Introduction to Genetics

Bruce Walsh lecture notes
Liege May 2011 course
version 22 May 2011

Topics

- Darwin and Mendel
- Mendel genetics
  - Mendel's experiments
  - Mendel's laws
- Genes and chromosomes
  - Linkage
  - Prior probability of linkage
- Genes and DNA
  - Basics of DNA structure
  - Types of Genetic markers

Darwin & Mendel

- Darwin (1859)
  - Instant Classic, major immediate impact
  - Problem: Model of Inheritance
    - Darwin assumed Blending inheritance
    - Offspring = average of both parents
    - $z_o = (z_m + z_f)/2$
    - Fleming Jenkin (1867) pointed out problem
      - $\mathrm{Var}(z_o) = \mathrm{Var}[(z_m + z_f)/2] = (1/2)\ \mathrm{Var}(\text{parents})$
      - Hence, under blending inheritance, half the variation is removed each generation and this must somehow be replenished by mutation.

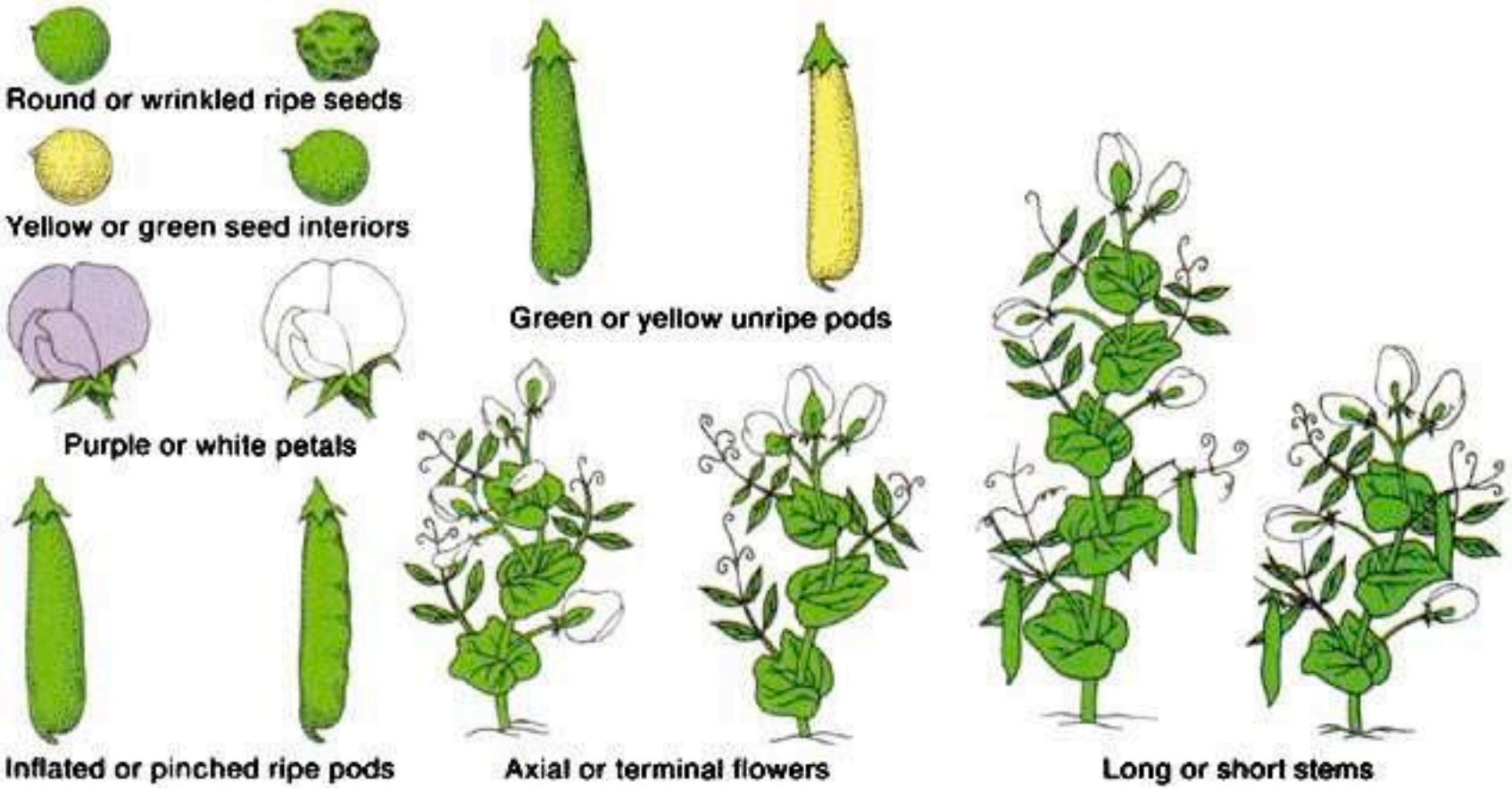
Mendel

Mendel

- Mendel (1865),
- No impact, paper essentially ignored
  - Ironically, Darwin had an apparently unread copy in his library
  - Why ignored? Perhaps too mathematical for 19th century biologists
- The rediscovery in 1900 (by three independent groups)
- Mendel's key idea: Genes are discrete particles passed on intact from parent to offspring

Mendel's experiments with the Garden Pea

7 traits examined



Mendel crossed a pure-breeding yellow pea line with a pure-breeding green line.

Let P1 denote the pure-breeding yellow (parental line 1)
P2 the pure-breed green (parental line 2)

The F1, or first filial, generation is the cross of P1 x P2 (yellow x green).

All resulting F1 were yellow

The F2, or second filial, generation is a cross of two F1's

In F2, 1/4 are green, 3/4 are yellow

This outbreak of variation blows the theory of blending inheritance right out of the water.

Mendel also observed that the P1, F1 and F2 Yellow lines behaved differently when crossed to pure green

P1 yellow x P2 (pure green) --> all yellow

F1 yellow x P2 (pure green) --> 1/2 yellow, 1/2 green

F2 yellow x P2 (pure green) --> 2/3 yellow, 1/3 green

Mendel's explanation

Genes are discrete particles, with each parent passing one copy to its offspring.

Let an allele be a particular copy of a gene. In Diploids, each parent carries two alleles for every gene

Pure Yellow parents have two Y (or yellow) alleles

We can thus write their genotype as YY

Likewise, pure green parents have two g (or green) alleles

Their genotype is thus gg

Since there are lots of genes, we refer to a particular gene by given names, say the pea-color gene (or locus)

Each parent contributes one of its two alleles (at random) to its offspring

Hence, a YY parent always contributes a Y, while a gg parent always contributes a g

An individual carrying only one type of an allele (e.g. yy or gg) is said to be a homozygote

In the F1, YY x gg --> all individuals are Yg

An individual carrying two types of alleles is said to be a heterozygote.

The phenotype of an individual is the trait value we observe

For this particular gene, the map from genotype to phenotype is as follows:

YY --> yellow

Yg --> yellow

gg --> green

Since the Yg heterozygote has the same phenotypic value as the YY homozygote, we say (equivalently)

Y is dominant to g, or

g is recessive to Y

Explaining the crosses

F1 x F1 -> Yg x Yg

Prob(YY) = yellow(dad)*yellow(mom) = (1/2)*(1/2)

Prob(gg) = green(dad)*green(mom) = (1/2)*(1/2)

Prob(Yg) = 1-Pr(YY) - Pr(gg) = 1/2

Prob(Yg) = yellow(dad)*green(mom) + green(dad)*yellow(mom)

Hence, Prob(Yellow phenotype) = Pr(YY) + Pr(Yg) = 3/4

Prob(green phenotype) = Pr(gg) = 1/4

Review of terms (so far)

- Gene
- Locus
- Allele
- Homozygote
- Heterozygote
- Dominant
- Recessive
- Genotype
- Phenotype

In class problem (5 minutes)

Explain why F2 yellow x P2 (pure green)
- -> 2/3 yellow, 1/3 green

F2 yellows are a mix, being either Y*g* or YY

$$\text{Prob(F2 yellow is } Yg) = \frac{\text{Pr(yellow} \mid Yg)*\text{Pr}(Yg \text{ in F2})}{\text{Pr(Yellow)}}$$

$$= (1*\ 1/2)/(3/4) = 2/3$$

2/3 of crosses are Y*g* x *gg* -> 1/2 Y*g* (yellow), 1/2 *gg* (green)

1/3 of crosses are YY x *gg* -> all Y*g* (yellow)

Pr(yellow) = (2/3)*(1/2) + (1/3) = 2/3

Dealing with two (or more) genes

For his 7 traits, Mendel observed Independent Assortment

The genotype at one locus is independent of the second

RR, Rr - round seeds, rr - wrinkled seeds

Pure round, green (RR*gg*) x pure wrinkled yellow (rrYY)

F1 --> RrY*g* = round, yellow

What about the F2?

Let R- denote RR and Rr. R- are round. Note in F2, Pr(R-) = 1/2 + 1/4 = 3/4

Likewise, Y- are YY or Yg, and are yellow

Phenotype	Genotype	Frequency
Yellow, round	Y-R-	(3/4)*(3/4) = 9/16
Yellow, wrinkled	Y-rr	(3/4)*(1/4) = 3/16
Green, round	ggR-	(1/4)*(3/4) = 3/16
Green, wrinkled	ggrr	(1/4)*(1/4) = 1/16

Or a 9:3:3:1 ratio

Probabilities for more complex genotypes

Cross AaBBCcDD X aaBbCcDd

What is Pr(aaBBCCDD)?

Under independent assortment,

= Pr(aa)*Pr(BB)*Pr(CC)*Pr(DD)

= (1/2*1)*(1*1/2)*(1/2*1/2)*(1*1/2) = $1/2^5$

What is Pr(AaBbCc)?

= Pr(Aa)*Pr(Bb)*Pr(Cc) = (1/2)*(1/2)*(1/2) = 1/8

Mendel was wrong: Linkage

Bateson and Punnet looked at

flower color: P (purple) dominant over p (red)

pollen shape: L (long) dominant over l (round)

Phenotype	Genotype	Observed	Expected
Purple long	P-L-	284	215
Purple round	P-ll	21	71
Red long	ppL-	21	71
Red round	ppll	55	24

Excess of PL, pl gametes over Pl, pL

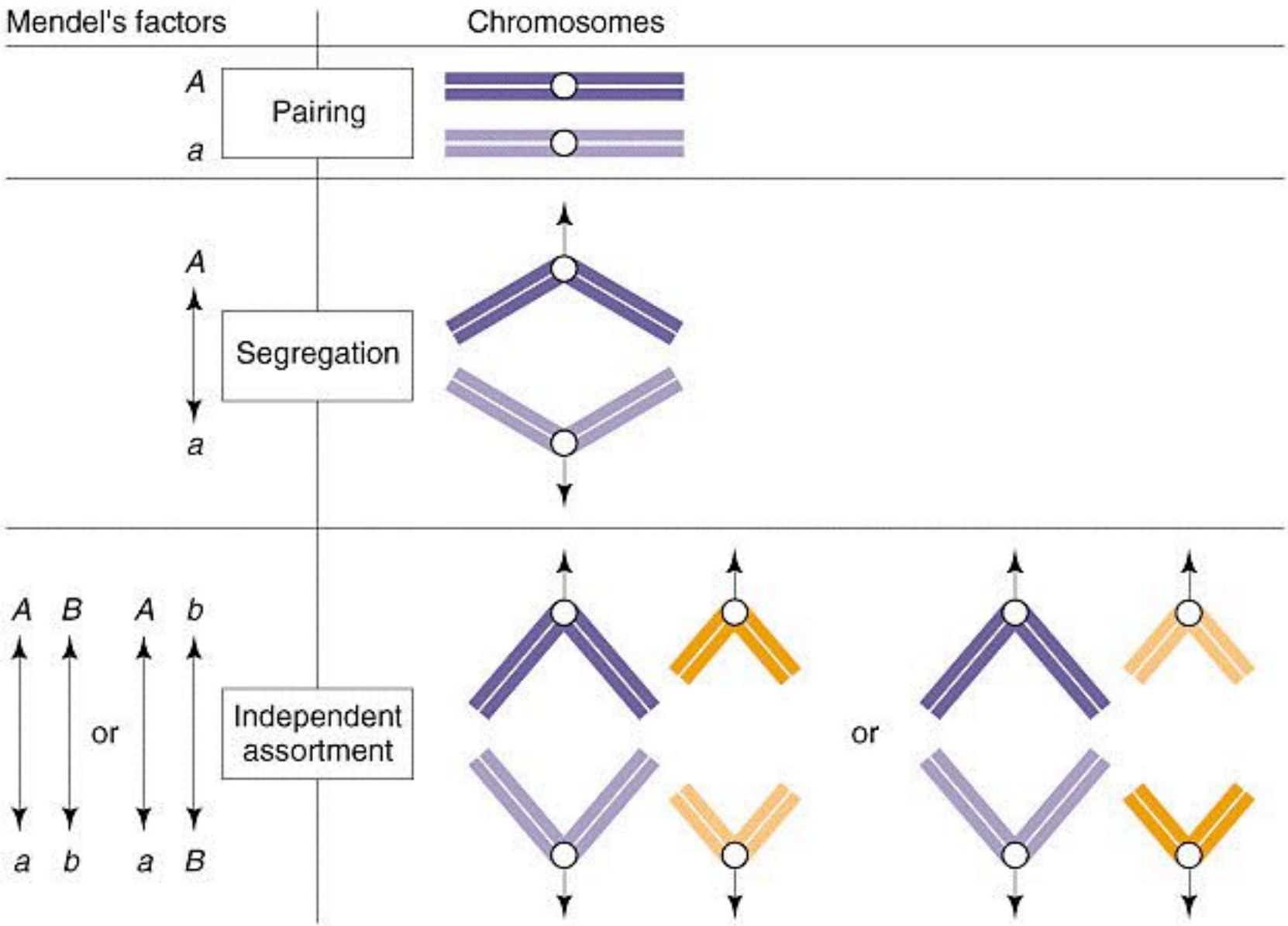
Departure from independent assortment

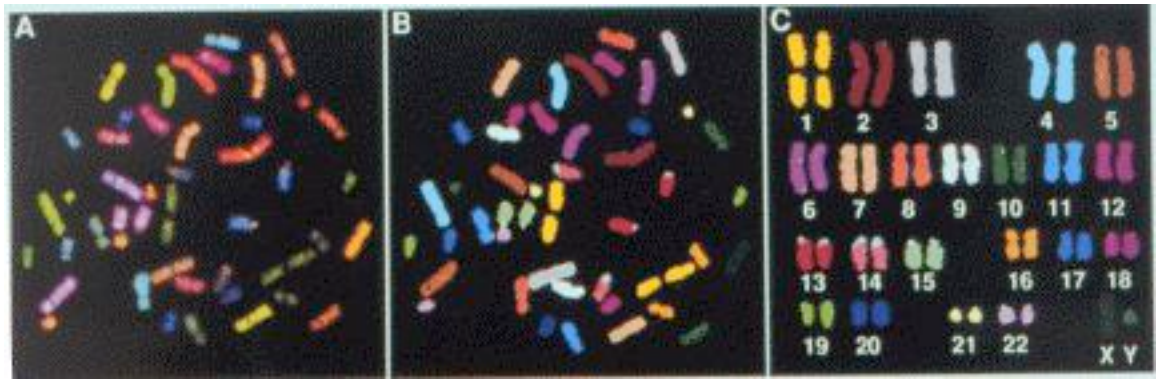
Interlude: Chromosomal theory of inheritance

Early light microscope work on dividing cells revealed small (usually) rod-shaped structures that appear to pair during cell division. These are chromosomes.

It was soon postulated that Genes are carried on chromosomes, because chromosomes behaved in a fashion that would generate Mendel's laws.

We now know that each chromosome consists of a single double-stranded DNA molecule (covered with proteins), and it is this DNA that codes for the genes.


Mendel's factors
Chromosomes
A
a
Pairing
A
a
Segregation
A B A b
or
a b a B
Independent assortment
or



Humans have 23 pairs of chromosomes (for a total of 46)

22 pairs of autosomes (chromosomes 1 to 22)

1 pair of sex chromosomes -- XX in females, XY in males

Humans also have another type of DNA molecule, namely the mitochondrial DNA genome that exists in tens to thousands of copies in the mitochondria present in all our cells

mtDNA is usual in that it is strictly maternally inherited. Offspring get only their mother's mtDNA.

Linkage

If genes are located on different chromosomes they (with very few exceptions) show independent assortment.

Indeed, peas have only 7 chromosomes, so was Mendel lucky in choosing seven traits at random that happen to all be on different chromosomes? Exercise: compute this probability.

However, genes on the same chromosome, especially if they are close to each other, tend to be passed onto their offspring in the same configuration as on the parental chromosomes.

Consider the Bateson-Punnet pea data

Let PL / pl denote that in the parent, one chromosome carries the P and L alleles (at the flower color and pollen shape loci, respectively), while the other chromosome carries the p and l alleles.

Unless there is a recombination event, one of the two parental chromosome types (PL or pl) are passed onto the offspring. These are called the parental gametes.

However, if a recombination event occurs, a PL/pl parent can generate Pl and pL recombinant chromosomes to pass onto its offspring.

Let c denote the recombination frequency --- the probability that a randomly-chosen gamete from the parent is of the recombinant type (i.e., it is not a parental gamete).

For a PL/pl parent, the gamete frequencies are

Gamete type	Frequency	Expectation under independent assortment
PL	(1-c)/2	1/4
pl	(1-c)/2	1/4
pL	c/2	1/4
Pl	c/2	1/4

Parental gametes in excess, as (1-c)/2 > 1/4 for c < 1/2

Gamete type	Frequency	Expectation under independent assortment
PL	(1-c)/2	1/4
pl	(1-c)/2	1/4
pL	c/2	1/4
Pl	c/2	1/4

Recombinant gametes in deficiency, as c/2 < 1/4 for c < 1/2

Expected genotype frequencies under linkage

Suppose we cross PL/pl X PL/pl parents

What are the expected frequencies in their offspring?

$$\Pr(PPLL) = \Pr(PL|father)*\Pr(PL|mother)$$
$$= [(1-c)/2]*[(1-c)/2] = (1-c)^2/4$$

Likewise, $\Pr(ppll) = (1-c)^2/4$

Recall from previous data that freq(ppll) = 55/381 =0.144

Hence, $(1-c)^2/4 = 0.144$, or $c = 0.24$

A (slightly) more complicated case

Again, assume the parents are both PL/pl.
Compute Pr(PpLl)

Two situations, as PpLl could be PL/pl or Pl/pL

$$\Pr(PL/pl) = \Pr(PL|dad)*\Pr(pl|mom) + \Pr(PL|mom)*\Pr(pl|dad)$$
$$= [(1-c)/2]*[(1-c)/2] + [(1-c)/2]*[(1-c)/2]$$

$$\Pr(Pl/pL) = \Pr(Pl|dad)*\Pr(pL|mom) + \Pr(Pl|mom)*\Pr(pl|dad)$$
$$= (c/2)*(c/2) + (c/2)*(c/2)$$

Thus, $\Pr(PpLl) = (1-c)^2/2 + c^2/2$

Generally, to compute the expected genotype probabilities, need to consider the frequencies of gametes produced by both parents.

Suppose dad = Pl/pL, mom = PL/pl

$$\Pr(PPLL) = \Pr(PL|dad)*\Pr(PL|mom)$$
$$= [c/2]*[(1-c)/2]$$

Notation: when PL/pl, we say that alleles P and L are in coupling

When parent is Pl/pL, we say that P and L are in repulsion

Genetic Maps and Mapping Functions

The unit of genetic distance between two markers is the recombination frequency, c (also called θ)

If the phase of a parent is AB/ab, then 1-c is the frequency of "parental" gametes (e.g., AB and ab), while c is the frequency of "nonparental" gametes (e.g.. Ab and aB).

A parental gamete results from an EVEN number of crossovers, e.g., 0, 2, 4, etc.

For a nonparental (also called a recombinant) gamete, need an ODD number of crossovers between A & b e.g., 1, 3, 5, etc.

Hence, simply using the frequency of "recombinant" (i.e. nonparental) gametes UNDERESTIMATES the m number of crossovers, with E[m] > c

In particular, c = Prob(odd number of crossovers)

Mapping functions attempt to estimate the expected number of crossovers m from observed recombination frequencies c

When considering two linked loci, the phenomena of interference must be taken into account

The presence of a crossover in one interval typically decreases the likelihood of a nearby crossover

Suppose the order of the genes is A-B-C.

If there is no interference (i.e., crossovers occur independently of each other) then

Probability(odd number of crossovers btw A and C)

Even number of crossovers btw A & B, Odd number between B & C

$$c_{AC} = c_{AB}(1 - c_{BC}) + (1 - c_{AB})\, c_{BC} = \boxed{c_{AB} + c_{BC} - 2c_{AB}\, c_{BC}}$$

odd number in A-B, even number in B-C

We need to assume independence of crossovers in order to multiply these two probabilities

When interference is present, we can write this as

$$c_{AC} = c_{AB} + c_{BC} - 2(1 - \delta) c_{AB}\, c_{BC}$$

Interference parameter

δ = 1 --> complete interference: The presence of a crossover eliminates nearby crossovers

δ = 0 --> No interference. Crossovers occur independently of each other

Mapping functions. Moving from c to m

Haldane's mapping function (gives Haldane map distances)

Assume the number k of crossovers in a region follows a Poisson distribution with parameter m

This makes the assumption of NO INTERFERENCE

Pr(Poisson = k) = λ^k Exp[-λ]/k!

λ = expected number of successes

$$c = \sum_{k=0}^{\infty} p(m, 2k+1) = e^{-m} \sum_{k=0}^{\infty} \frac{m^{2k+1}}{(2k+1)!} = \frac{1 - e^{-2m}}{2}$$

Prob(Odd number of crossovers)

$$c = \sum_{k=0}^{\infty} p(m, 2k+1) = e^{-m} \sum_{k=0}^{\infty} \frac{m^{2k+1}}{(2k+1)!} = \frac{1 - e^{-2m}}{2}$$

Odd number

Relates recombination fraction c to expected number of crossovers m

This gives the estimated Haldane distance as

$$m = -\frac{\ln(1-2c)}{2}$$

Usually reported in units of Morgans or Centimorgans (Cm)

One morgan --> m = 1.0. One cM --> m = 0.01

The Prior Probability of Linkage

Morton (1955), in the context of linkage analysis in humans, introduced the concept of a Posterior Error Rate, or PER

PER = probability that a test declared significant is a false positive, PER = Pr(false positive | significant test)

The screening paradox: type I error control may not lead to a suitably low PER

With PER, conditioning on the test being significant
As opposed to conditioning on the hypothesis being a null, as occurs with type I error control (α)

Let α be the Type 1 error, β the type 2 error (1- β = power)
And π be the fraction of null hypothesis, then from
Bayes' theorem

$$\text{PER} = \Pr(\text{false positive} \mid \text{significant})$$

$$\text{PER} = \frac{\Pr(\text{false positive} \mid \text{null True}) * \Pr(\text{null})}{\Pr(\text{significant test})}$$

Since there are 23 pairs of human chromosomes, Morton argued that two randomly-chosen genes had a 1/23 (roughly 5%) prior probability of linkage, i.e. $\pi = 0.95$

Assuming α type I error of a = 0.05 and 80% power (β = 0.2), the expected PER is

$$\frac{0.05*0.95}{0.05*0.95 + 0.8*0.05} = 0.54$$

Hence, even with a 5% type-I error control, a random significant test has a 54% chance of being a false-positive.

This is because most of the hypotheses are expected to null. If we draw 1000 random pairs of loci, 950 are expected to be unlinked, and we expect 950 * 0.05 = 47.5 of these to show a false-positive. Conversely, only 50 are expected to be linked, and we would declare 50 * 0.80 = 40 of these to be significant, so that 47.5/87.5 of the significant results are due to false-positives.

Molecular Markers

You and your neighbor differ at roughly 22,000,000 nucleotides (base pairs) out of the roughly 3 billion bp that comprises the human genome

Hence, LOTS of molecular variation to exploit

SNP -- single nucleotide polymorphism. A particular position on the DNA (say base 123,321 on chromosome 1) that has two different nucleotides (say G or A) segregating

STR -- simple tandem arrays. An STR locus consists of a number of short repeats, with alleles defined by the number of repeats. For example, you might have 6 and 4 copies of the repeat on your two chromosome 7s

SNPs vs STRs

SNPs

Cons: Less polymorphic (at most 2 alleles)

Pros: Low mutation rates, alleles very stable
Excellent for looking at historical long-term associations (association mapping)

STRs

Cons: High mutation rate

Pros: Very highly polymorphic
Excellent for linkage studies within an extended Pedigree (QTL mapping in families or pedigrees)