

Lecture 2 (part a)

Basic Plant Genetics

Bruce Walsh lecture notes
Tucson Winter Institute
7 - 9 Jan 2013

Overview

- Ploidy
- Linkage
- Linkage disequilibrium (LD)
- Genetic markers
- Mapping functions
- Organelle inheritance
- Mating systems and types of crosses
- Gene actions
 - Dominance and Epistasis
 - Pleiotropy

Ploidy

- Most animals are diploid ($2n$), with their gametes (eggs, sperm) containing a haploid set of n chromosomes
- Polyploids are much more common in plants.
- Allopolyploids consist of haploid sets from two (or more) species
 - e.g., an allotetraploid is $AABB$,
 - One allohexaploid is $AABBCC$
 - Generally speaking, allopolyploids largely behave as diploids, i.e., each pollen/egg gets a haploid set from each of the founding species
- Autopolyploids have multiple haploid sets from the same species
 - Autoteraploids ($4n$) and autohexaploids ($6n$)
 - these give pollen and eggs with two ($2n$) and three ($3n$) (respectively) copies of each homologous chromosome

Plant	Type of Polyploidy	Ploidy	Chromosome Number
Potato	Autopolyploid	$4n$	48
Banana	Autopolyploid	$3n$	33
Peanut	Autopolyploid	$4n$	40
Sweet potato	Autopolyploid	$6n$	90
Tobacco	Allopolyploid	$4n$	48
Cotton	Allopolyploid	$4n$	52
Wheat	Allopolyploid	$6n$	42
Oats	Allopolyploid	$6n$	42
Sugar cane	Allopolyploid	$8n$	80
Strawberry	Allopolyploid	$8n$	56

Linkage

- Independent assortment for unlinked genes
- Linkage
- Computing expected genotypic frequencies from linkage

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 (RRgg) x pure wrinkled yellow (rrYY)

F₁ --> RrYg = 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

Mendel was wrong: Linkage

Bateson and Punnett 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

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?

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-Punnett 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?

$$\begin{aligned}\Pr(\text{PPLL}) &= \Pr(\text{PL}|\text{father}) * \Pr(\text{PL}|\text{mother}) \\ &= [(1-c)/2] * [(1-c)/2] = (1-c)^2/4\end{aligned}$$

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

Recall from previous data that $\text{freq}(\text{ppll}) = 55/381 = 0.144$

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

A (slightly) more complicated case

Again, assume the parents are both PL/pl.

Compute $\Pr(\text{PpLl})$

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

$$\begin{aligned}\Pr(\text{PL/pl}) &= \Pr(\text{PL}|\text{dad})\Pr(\text{pl}|\text{mom}) + \Pr(\text{PL}|\text{mom})\Pr(\text{pl}|\text{dad}) \\ &= [(1-c)/2]*[(1-c)/2] + [(1-c)/2]*[(1-c)/2]\end{aligned}$$

$$\begin{aligned}\Pr(\text{Pl/pL}) &= \Pr(\text{Pl}|\text{dad})\Pr(\text{pL}|\text{mom}) + \Pr(\text{Pl}|\text{mom})\Pr(\text{pl}|\text{dad}) \\ &= (c/2)*(c/2) + (c/2)*(c/2)\end{aligned}$$

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

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

$$\begin{aligned}\Pr(\text{PPLL}) &= \Pr(\text{PL}|\text{dad}) * \Pr(\text{PL}|\text{mom}) \\ &= [c/2] * [(1-c)/2]\end{aligned}$$

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**

Linkage Disequilibrium

- Under linkage equilibrium, the frequency of gametes is the product of allele frequencies,
 - e.g. $\text{Freq}(AB) = \text{Freq}(A) * \text{Freq}(B)$
 - A and B are **independent** of each other
- If the linkage phase of parents in some set or population departs from random (alleles not independent) , linkage disequilibrium (LD) is said to occur
- The amount D_{AB} of disequilibrium for the AB gamete is given by
 - $D_{AB} = \text{Freq}(AB) \text{ gamete} - \text{Freq}(A) * \text{Freq}(B)$
 - $D > 0$ implies AB gamete from frequent than expected
 - $D < 0$ implies AB less frequent than expected

Dynamics of D

- Under random mating in a large population, allele frequencies do not change. However, gamete frequencies do if there is any LD
- The amount of LD decays by $(1-c)$ each generation
 - $D(t) = (1-c)^t D(0)$
- The expected frequency of a gamete (say AB) is
 - $\text{Freq}(AB) = \text{Freq}(A) * \text{Freq}(B) + D$
 - $\text{Freq}(AB \text{ in gen } t) = \text{Freq}(A) * \text{Freq}(B) + (1-c)^t D(0)$

Molecular Markers

In the molecular era, genetic maps are based not on alleles with large phenotypic effects (i.e., green vs. yellow peas), but rather on molecular markers

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

Even with whole-genome sequencing, sites are still classified into these two classes (plus other types) ¹⁸

SNPs

SNPs vs STRs

Cons: Less polymorphic (at most 2 alleles)

Pros: Low mutation rates, alleles very stable

Excellent for looking at historical long-term associations (association mapping)

Cheap to score 100,000s (+) on a single SNP Chip

STRs

Cons: High mutation rate

Pros: Very highly polymorphic (more information/site)

Excellent for linkage studies within an extended pedigree (QTL mapping in families or pedigrees)

Genetic maps

- Published genetic maps give the distances between molecular markers along a chromosome in terms of map units (m , expected number of crossovers between them), rather than their recombination frequencies c . Why?
 - c is not additive over loci, while m is
 - Hence, m is a more natural metric
 - Transition from an observed c to an estimated m requires a **mapping function**, which requires an assumption about how interference works

Genetic Maps and Mapping Functions

The unit of genetic distance between two markers is the **recombination frequency**, c

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, $c = \text{Prob}(\text{odd number of crossovers})$

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

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

↑
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

$\delta = 1$ --> complete interference: The presence of a crossover eliminates nearby crossovers

$\delta = 0$ --> 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

$$\Pr(\text{Poisson} = k) = \lambda^k \text{Exp}[-\lambda]/k!$$

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)

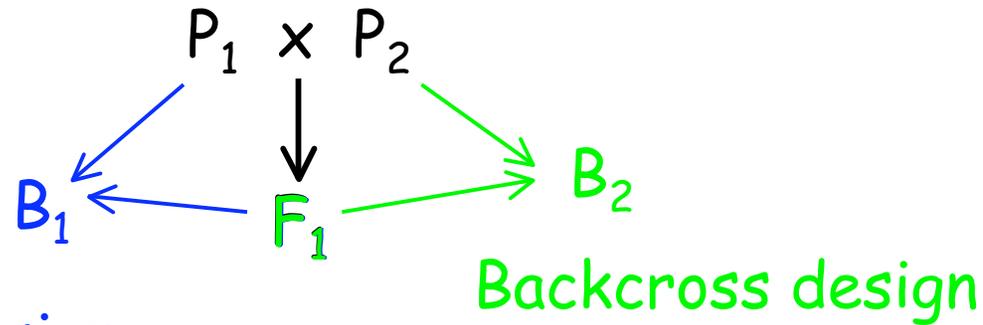
One morgan --> $m = 1.0$. One cM --> $m = 0.01$

Organelle genetics

- With autosomal loci, each parent contributes an equal number of chromosomes
- However, the mitochondrial and chloroplast genomes are only passed from the mother.
 - While these have a small number of genes (mtDNA ~ 20, cpDNA ~ 50-100), they can still have phenotypic effects
 - Example: cytoplasmic sterility factors on mtDNA used in maize to avoid having to detassel pollen plants

Systems of matings and types of crosses

- Types of crosses
 - F_1 , F_2 , Backcrosses
 - F_k , Advanced intercross (AIC) lines
 - Isogenic/inbred lines
 - Recombinant inbred lines (RILS)
 - Selfing
 - S_k lines
 - Doubled haploids



Backcross design

$B_1(2) = B_1 \times P_1$ Repeating backcrossing to the P_1 gives

$$B_1(k) = B_{k-1} \times P_1 \text{ lines}$$

Fraction genetic contributions from each parent

Cross	% P_1	% P_2
F_1	50	50
B_1	75	25
B_2	25	75
$B_1(k)$	$1 - (1/2)^{k+1}$	$(1/2)^{k+1}$
$B_2(k)$	$(1/2)^{k+1}$	$1 - (1/2)^{k+1}$

F₂-based crosses

- Randomly mating the F₁ generates the F₂
 - These can also be generated by selfing each F₁
- **Isogenic** (or **inbred**) lines are created by taking a set of F₁ individuals and selfing each for 5-10 generations to create a series of inbred lines
 - Generates a series of pure lines that capture some of the initially segregating variation
 - When generated following a cross, also called RILs
- **Advanced intercross lines** (AIC) are created by randomly-mating the F₂ line for multiple generations (AIC(k) = F_k = k generations of random mating)
 - Has the effect of expanding the genetic map in the AIC(k), recombination rate between two markers $\sim c \cdot k$

Selfing, Doubled Haploids

- If two inbred lines are crossed, all of the F_1 are heterozygotes. If we self the F_1 for k generations, then the fraction of loci that are heterozygotes is $(1/2)^k$.
 - Less than 1% in the F_7 , 0.09% in F_{10}
 - S_k lines refer to k generations of selfing an F_2 , e.g., with only selfing $S_k = F_{k+2}$
 - $S_0 =$ The F_2 from selfing an F_1 line,
- **Doubled haploids, DH**, (the doubling of a haploid set in a gamete) produces fully inbred individuals in one generation
 - DH lines capture most of the initial LD (only a single generation of recombination)
 - Selfed-generated lines further decay some LD, but not as efficiently as random mating.

Selfing and Favorable Alleles

- Suppose inbred lines 1 and 2 are each fixed for five favorable alleles not found in the other.
 - In the F_1 , all individuals carry at least one favorable allele at each locus
 - In the F_2 , the probability a locus contains at least one favorable allele is $\text{Pr}(\text{favorable homozygote}) + \text{Pr}(\text{favorable heterozygote}) = (1/4) + (1/2) = 3/4$.
 - $\text{Pr}(\text{all 10 loci do}) = (3/4)^{10} = 0.056$
 - If fully inbred, $\text{Pr}(\text{at least one favorable allele at a locus}) = 1/2$
 - $\text{Pr}(\text{all 10 loci fixed for favorable allele}) = (1/2)^{10} = 1/1024$,
 - Roughly 57 times less likely than an F_2 .
 - Hence, while inbred lines build up loci fixed for both favorable alleles, they have less loci with favorable alleles than an F_1 or F_2 .

Effects of selection

- Now suppose that selection increases the frequency of each favorable allele from 0.5 to 0.9
 - For an inbred, now $\text{Pr}(\text{all loci fixed for favorable alleles}) = 0.9^{10} = 0.35$
 - For a random-mating population, $\text{Pr}(\text{all loci contain a favorable allele}) =$
 - $(1-0.1^2)^{10} = 0.904$

Types of Gene Action

- At a single gene, we can see **dominance**
 - The heterozygote has a phenotype that is different from the average of the two homozygotes
 - Interaction between the two alleles at a locus
- We can also have **pleiotropy**, where a single gene influences two or more **traits**.
- When two (or more) genes influence the same trait, the possibility of epistasis exists
 - The two-locus phenotype is not simply the sum of the two single-locus phenotype

Epistasis

- Consider the two-locus genotype $A_i A_j B_k B_l$
- Let $G_{ij} = G_{ij}$ denote the average deviation between an $A_i A_j$ individual and the population mean, same for G_{kl}
- If $G_{ijkl} = u + G_{ij} + G_{kl}$, i.e., the two-locus genotypic value is the sum of each single locus genotypic values (based on deviations from the mean u), then we have additive genotypic values across loci (while dominance might still occur at either locus)
 - If this is NOT the case, we have that epistasis occurs --- the two-locus genotype departs from the average contribution of both single loci.
 - Dominance = interaction between alleles at the SAME locus
 - Epistasis = interaction between alleles at DIFFERENT loci

Example

		0	5	10
		AA	Aa	aa
10	BB	10	15	20
10	Bb	10	15	20
5	bb	5	10	15

B is dominant to b, A is additive (no dominance)
However, no epistasis, as phenotypic value is
(B phenotype) + 5*(# of a alleles), namely the sum
of the two genotypic values at each locus