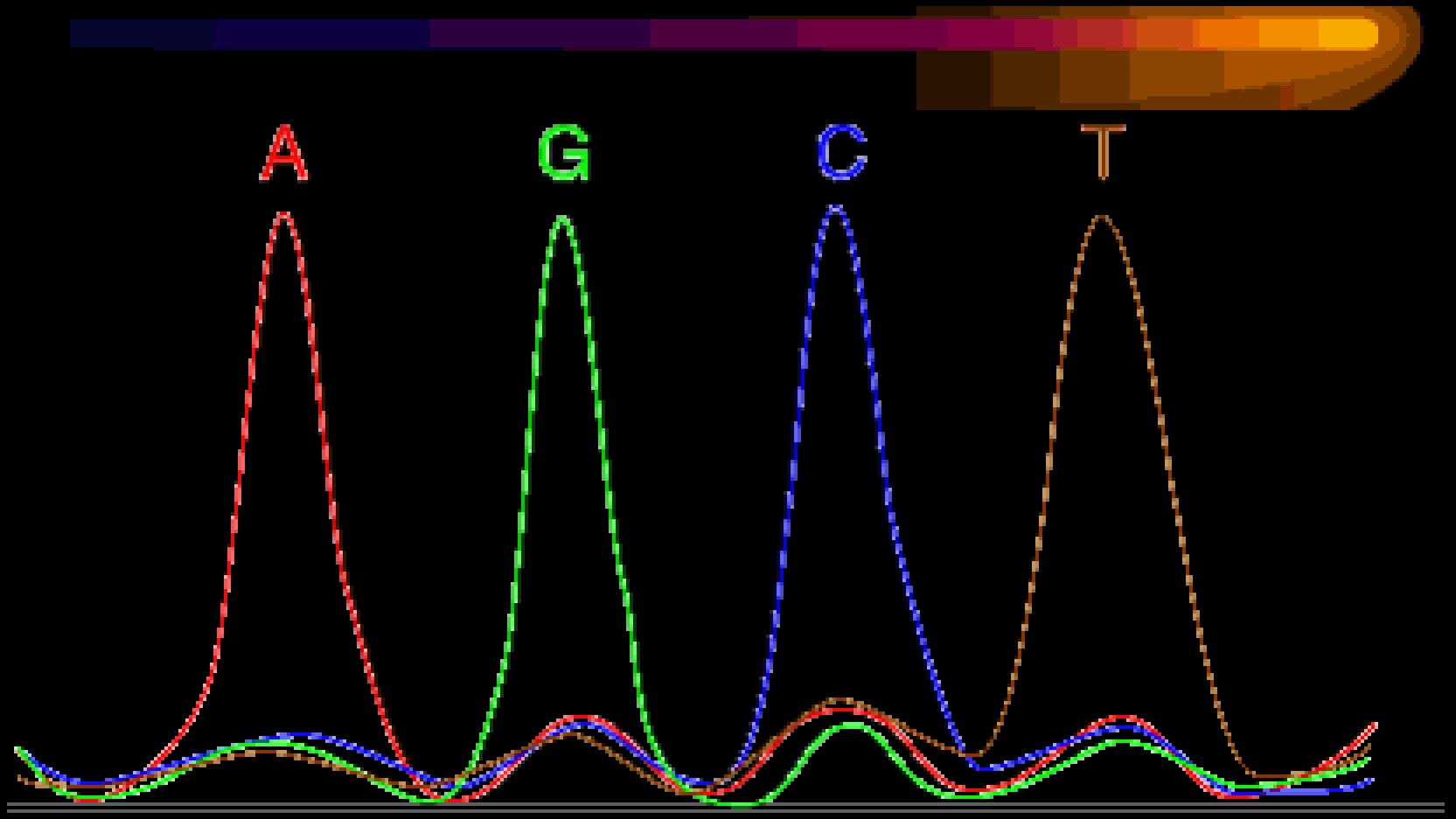
BENG 183

Trey Ideker

Genetics 101: The Basis of Genome Association



Review of Mendelian genetics

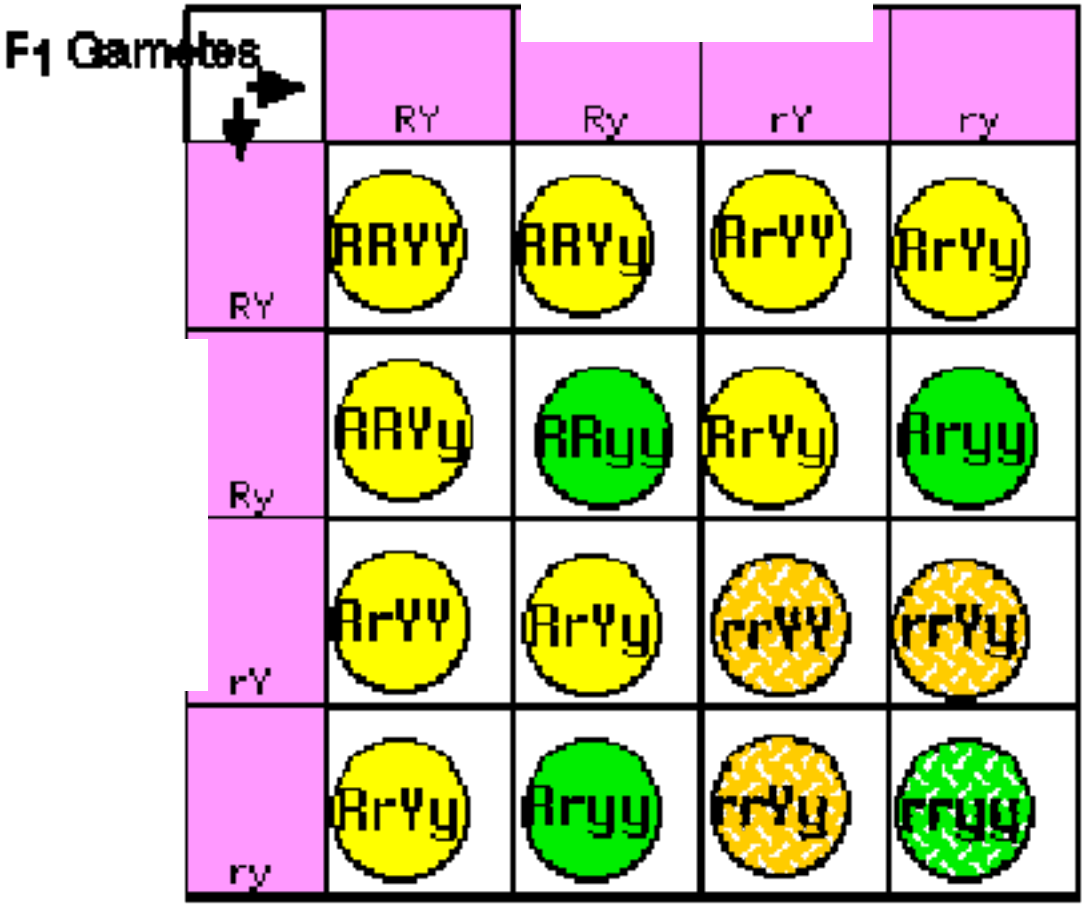
- Gregor Mendel analyzed the patterns of inheritance of seven pairs of contrasting traits in the domestic pea plant. As an example pair:
- P1: He mated a plant that was homozygous for round (**RR**) yellow (**YY**) seeds with one that was homozygous for wrinkled (**rr**) green (**yy**) seeds.
- F1: All the offspring were dihybrids, i.e., heterozygous for each pair of alleles (**RrYy**).
- All seeds were round and yellow, showing that the genes for round and yellow are **dominant**.

Mendelian genetics (2)

- F2: Mendel then crossed the RrYy dihybrids.
- If round seeds must always be yellow and wrinkled seeds must be green (linked genes), then this would have produced a typical **monohybrid** cross →→→
- But in fact, the F2s had seeds with all combinations:

Round-yellow	9/16
Round-green	3/16
Wrinkled-yellow	3/16
Wrinkled-green	1/16

	RY	ry
RY	RRYY	RrYy
ry	RrYy	rryy



Gregor Mendel

Results

round-yellow : round-green : wrinkled-yellow : wrinkled-green

9 : 3 : 3 : 1

Percentage of recombinants = 50%

Mendel's Rule of Independent Assortment

The inheritance of one pair of factors (genes) is independent of the inheritance of the other pair.

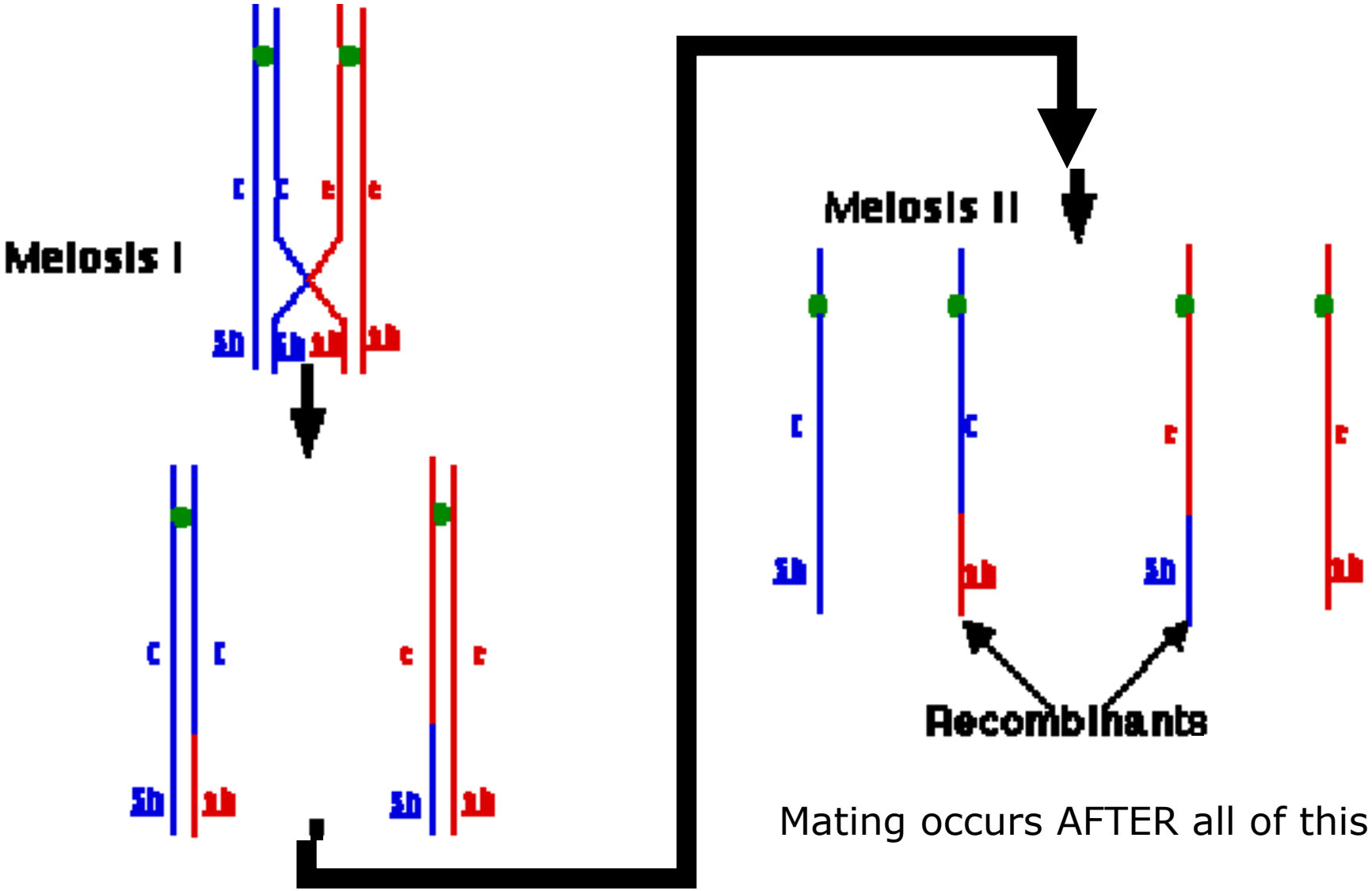
- In other words, the percentage of recombinants is 50%
- Today, we know that this rule holds only if one of two conditions is met:
  - The genes are on separate chromosomes
  - The genes are widely separated on the same chromosome.

- Mendel was lucky in that every pair of genes he studied met one requirement or the other!!!
- In fact, the rule does not apply to many matings of dihybrids. In many cases, two alleles inherited from one parent show a strong tendency to stay together as do those from the other parent.
- This phenomenon is called **linkage**.

Genetic distance in centiMorgans

- The percentage of recombinants formed by F1 individuals can range from 0-50%.
- 0% is seen if two loci map to the same gene.
- 50% is seen for two loci on separate chromosomes (independent assortment).
- Between these extremes, the higher the percentage of recombinants, the greater the genetic distance separating the two loci.
- The percent of recombinants is arbitrarily chosen as the genetic distance in **centimorgans** (**cM**), named for the pioneering geneticist Thomas Hunt Morgan.

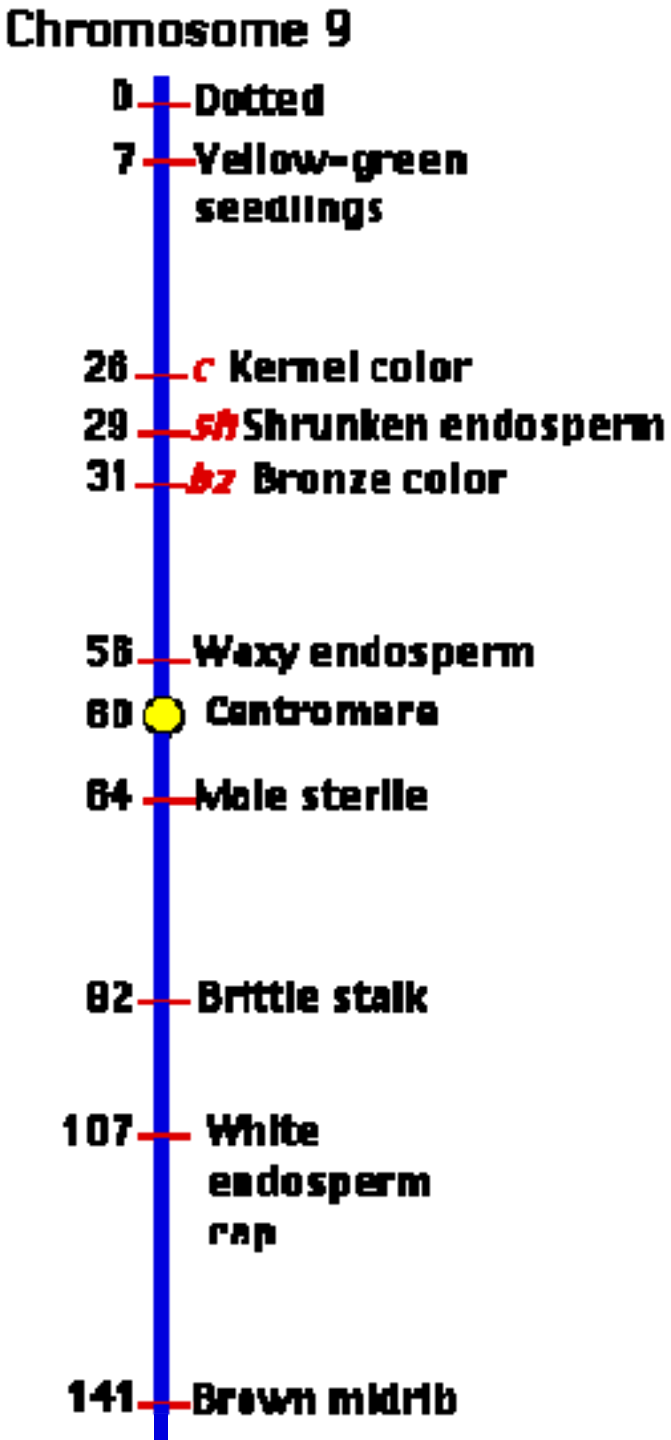
Genetic recombination is due to the physical process of *crossing-over*



Mating occurs AFTER all of this

Genetic Maps

- Chromosome maps prepared by counting phenotypes are called ***genetic maps***.
- Maps have been prepared for many eukaryotes, including corn, Drosophila, the mouse, and tomato.
- Controlled matings are not practicable in humans, but map positions are estimated by examining family trees (pedigrees)
- A genetic map of chr. 9 of the corn plant (**Zea mays**) is shown on the right with distances in cM.
- Note distances >50cM. How?



Example

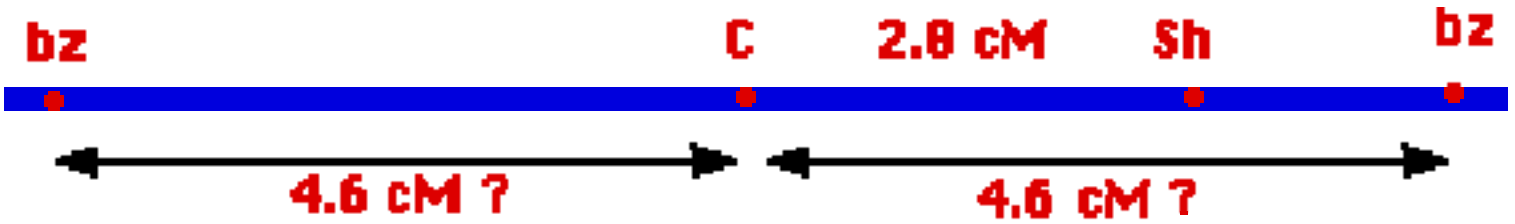
- Corn plants are scored for three traits:

C/c	colored/colorless seeds
Bz/bz	bronze/non-bronze stalk
Sh/sh	Smooth/shrunken

- The following F1 heterozygote self-crosses are performed:

(C/c; Bz/bz)	X	(C/c; Bz/bz)	4.6 cM
(Sh/sh; C/c)	X	(Sh/sh; C/c)	2.8 cM

- What does the genetic map look like?



Gather more data– now can the map be determined?

(Sh/sh; Bz/bz)	X	(Sh/sh; Bz/bz)	1.8 cM

Problems with genetic maps

- Recombination frequency underestimates genetic distance for larger distances, due to higher order cross-over events, i.e. double and triple crossovers between markers
  This is overcome with 3 pt. mapping (homework)
- The probability of a crossover is not uniform along the entire length of the chromosome.
  - Crossing over is inhibited in some regions (e.g., near the centromere).
  - Some regions are "hot spots" for recombination (for reasons that are not clear). Approximately 80% of genetic recombination in humans is confined to just one-quarter of our genome.
- In humans, the frequency of recombination of loci on most chromosomes is higher in females than in males. Therefore, genetic maps of female chromosomes are longer than those for males.

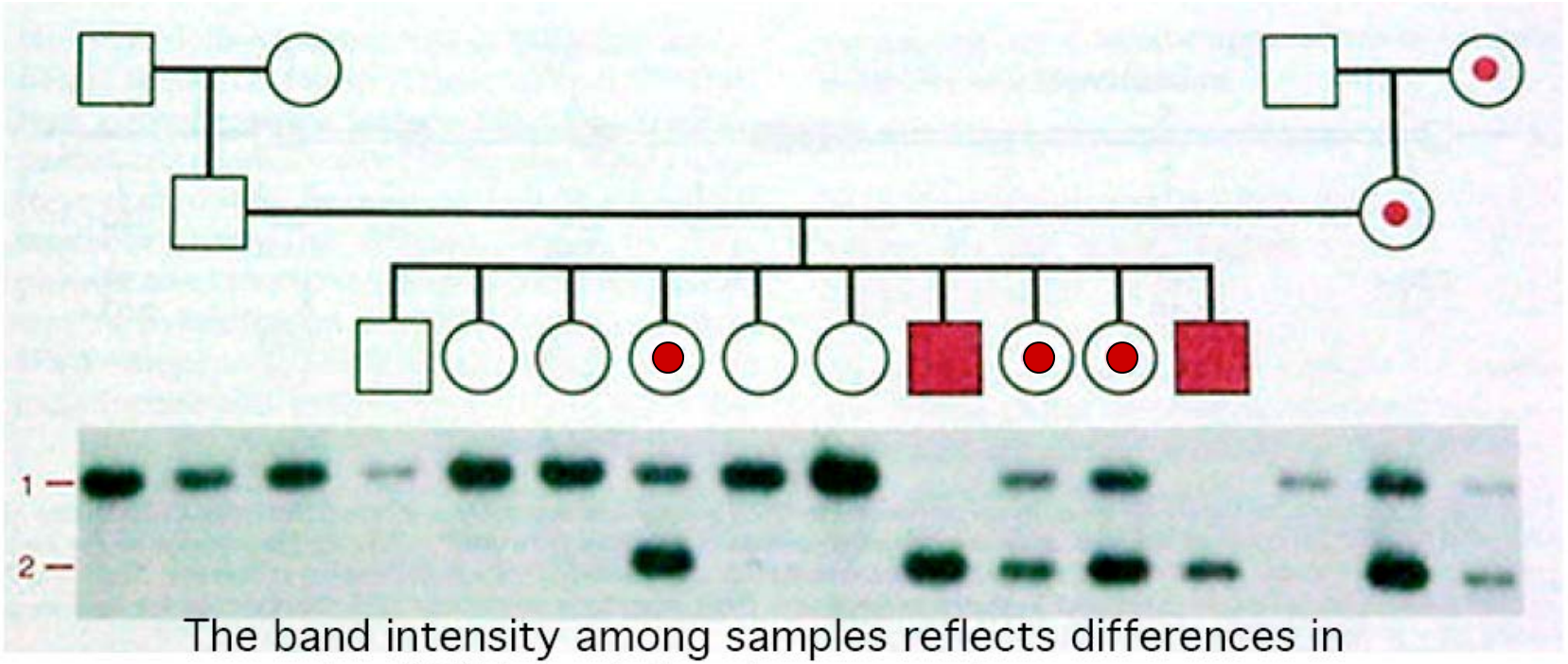
Genetic vs. physical maps

- The loci in **genetic maps** are simply parts of the DNA molecule that create the observed phenotype.
- Knowing the DNA sequence (or at least the ordering of contigs) directly gives the order/spacing of genes.
- Maps drawn in this way are called **physical maps**.
- As a very rough rule of thumb,
  1 cM genetic distance ~ 1 MB of DNA.

Gene linkage mapping

- Tries to find a common inheritance pattern between a chromosomal region/marker and a disease phenotype
- Requires genotyping on large, multigeneration pedigrees
- Coarse mapping with sparse markers <10 Mb
- At greater distances linkage generally does not occur due to frequent recombination events
- At lesser distances all loci are typically linked and thus are indistinguishable from one another

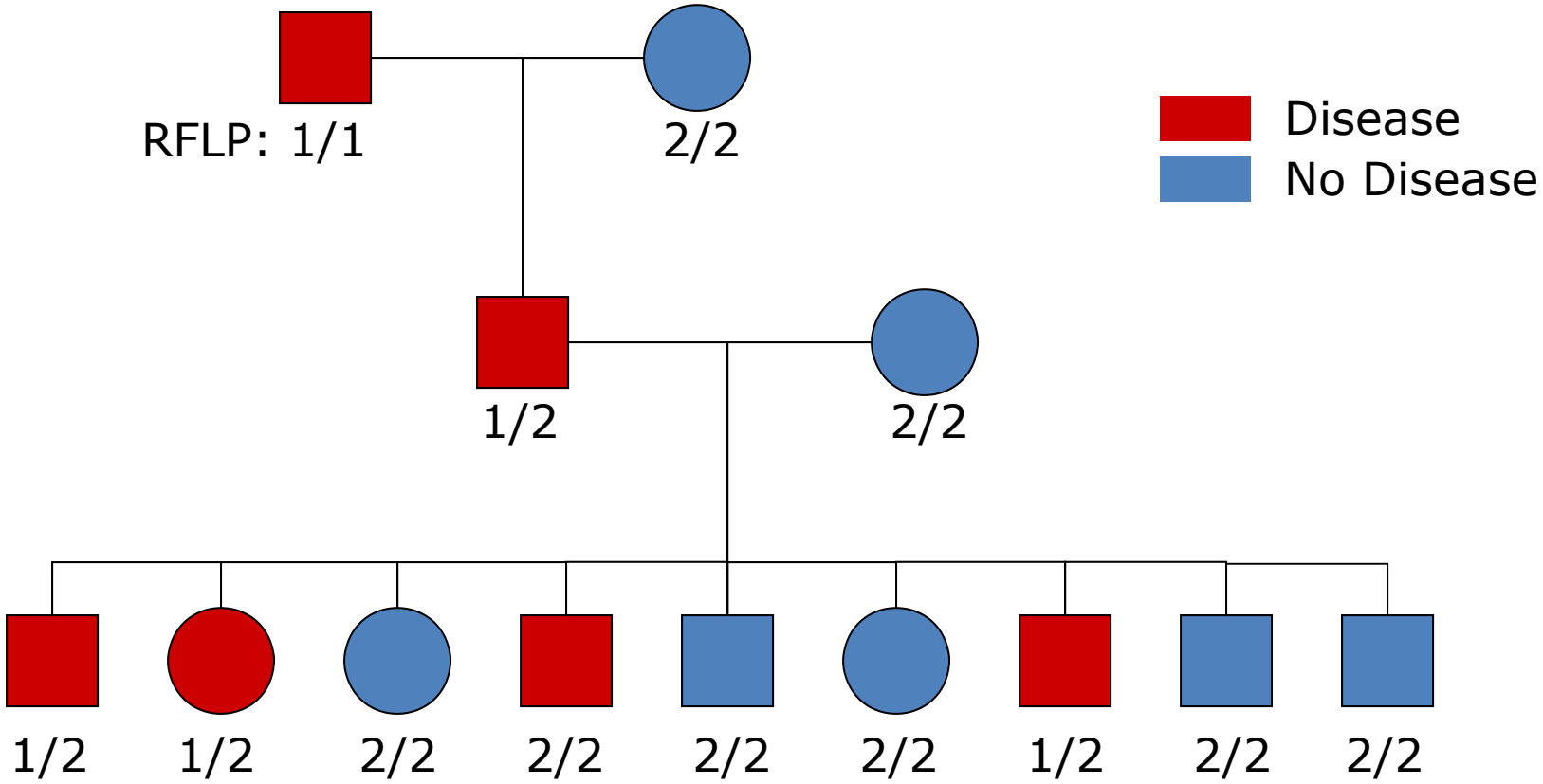
Linkage analysis in a 3 generation pedigree



The band intensity among samples reflects differences in amounts of DNA loaded, not copy number.

Solid red indicates the disease phenotype; dot means carrier
The gel is the result of RFLP analysis—note variants 1 and 2
Is this a recessive or dominant gene? Autosomal or sex linked?
What is the *penetrance*? This is Pr(disease phenotype | disease genotype)

Autosomal dominant linkage



Computing a Log Odds (LOD) score

- From the last slide: 3 affected offspring carry RFLP1, while 1 affected and 5 unaffected offspring do not carry it.
- If the two loci (RFLP1 and the disease gene) are unlinked, the probability of the above observation is $(0.5)^9 = 0.002$
- If the two loci are in fact linked and the chance of crossover is 10% (called the *recombination fraction*), the probability of the observed pattern of disease is $(0.9)^8(0.1)^1 = 0.04$
- For each individual we are computing:

  Pr(D|Model) = Pr(disease state | RFLP1 state ^ linkage)

  Pr(D|Random) = Pr(disease state | RFLP1 state ^ non-linkage)

Computing a LOD score

$$LOD = \log_{10} \frac{\prod_{children} \Pr(D \mid M)}{\prod_{children} \Pr(D \mid R)}$$

$$= \sum_{children} \log\left(\Pr(D \mid M)\right) - \sum_{children} \log\left(\Pr(D \mid R)\right)$$

- In the prev. example the LOD score is $\log_{10}(0.04 / 0.002) = 1.3$
- A LOD > 3.0 is generally considered significant
- Alternatively, *parametric* analysis models modes of inheritance (domnt,recssv,x-linked,etc.)

Table of LOD scores

If recombination fraction is unknown, optimize this parameter

	Recombination Fraction (%)				
	0	10	20	30	40
Family A	2.7	2.3	1.8	1.3	0.7
Family B	-∞	1.0	0.9	0.6	0.3
Total	-∞	3.3	2.7	1.9	1.0

Box 12.1 from Primrose and Twyman

Gene association mapping

- Also looks at common inheritance but in populations of unrelated individuals – No pedigrees required
- Fine mapping with dense markers at least every 60 kb
- Beyond this distance loci are generally in linkage equilibrium
- Also called Linkage Disequilibrium mapping
- Can be used in conjunction with the coarser grained map of linkage analysis

Odds Ratio (OR)

- The ratio of the odds of an event occurring in one population (the “cases”) to the odds of it occuring in another population (“the controls”).

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1}$$

Logistic Regression

- The log odds is also called the logit function.
- The logistic function is the inverse logit:

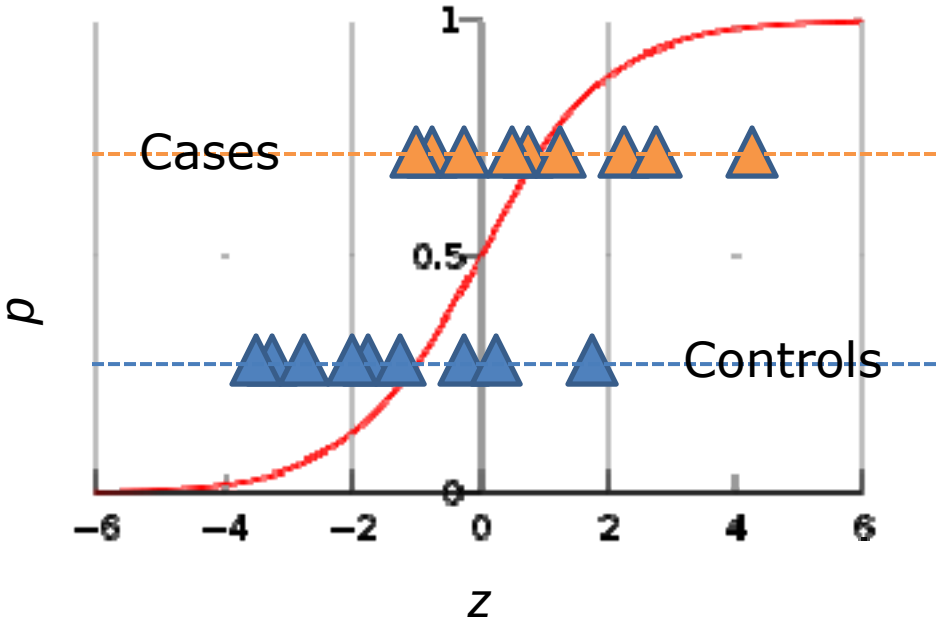
$$z = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$p = \text{logit}^{-1}(z) = \frac{1}{1+e^{-z}}$$

$$z = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$p = \frac{1}{1+e^{-z}}$$

The input is z and the output is $f(z)$. The logistic function is useful because it can take as an input any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1. The variable z represents the exposure to some set of independent variables, while $f(z)$ represents the probability of a particular outcome, given that set of explanatory variables.



Linkage and LD analysis in tandem

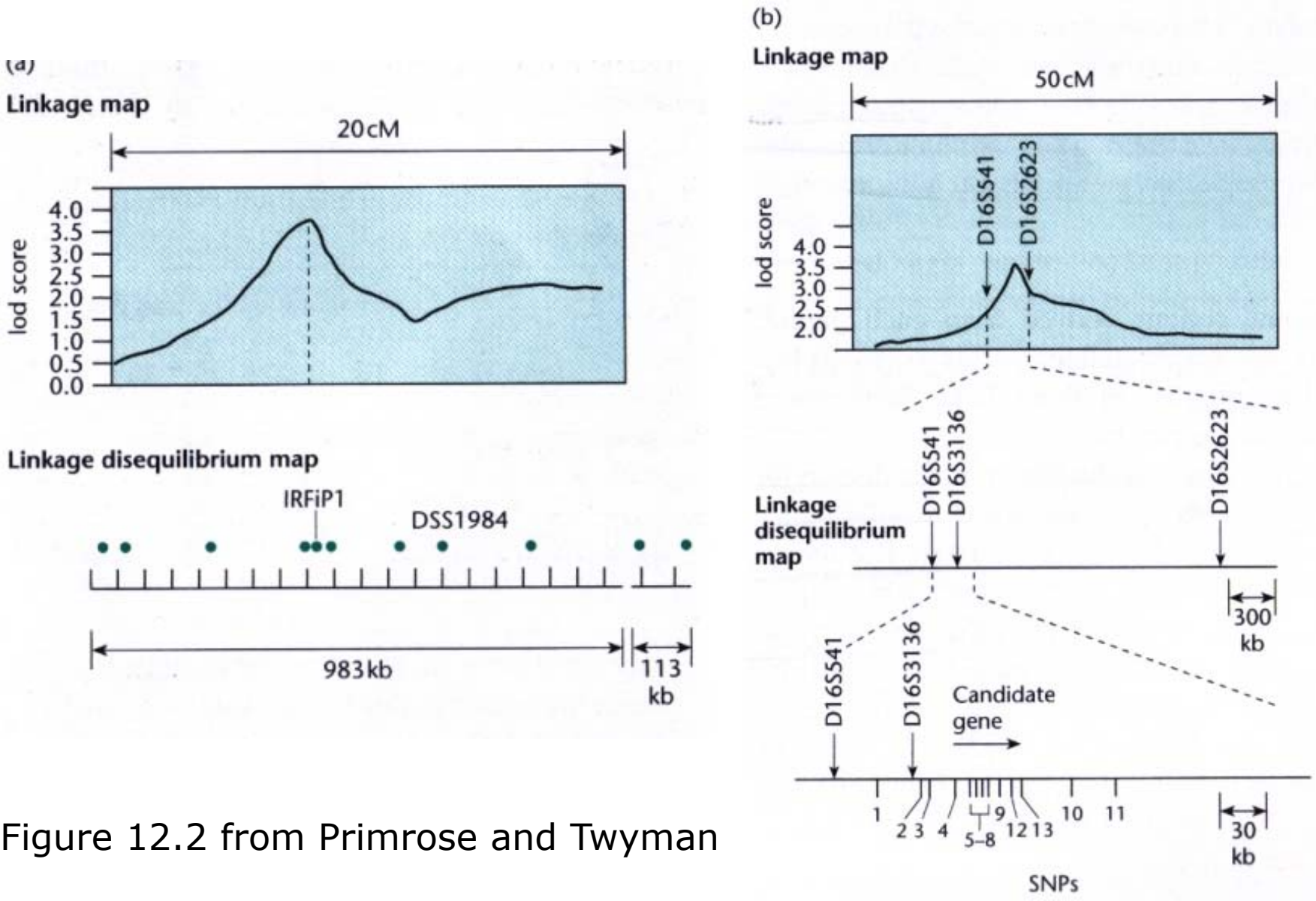


Figure 12.2 from Primrose and Twyman

LD mapping elucidates our evolutionary origins

- In Northern European populations, LD extends for ~60kb
- In a Nigerian African population, LD extends for ~5kb, a much shorter distance
- What do we conclude from these findings?

Haplotype mapping

- A *haplotype* is a pattern of SNPs in a contiguous stretch of DNA
- Due to linkage disequilibrium, SNPs are typically inherited in discrete haplotype blocks spanning 10-100kb
- Greatly simplifies LD analysis, because rather than screen all SNPs in a region, we just need to screen a few and the rest can be inferred
- A complete human haplotype map is still underway

Example haplotype map

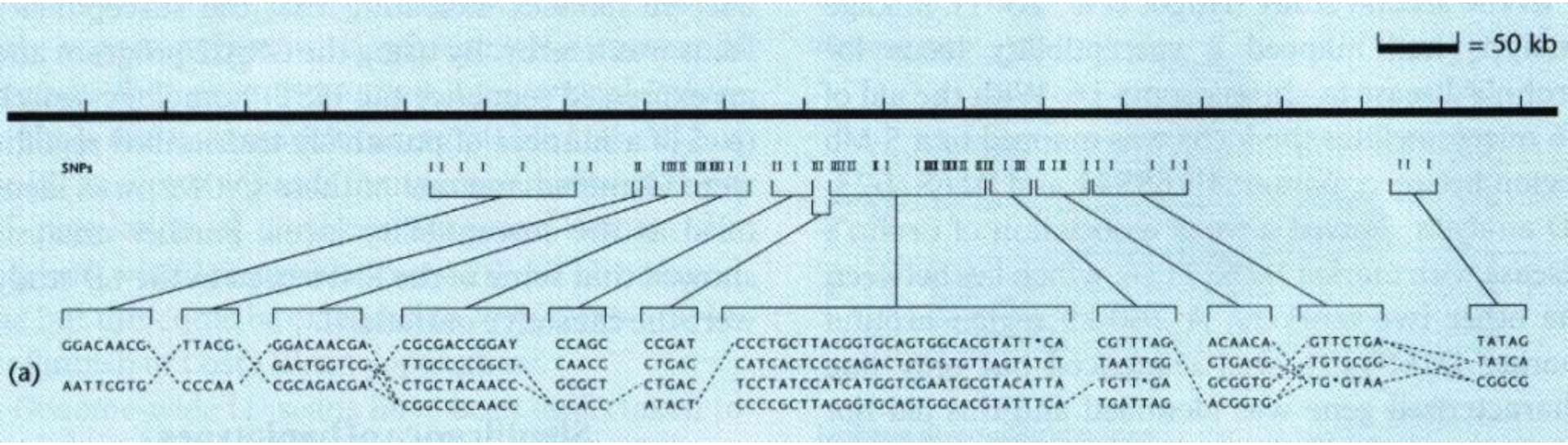
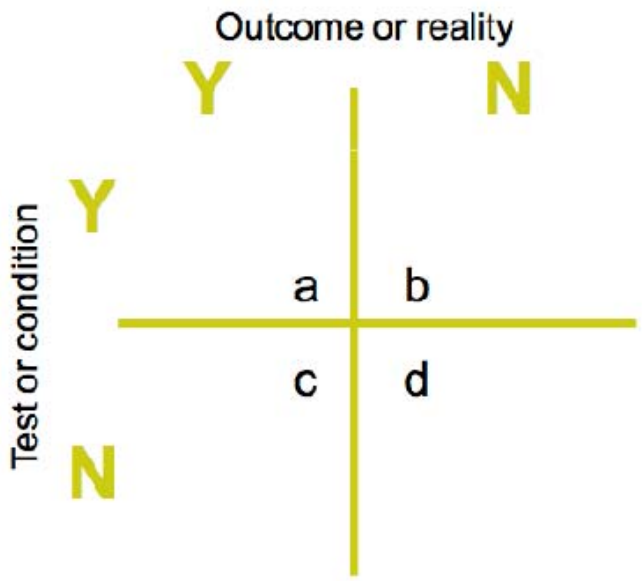


Figure 12.4 from Primrose and Twyman

Sensitivity, specificity, odds ratio, likelihood ratio, and all that…



b= type 1 error
c= type 2 error

Parameter	expression
Sensitivity	a/a+c
Specificity	d/d+b
Prevalence	a+b+c+d
NPV	d/d+c
PPV	a/a+b
OR	ad/cb
OR	(a/b) / (c/d)
RR	(a/a+b) / (c/c+d)
LR+	sen/1-spec
LR-	1-sen/spec