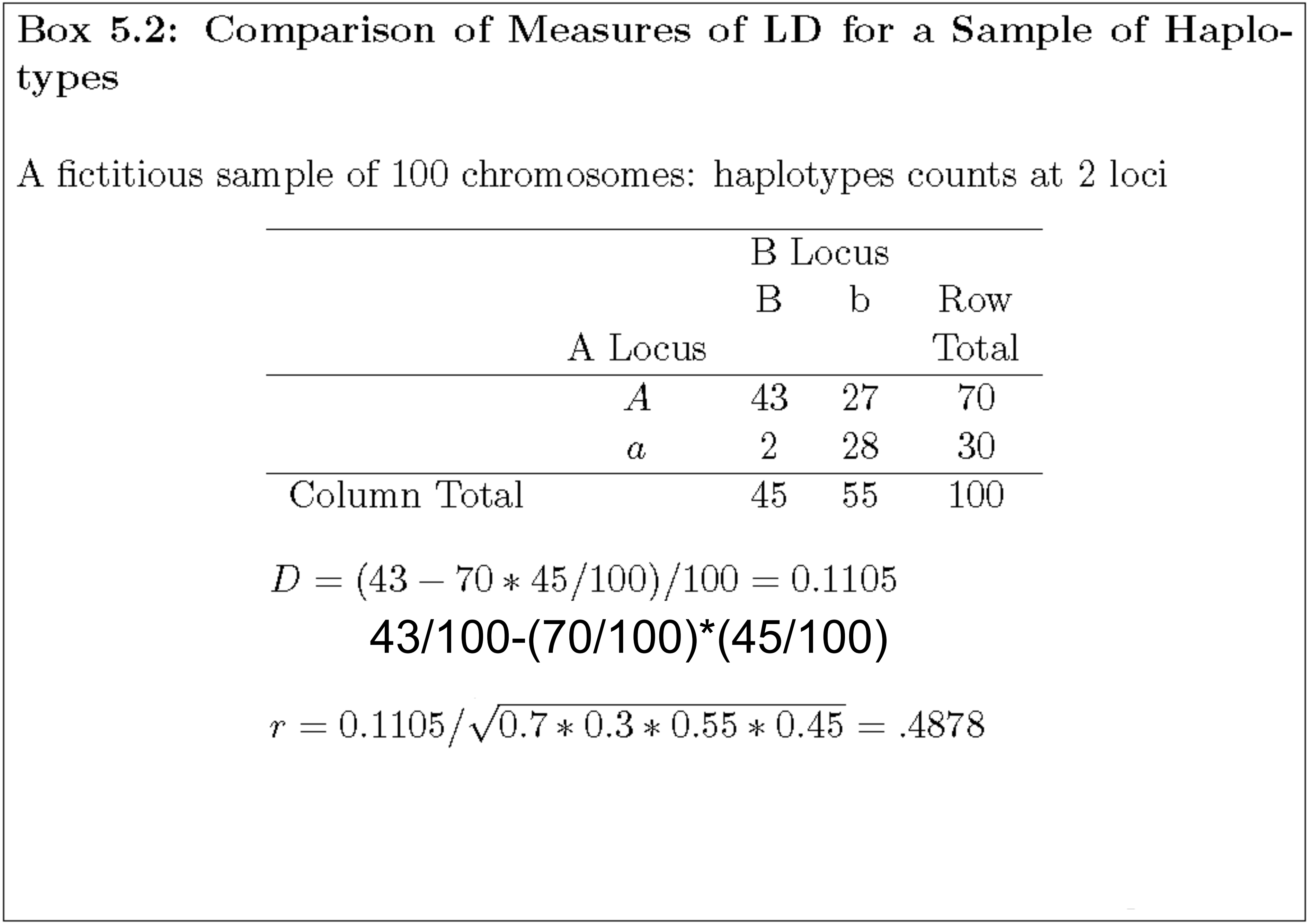
$$D = p_{\mathbf{AB}} - p_{\mathbf{A}} p_{\mathbf{B}}$$

Problem: Range of D depends on margins

Alternative:

correlation, $r = D/\sqrt{\ } p_{\mathbf{A}} p_{\mathbf{B}} p_{\mathbf{a}} p_{\mathbf{b}}$

nr^2 is effective sample size for testing marker instead DSL



Box 5.2: Comparison of Measures of LD for a Sample of Haplotypes

A fictitious sample of 100 chromosomes: haplotypes counts at 2 loci

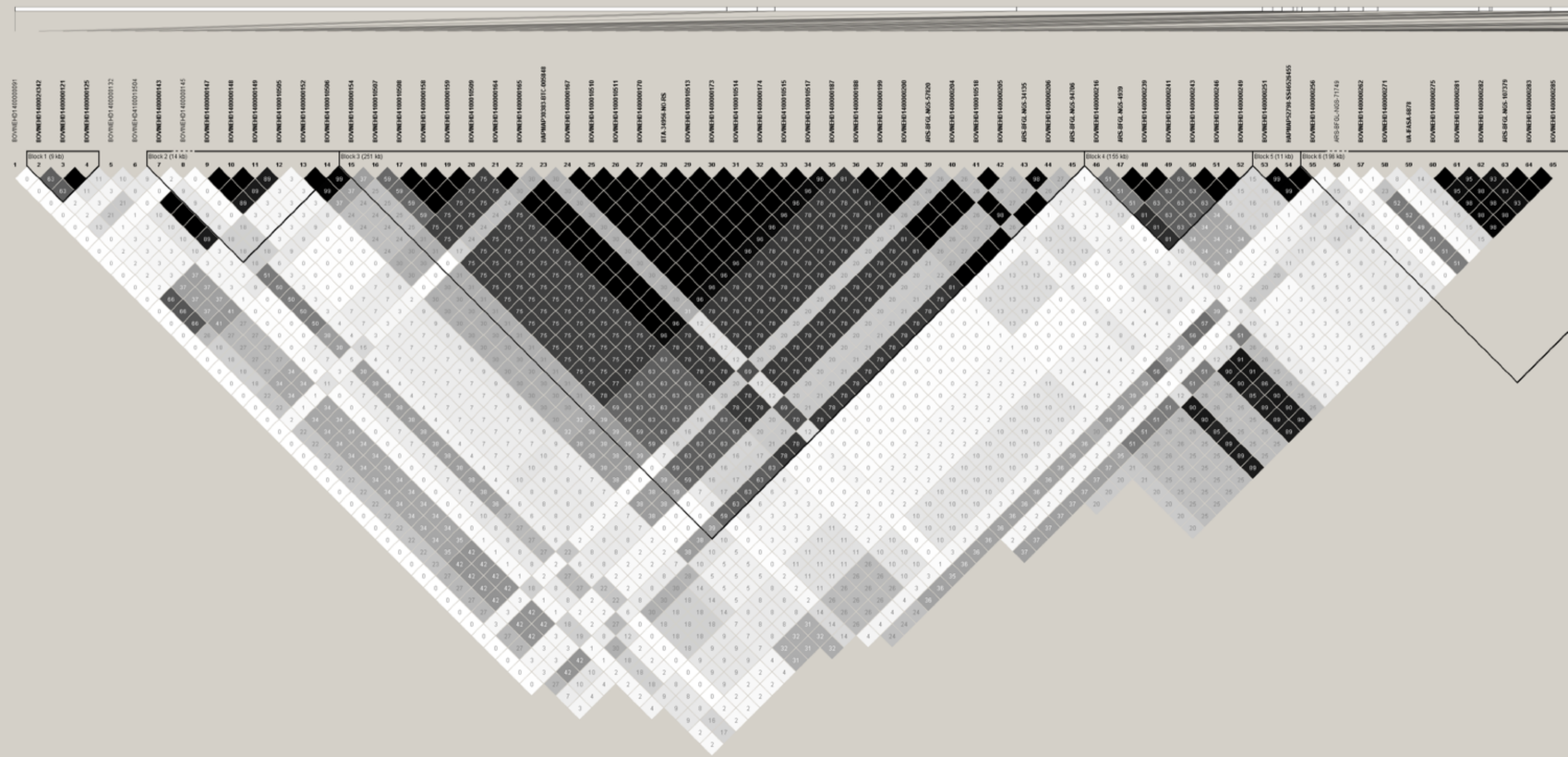
		B Locus		
		B	b	Row
	A Locus			Total
	A	43	27	70
	a	2	28	30
Column Total		45	55	100

$$D = (43 - 70 * 45/100)/100 = 0.1105$$

43/100-(70/100)*(45/100)

$$r = 0.1105/\sqrt{0.7 * 0.3 * 0.55 * 0.45} = .4878$$

Linkage Disequilibrium



0 correlation is white
100 is black

Bart Buitenhuis, BMC Genetics, 2016
https://doi.org/10.1186/s12863-016-0421-2

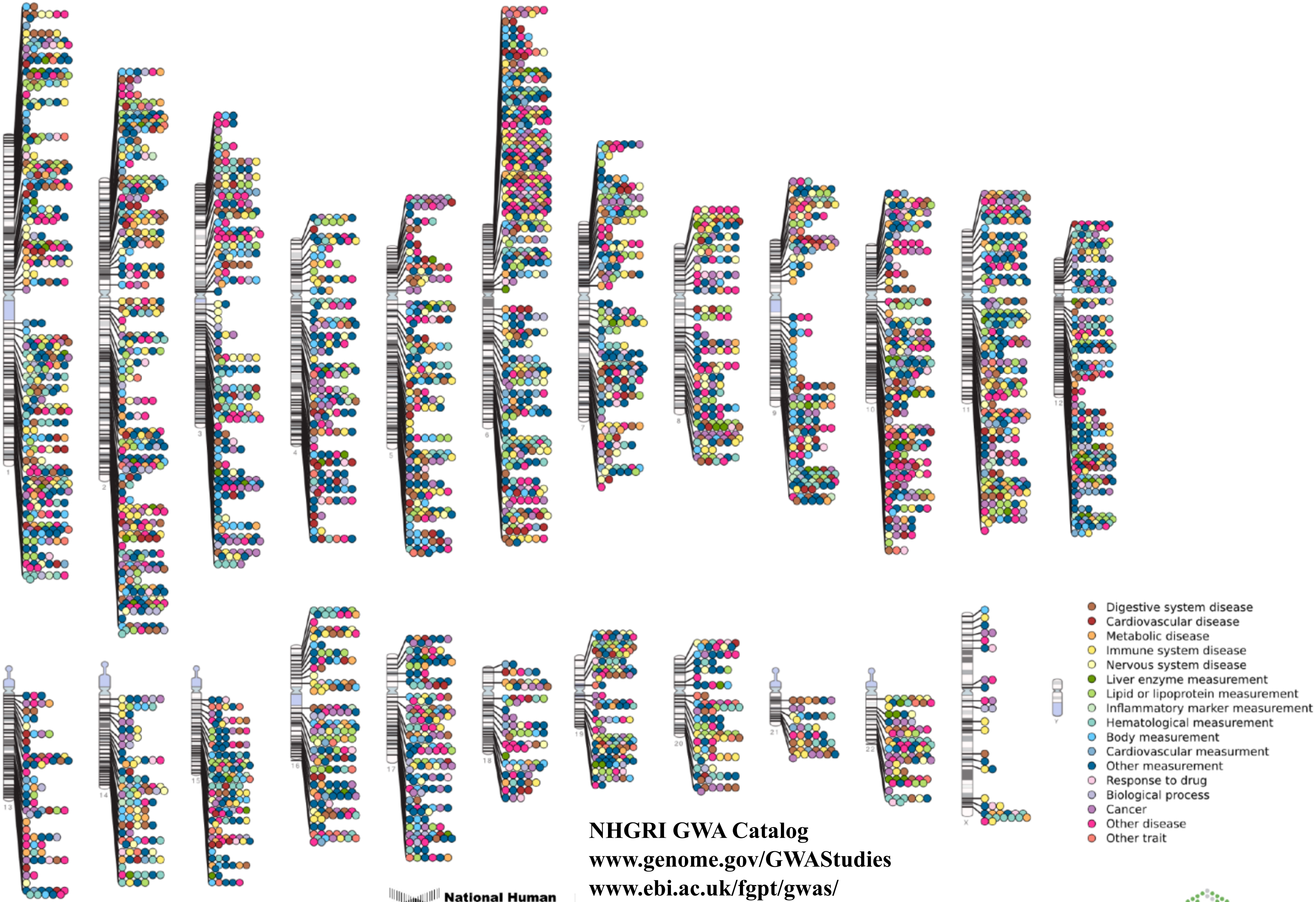
Genetic Association Analysis

- LD is the basis of Association Analysis
- For association analysis we observe a trait and a marker locus, usually not the DSL.
- Test association between marker and a trait using ordinary association analysis methods.
- Null hypothesis is no association between marker and the trait---Guilt by Association. Rejection implies DSL is in LD with the marker.
- ‘Spurious’ association occurs when 2 loci are not linked, i.e., association between two loci that are on different chromosomes (possibly due to population substructure)

Features of Genetic Association Analysis

- Can use isolated cases and unrelated individuals or families
- Much more powerful than linkage analysis
- Requires hundreds of thousands of markers for a whole genome analysis (500k-1,000k +); a severe multiple comparison problem

Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human
Genome Research
Institute

EMBL-EBI