4-30 Recitation

DG Lectures 19 & 20

QTLs & Human Genetics

Announcements

- Pset 5 due this Thursday (5-1)
- Exam 2 next Tuesday (5-6)
  - 2 double-sided sheets of notes
- Office Hours next Monday instead of Tuesday
- No recitations or regular OHs after exam
- Project Presentations May 13 and 15 – all students will peer review

Outline

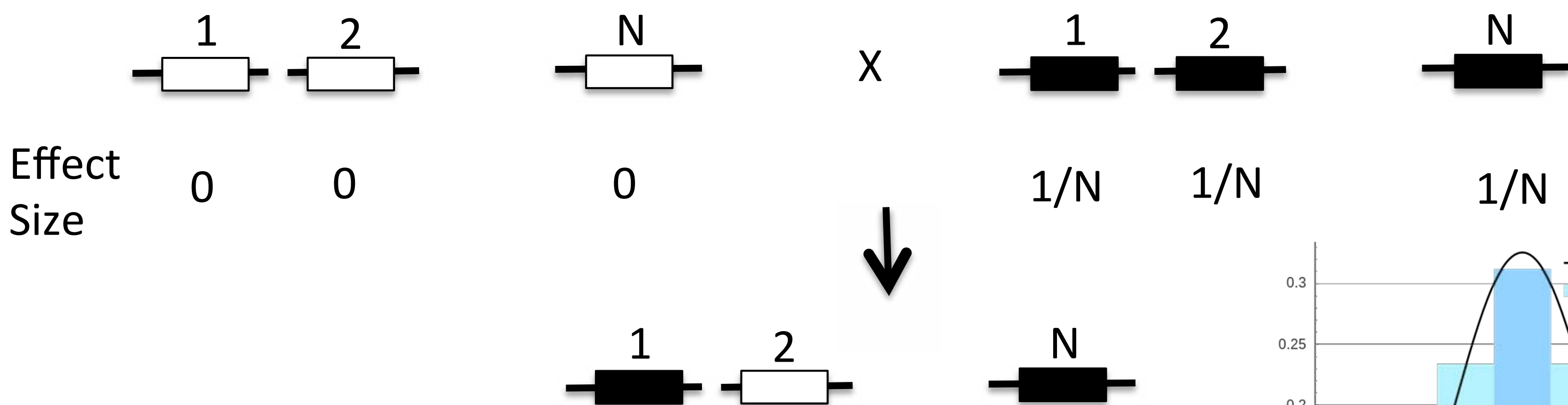
- Quantitative Trait Loci
  - Simple genetic model (haploid, unlinked)
  - Genotype-phenotype interactions
    - Broad-sense and narrow-sense heritability, sources of variance
  - LOD scores
  - Bloom *et al.* 2013 & missing sources of heritability
- Human Genetics
  - Testing for SNP/phenotype associations
  - Linkage Disequilibrium
  - Variant Phasing
  - Hardy-Weinberg Equilibrium

Genotype to Phenotype

- Phenotype: organisms observable characteristics or traits
  - Qualitative: dead/alive, tall/short
  - Quantitative: Growth rate, height, gene expression
- Quantitative Trait locus (loci) – a marker that is associated with a quantitative trait
  - eQTL (expression quantitative trait locus) – marker associated with gene expression
  - eQTLs are often SNPs (single nucleotide polymorphisms) in the population
    - Can be in *cis* (within ~Kbs on the same chromosome) or in *trans* (1+Mb away or on different chromosome)
    - Often cell-type specific

Haploid, unlinked genetic model

- N loci that each contribute equally (1/N) to the trait
- Haploid = organism has 1 copy of each allele
- Unlinked = loci are on different chromosomes or far enough apart on the same chromosome so crossing over (recombination) can always occur
  - Each locus is therefore inherited independently
- Child randomly inherits maternal or paternal copy



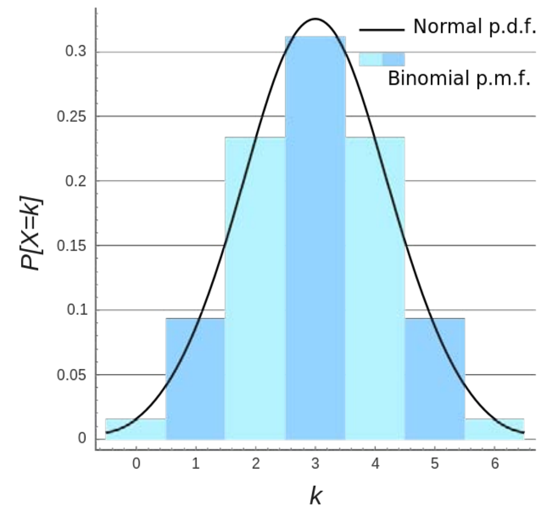
Binomial model of # of black alleles x inherited:

$$p(x,N) = \binom{N}{x}(1-.5)^{N-x}.5^{x}$$

Here x is the phenotypic value from 0 (no alleles) to 1 (all black alleles)

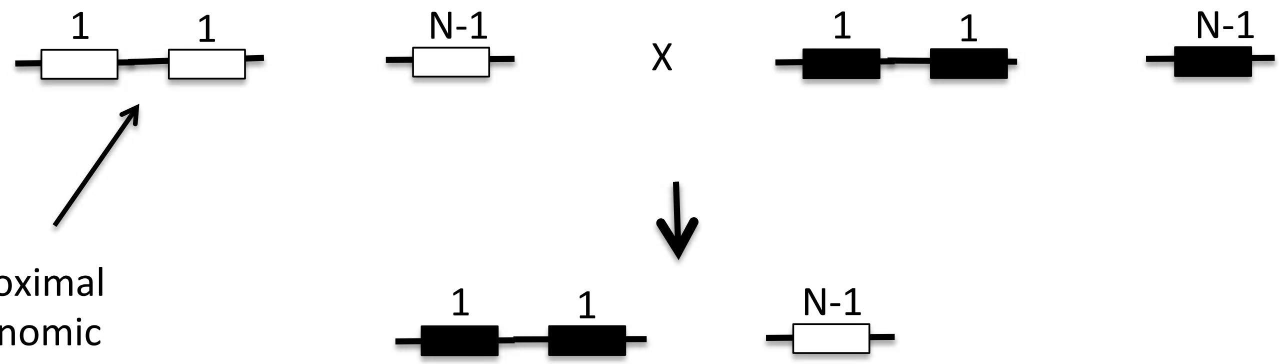
$$E[x] = .5$$

$$\sigma_x^2 = .25/N$$



Situation is more complex if loci are linked

Genetic linkage causes marker correlation



Proximal genomic locations makes crossing over unlikely during meiosis

- Assumption that each allele is inherited independently no longer holds – models more complex than binomial needed to capture this dependence

Genotype – Phenotype interactions

- i – individual in [1 .. N]
- g_i – genotype of individual i
- p_i – quantitative phenotype of individual i (single trait)
- e_i – environmental contribution to p_i

$$p_i = f\left(g_i\right) + e_i$$

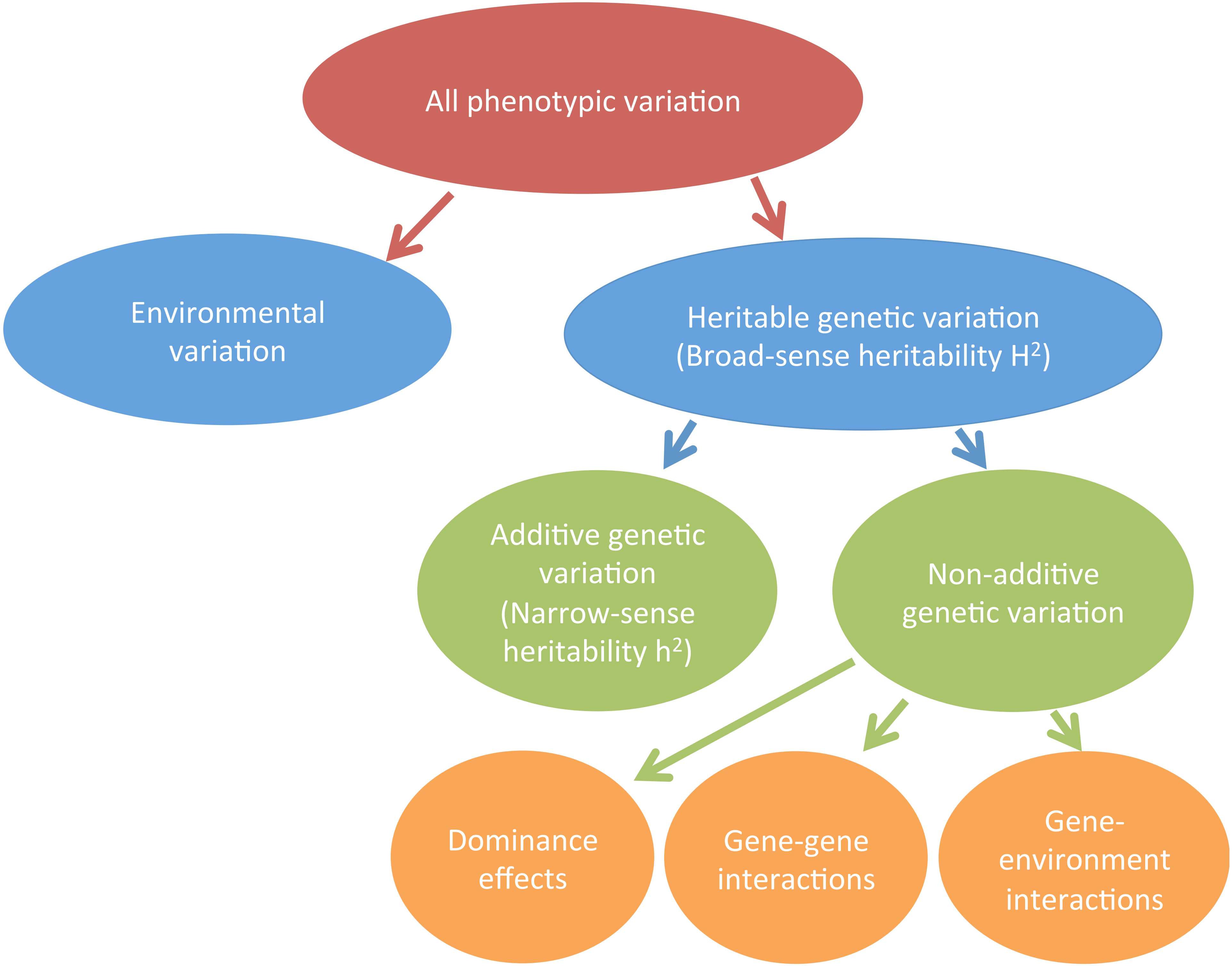
Phenotype is a function of genotype plus an environmental component

$$E\left[e_i\right] = 0 \qquad E\left[e^2\right] = \sigma_e^2$$

Environmental component is unbiased but introduces noise from genotype to phenotype

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + 2\sigma_{ge}^2 \quad \Rightarrow \quad \sigma_p^2 = \sigma_g^2 + \sigma_e^2$$

Assume environment affects all genotypes equally -> g and e are independent and their covariance is 0


All phenotypic variation
Environmental variation
Heritable genetic variation (Broad-sense heritability H2)
Additive genetic variation (Narrow-sense heritability h2)
Non-additive genetic variation
Dominance effects
Gene-gene interactions
Gene-environment interactions

2 types of heritability

- Broad-sense (H^2) and narrow-sense (h^2)
- Broad-sense
  - Fraction of phenotypic variance explained by genetic components

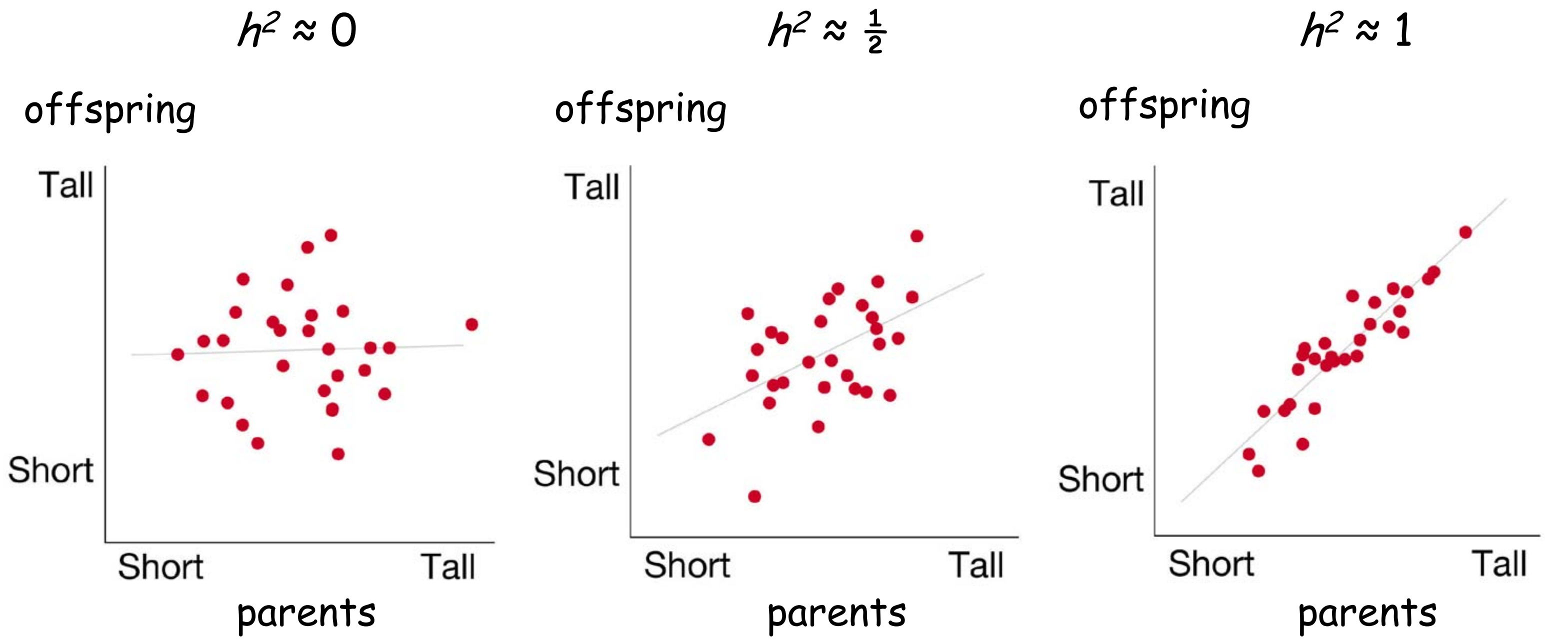
$$H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_p^2 - \sigma_e^2}{\sigma_p^2}$$

Can be estimated from identical twins or clones

Can be observed from all individuals in population

  - The upper bound for phenotypic prediction by optimal arbitrary (not necessarily linear) model
- Narrow-sense
  - The upper bound for phenotypic prediction by *linear* model (= fraction of total phenotypic variance that is caused by the additive effects of genes)
  - Determines the resemblance of offspring to their parents and the population's evolutionary response to selection

Narrow-sense heritability (h^2) is the regression (slope) of offspring on parents



- Regression slope is: Cov(x,y)/Variance(x) or Cov(parents, offspring)/Variance(parents)
  - x is the "mid-parent"

- The higher the slope, the better the offspring resemble their parents.
- In other words, the higher the heritability, the better the offspring trait values are predicted by parental trait values.

http://content.csbs.utah.edu/~rogers/ant5221/lecture/QTs2.pdf

Narrow-sense heritability: additive model of phenotype

- $g_{i,j}$ is a binary {0,1} variable of QTL j in individual i
- Each QTL in the genotype contributes independently & linearly to the phenotype:

$$f_a(g_i) = \sum_{j \in QTL} \beta_j g_{ij} + \beta_0$$

- β_j is the effect of QTL j on the phenotype (higher -> QTL has greater impact)
- For additive markers, children are expected to be the midpoint of their parents since they get an average of ½ loci from each parent:

$$E[f_a(g_i)] = \frac{f_a(p_1)}{2} + \frac{f_a(p_2)}{2}$$

Narrow-sense heritability: additive model of phenotype

$$f_a(g_i) = \sum_{j \in QTL} \beta_j g_{ij} + \beta_0$$

$$p_i = f_a(g_i) + e_i$$

$$\sigma_a^2 = \sigma_p^2 - \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(p_i - f_a(g_i)\right)^2}_{\text{Variance that remains after linear model – one source of “missing” heritability in studies}}$$

Additive genetic variance

Total phenotypic variance

Narrow-sense heritability: $h^2 = \dfrac{\sigma_a^2}{\sigma_p^2}$

Using LOD scores to discover QTLs for a trait (e.g. gene expression)

$$LOD = \log_{10}\prod_{i=1}^{N}\frac{P\left(p_i \mid g_{ij},\mu_0,\mu_1,\sigma\right)}{P\left(p_i \mid \mu,\sigma\right)}$$

LOD = Logarithm of the ODds
i = individual

"Null" model: locus does not affect gene's expression, and the probability of expression value p_i simply follows a Normal(μ,σ^2) distribution

"Alternative" model: locus affects a gene's expression (is a QTL), and there are different mean expression values μ_0 and μ_1 depending on which genotype is present at the locus (if g_{ij}=0 or 1)

- If the alternative model (that the locus is a QTL for the gene) doesn't explain the expression values any better than the null model, the probability ratios are 1 and the LOD score is 0
  - If alternative model better explains the data, LOD score > 0
- If the locus is a QTL, the LOD score will get higher with increasing number of individuals (*N*) – with larger sample samples we have greater power to detect loci as being statistically significant QTLs. This is referred to as "power" – a study with too few people to determine statistical significance at some loci is "underpowered".

Using LOD scores to discover QTLs for a trait (e.g. gene expression)

$$LOD = \log_{10} \prod_{i=1}^{N} \frac{P\left(p_i \mid g_{ij}, \mu_0, \mu_1, \sigma\right)}{P\left(p_i \mid \mu, \sigma\right)}$$

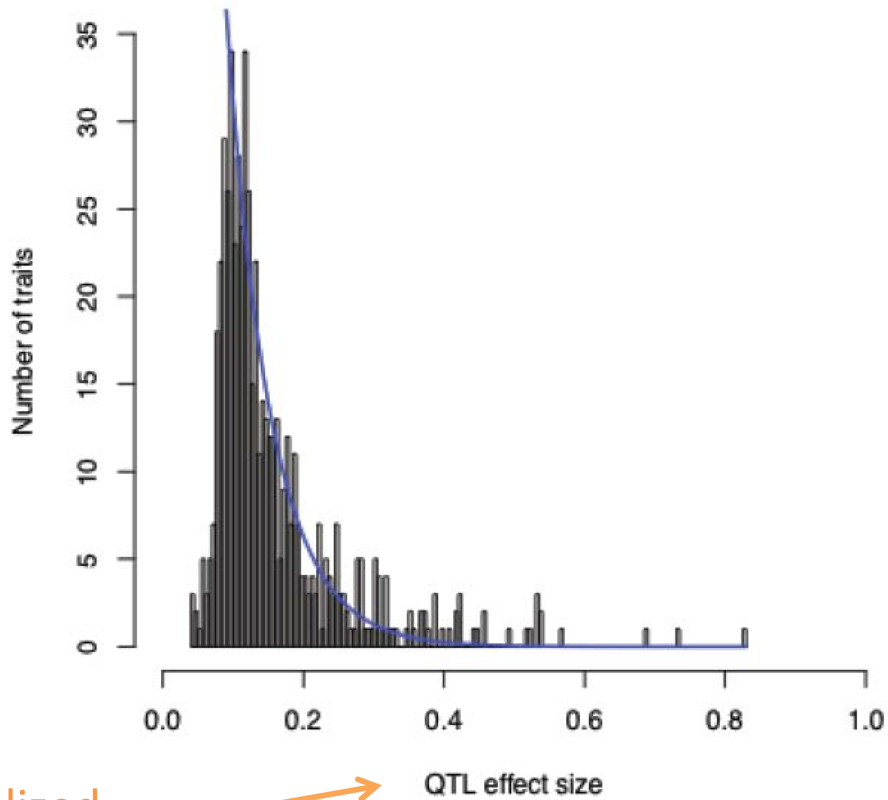
- How to determine if a LOD score is significant?
  - Permute genotypes (so the marker g_{ij} and expression values are mixed up) 1000 times and compute LOD scores to get empirical null distribution
  - Determine the null LOD score that corresponds to FDR = 0.05
  - Use this threshold on unpermuted LOD scores to find QTLs for each gene
  - Since all loci are included in the permuted null distribution, no multiple hypothesis correction needed

- Fit a linear model to discovered QTLs to determine each QTL's contribution (β_j)

- Once this has been done to find the set of statistically significant QTLs from the first pass, you can repeat to find QTLs in the residuals from the existing model that may have been below the threshold in the first pass (3 times)

Bloom et al. 2013: "Finding the sources of missing heritability in a yeast cross"

- 5-29 QTLs per trait (median of 12), although most QTLs have small effect size

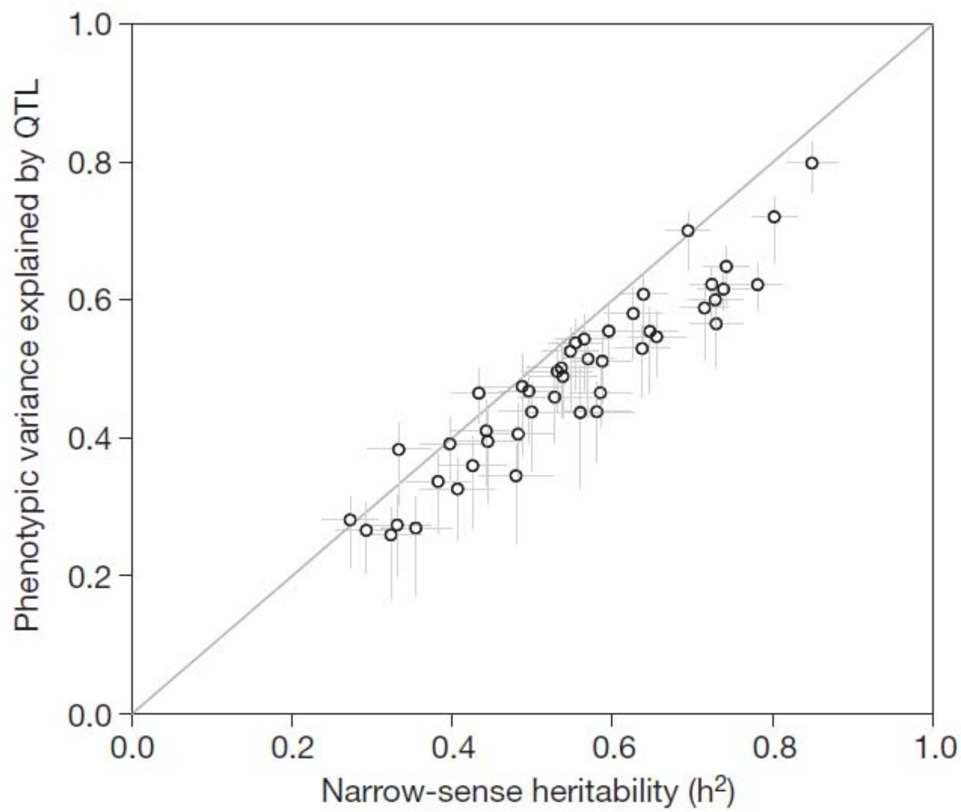


Absolute value of normalized difference in means between genotypes

Source: Bloom, Joshua S., Ian M. Ehrenreich, et al. "Finding the Sources of Missing Heritability in a Yeast Cross." *Nature* 494, no. 7436 (2013): 234-7.

Bloom et al. 2013: "Finding the sources of missing heritability in a yeast cross"

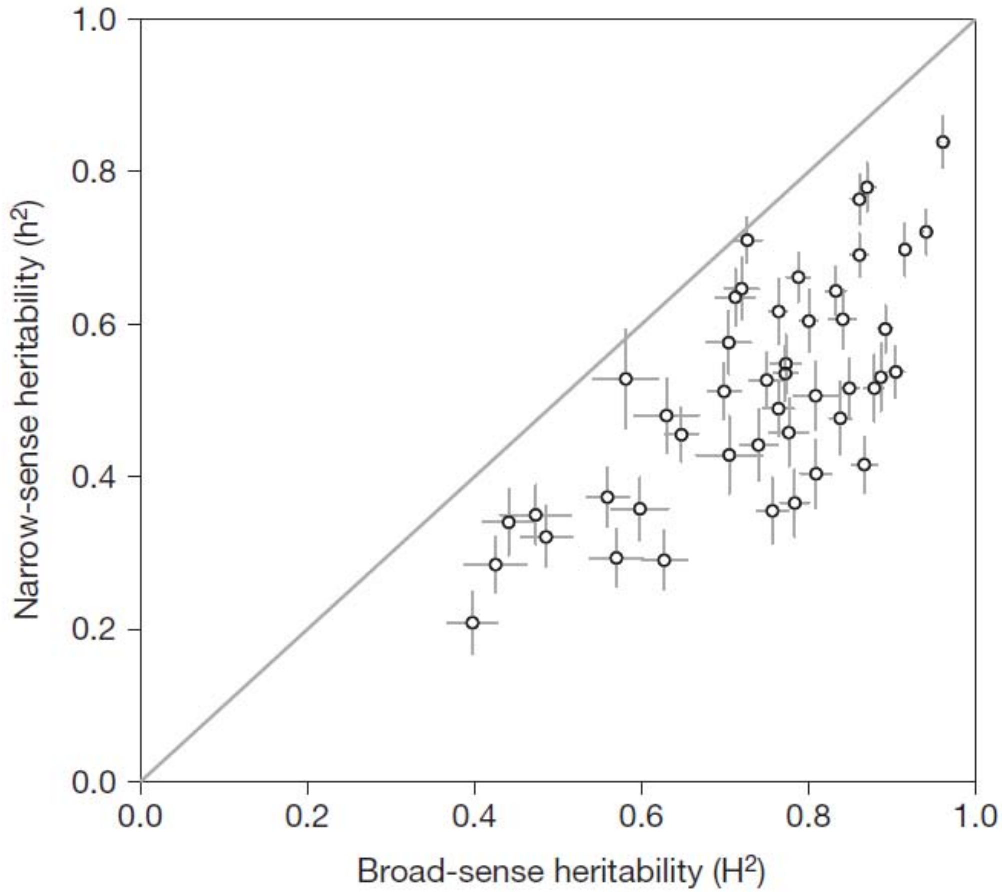
- Good news: most additive heritability (narrow-sense) is explained by detected QTLs



Source: Bloom, Joshua S., Ian M. Ehrenreich, et al. "Finding the Sources of Missing Heritability in a Yeast Cross." *Nature* 494, no. 7436 (2013): 234-7.

Bloom et al. 2013: "Finding the sources of missing heritability in a yeast cross"

- Bad news: There is still much heritability missing from our additive linear model



Source: Bloom, Joshua S., Ian M. Ehrenreich, et al. "Finding the Sources of Missing Heritability in a Yeast Cross." *Nature* 494, no. 7436 (2013): 234-7.

Bloom et al. 2013: "Finding the sources of missing heritability in a yeast cross"

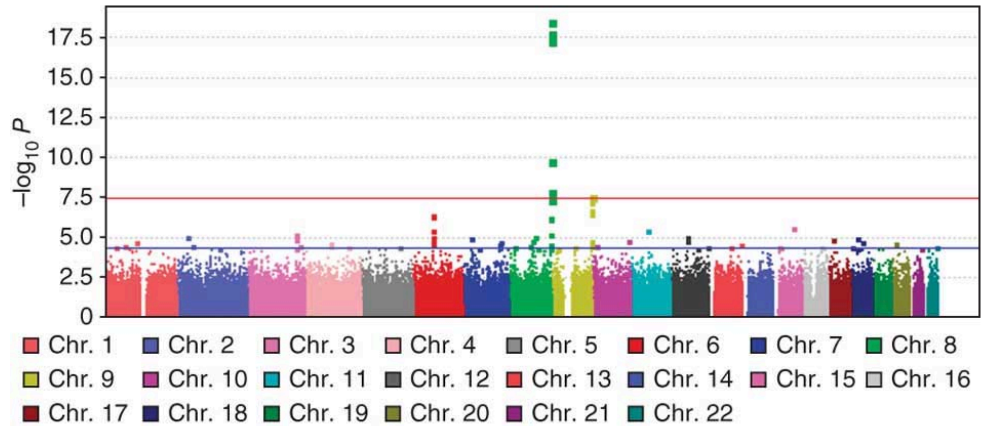
- What could cause the missing heritability?
  - Incorrect heritability estimates
  - Rare variants that the study is underpowered to detect
  - Structural variants (insertions or deletions – these studies typically only measure SNPs)
  - Epigenetic interactions
  - Epistatic effects
    - When the effect of a gene depends on the presence of one or more modifier genes (the genetic background)
    - Example: locus A and locus B each only cause a 5% decrease if one of the variants is present, but a 50% decrease if both are present
    - Since all pairwise interactions is too large of a search space (100,000 x 100,000), can only consider all interactions that involve at least of the detected QTLs (20 x 100,000)

Human Genetics

- We want to find human variants (SNPs, etc.) that are associated with a particular phenotype (e.g. a disease)

"Manhattan plot"

http://www.nature.com/ng/journal/v44/n4/images/ng.1109-F1.jpg



Source: Tanikawa, Chizu, Yuji Urabe, et al. "A Genome-wide Association Study Identifies Two Susceptibility Loci for Duodenal Ulcer in the Japanese Population." *Nature Genetics* 44, no. 4 (2012): 430-4.

- We need a way to test whether a SNP is significantly associated with a phenotype:
  - Chi-squared test
    - Asymptotic approximation, so not appropriate if counts are small (should be at least 5 counts per category)
  - Fisher's exact test
    - An "exact" calculation (not asymptotic approximation), but involved factorials so computationally difficult when counts become large (but this is exactly when the Chi-square test is appropriate)

Testing for SNP/phenotype association

- Testing for association between a SNP and a disease (or some other trait) – we are given the following counts:

Allele	Cases	Controls	Total Counts
C	62	80	142
A	108	250	358
Total Counts	170	330	500

- Calculate expected counts under null hypothesis that the proportion/ ratio of cases to controls is the same regardless of whether an individual is C or A:
  - 1) calculate total proportion of **cases** regardless of A/C = 170/500 = 0.34
  - 2) calculate what proportion of the 142 Cs should be cases according to the total proportion of cases = 142(0.34) = 48.28, controls = 142(1-0.34) = 93.72
  - 3) same for the As: what proportion of the 358 As should be cases/controls according to null model?

for A individuals, expected cases = 358(0.34) = 121.72, controls = 358(1-0.34)=236.28

Testing for SNP/phenotype association

- Testing for association between a SNP and a disease (or some other trait) – we are given the following counts:

Observed

Allele	Cases	Controls	Total Counts
C	62	80	142
A	108	250	358
Total Counts	170	330	500

Expected

Allele	Cases	Controls	Total Counts
C	48.28	93.72	142
A	121.72	236.28	358
Total Counts	170	330	500

Using a Chi-squared test:

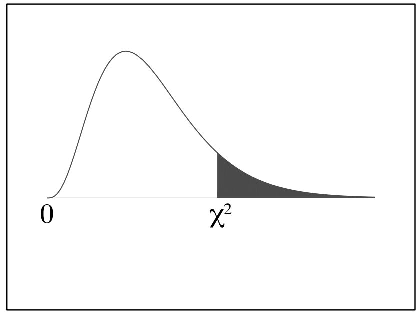
$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \frac{(62-48.28)^2}{48.28} + \frac{(80-93.72)^2}{93.72} + \frac{(108-121.72)^2}{121.72} + \frac{(250-236.28)^2}{236.28} = 8.25$$

df = (# rows -1)(# cols – 1) = 1

Testing for SNP/phenotype association

http://sites.stat.psu.edu/~mga/401/tables/Chi-square-table.pdf

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

Since our statistic (8.25) is higher than the cut-off for P = 0.005, the P-value is less than 0.005

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	$\boxed{7.879}$
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Using a Chi-squared test:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \frac{(62-48.28)^2}{48.28} + \frac{(80-93.72)^2}{93.72} + \frac{(108-121.72)^2}{121.72} + \frac{(250-236.28)^2}{236.28} = 8.25$$

df = (# rows -1)(# cols – 1) = 1, and $P(X_1^2 \geq 8.25) = 0.0041$ so we reject H_0 (=>SNP is associated)

Testing for SNP/phenotype association

- Testing for association between a SNP and a disease (or some other trait) – we are given the following counts:

Observed

Allele	Cases	Controls	Total Counts
C	62	80	142
A	108	250	358
Total Counts	170	330	500

Fisher's Exact Test:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

Sum all probabilities for observed and all more extreme values with same marginal totals to compute probability of null hypothesis

Let our 1 degree of freedom be a, the number of cases with "C"

Upper-tail one-sided P-value: $\sum_{a=62}^{142} \frac{\binom{142}{a}\binom{358}{170-a}}{\binom{500}{170}} \approx .003$

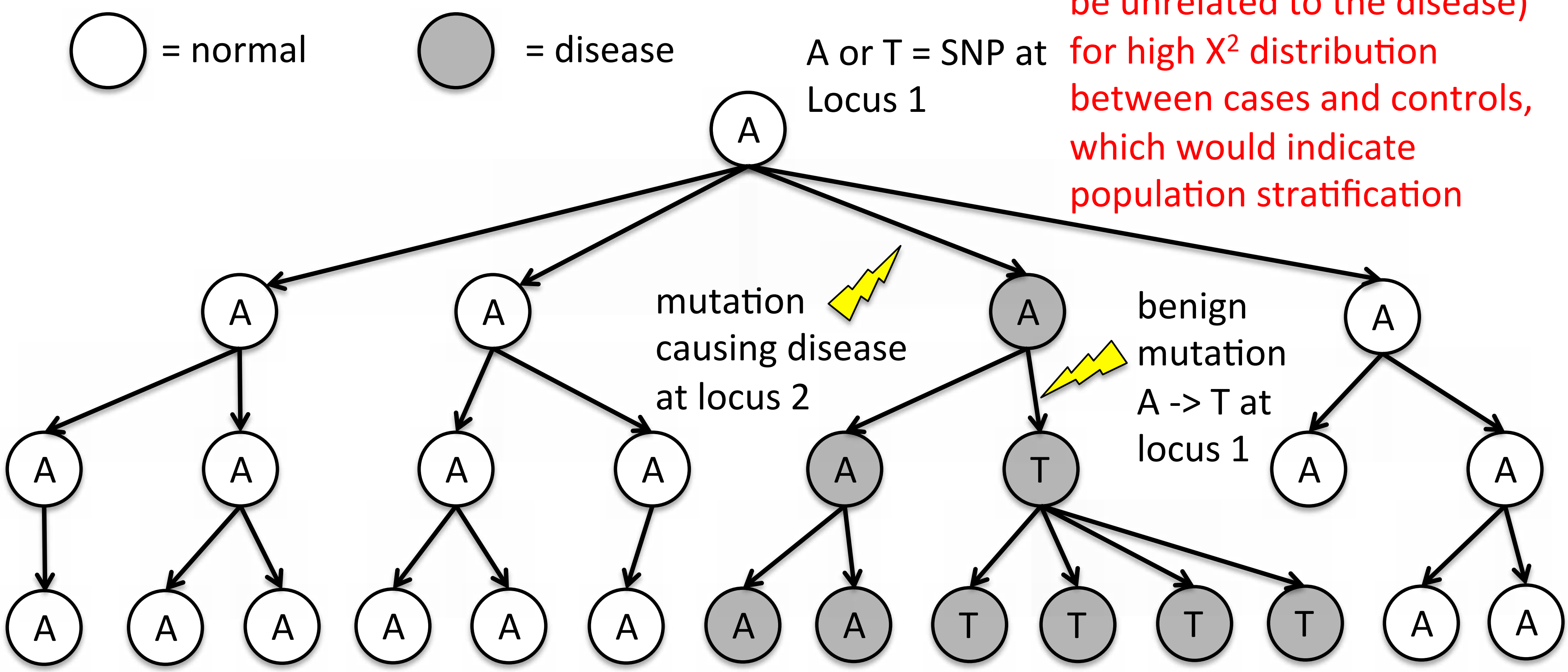
Since the expected count of a (= Cases with C) was ~48, since 62 = 48 + 14, the lower tail goes up to 48 - 14 = 34. The two-sided P-value is:

$$\sum_{a=0}^{34} \frac{\binom{142}{a}\binom{358}{170-a}}{\binom{500}{170}} + \sum_{a=62}^{142} \frac{\binom{142}{a}\binom{358}{170-a}}{\binom{500}{170}} \approx .0047$$

Human Genetics

After doing a Chi-square test and seeing that a SNP is significantly enriched in a disease population, we might believe that the SNP is linked to the disease. But <u>population structure</u> can confound these results (methods for correcting for this are beyond the scope of this class)

Test control SNPs (known to be unrelated to the disease) for high X^2 distribution between cases and controls, which would indicate population stratification

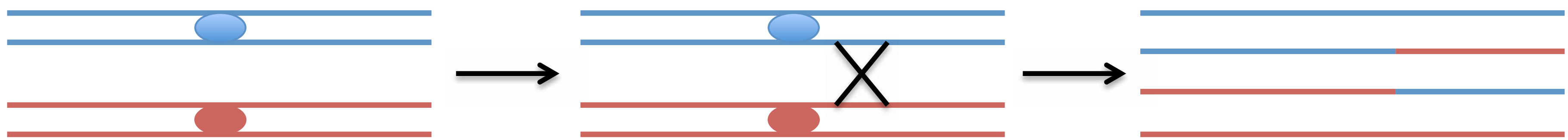


In 4th generation, fraction of Ts in population = 4/14, but in diseased group = 4/6

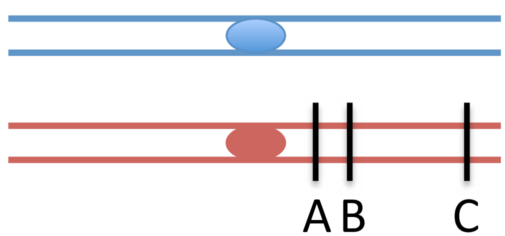
But once we see the family tree, we see that the SNP at locus 1 is unrelated to the disease

Linkage Disequilibrium

- Recombination during meiosis "shuffles" alleles between the homologous maternal and paternal chromosomes



Over time and after many crossover events have occurred, loci that are physically close together on the chromosome will tend to remain together, so the probability of two loci occurring together is a function of their distance along the chromosome



If a crossover event is equally likely to occur at any position along the chromosome, the probability that it will separate loci A and B is much smaller than A and C or B and C

We have so far generally assumed that inheriting a particular allele at one locus won't affect the probability of inheriting an allele at a different locus. Such loci are in <u>linkage equilibrium</u>.

Loci are considered in <u>linkage disequilibrium</u> if genotypes at two loci are not independent of one another (e.g. inheriting A at locus 1 influences probability of inheriting B at locus 2)

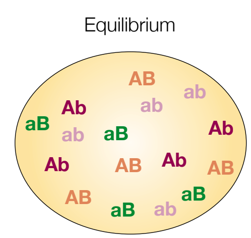
Linkage Disequilibrium

- Measuring linkage disequilibrium: consider two loci A and B, where locus A has two possible alleles A and a, and locus B has two alleles B and b:
  - then gametes can have one of four possible combinations:

Gamete	Frequency
AB	p_{AB}
Ab	p_{Ab}
aB	p_{aB}
ab	p_{ab}

Allele	Frequency
A	$p_A = p_{AB+}p_{Ab}$
a	$p_a = p_{aB+}p_{ab}$
B	$p_B = p_{aB+}p_{AB}$
b	$p_b = p_{ab+}p_{Ab}$

- Then if alleles are randomly associated w/ one another, the frequencies of the four gametes should be the product of the allele frequencies:
  - ex. $p_{AB} = p_A p_B = (p_{AB}+p_{Ab})(p_{aB}+p_{AB})$



http://www.nature.com/nrg/journal/v2/n1/pdf/nrg0101_011a.pdf

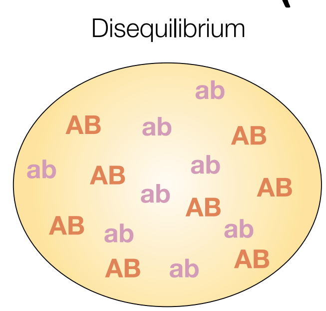
Source: Mackay, Trudy FC. "Quantitative Trait Loci in Drosophila." *Nature Reviews Genetics* 2, no. 1 (2001): 11-20.

Linkage Disequilibrium

Gamete	Frequency
AB	p_{AB}
Ab	p_{Ab}
aB	p_{aB}
ab	p_{ab}

Allele	Frequency
A	$p_A=p_{AB}+p_{Ab}$
a	$p_a=p_{aB}+p_{ab}$
B	$p_B=p_{aB}+p_{AB}$
b	$p_b=p_{ab}+p_{Ab}$

- If they are <u>not</u> randomly associated (and therefore in linkage disequilibrium) then there will be a deviation (D) in the expected frequencies:
  - $p_{AB} = p_A p_B + D$
  - $p_{Ab} = p_A p_b - D$
  - $p_{aB} = p_a p_B - D$
  - $p_{ab} = p_a p_b + D$



http://www.nature.com/nrg/journal/v2/n1/pdf/nrg0101_011a.pdf

Source: Mackay, Trudy FC. "Quantitative Trait Loci in Drosophila." *Nature Reviews Genetics* 2, no. 1 (2001): 11-20.

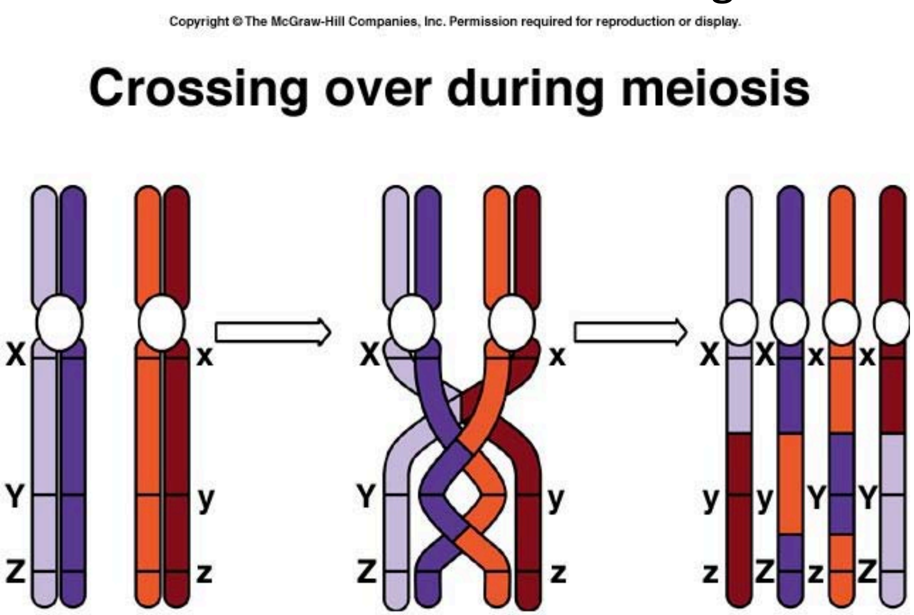
- Where D is given by:
  - $D = p_{AB}p_{ab} – p_{Ab}p_{aB}$ ($D = 0$ => no disequilibrium)
- AB and ab are the "coupling" gametes (AB on one parental chromosome, ab on the other), Ab and aB are the "repulsion" gametes (crossing over event must occur between the loci) – D is the difference between these types.

Variant Phasing

- To determine which genes are linked together (and therefore likely to be inherited together in the next generation), you need to figure out which alleles (which variant SNPs) are on the same chromosome = "phasing"
  - Why does this matter?

    - If you have 2 different mutations in the same copy of a gene (phased), the 2nd copy (no mutations) may be enough for normal activity

    - If there's one mutation in each (unphased), both copies of the gene may be nonfunctional
- Often rely on family data (e.g. parents) to determine which "parental" chromosome segments were inherited together in the child
- Can be used to identify haplotypes = combinations of alleles at adjacent locations in a chromosome that are inherited together over many generations



http://www.uic.edu/classes/bios/bios100/lecturesf04am/crossingover01.jpg

X, Y, and Z are "in phase" on this chromosome

x, y, and z are "in phase" on this chromosome

Due to crossing over, the phasing has changed

Variant Phasing

Longer reads will help – two SNPs present in the same read are definitely on the same chromosome

Hardy-Weinberg Equilibrium (HWE)

- Assume only two alleles: A and a
- If P(A) = ψ = frequency of A in the population, and the population is in HWE, then:
  - P(AA) = ψ^2
  - P(Aa) = $2\psi(1-\psi)$
  - P(aa) = $(1-\psi)^2$

gamete	**A** (ψ)	**a** ($1-\psi$)
A (ψ)	**AA** (ψ^2)	**Aa** ($\psi(1-\psi)$)
a ($1-\psi$)	**Aa** ($\psi(1-\psi)$)	**aa** ($(1-\psi)^2$)

- HWE states that allele and genotype frequencies in a population will be constant from generation to generation in the absence of other evolutionary forces; assuming the following:
  - random mating
  - population size is infinite
  - no migration, mutation or selection (so allele frequencies won't change)

HWE and Likelihood ratio tests

- Testing whether a population is in HWE using a likelihood ratio test (LRT):
  - say we observe N = 200 individuals with the following genotypes: **25** aa, **90** Aa, **85** AA
  - is this population in HWE?
- Recall that the likelihood ratio is given by:

$$\lambda = \frac{P(Data \mid H_0)}{P(Data \mid H_1)}$$

likelihood of the data under the null model

likelihood of the data under the alternative model

- Then the following test statistic is approximately Chi-square distributed:

$$-2\ln(\lambda) \sim X^2_{df}$$

  - *df* = (# free parameters in H_1) – (# free parameters in H_0)

HWE and Likelihood ratio tests

- We observe n = 200 individuals with the following genotypes: **25** aa, **55** Aa, **120** AA
  - is this population in HWE?
- Here, under the unconstrained model H_1, the parameters are p_{AA}, p_{Aa} and p_{aa} (df = 2)
  - for this example: $p_{AA} = 120/200 = 0.6$, $p_{Aa} = 55/200 = 0.275$, $p_{aa} = 25/200 = 0.125$
- Under the constrained model H_0, we only need p_A (fraction of A alleles in population) and if HWE holds:
  - $p_A = (2n_{AA}+n_{Aa})/2n = (2(120) + 55)/400 = 295/400 = 0.7375$
  - $p_{AA} = (p_A)^2 = (0.7375)^2 = 0.5439$
  - $p_{Aa} = 2p_A(1-p_A) = 0.3872$
  - $p_{aa} = (1-p_A)^2 = 0.0689$

could also do a Chi-square goodness of fit test with these probabilities * n as the expected counts instead of LRT

HWE and Likelihood ratio tests

- We observe N = 200 individuals with the following genotypes: **25** aa, **90** Aa, **85** AA
  - is this population in HWE?
- Therefore, our test statistic is:

$$-2\ln(\lambda) = -2\ln\frac{P(Data \mid p_A^2, 2p_A(1-p_A), (1-p_A)^2)}{P(Data \mid p_{AA}, p_{Aa}, p_{aa})}$$

$$= -2\ln\frac{P(Data \mid 0.5439, 0.3872, 0.0689)}{P(Data \mid 0.6, 0.275, 0.125)}$$

- Note that P(Data|H) follows a multinomial distribution (generalized binomial for more than 2 categories):

$$P(x_1,\ldots,x_k; n, p_1,\ldots,p_k) = \frac{n!}{x_1!,\ldots,x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

Note that the factorials will drop out of LRT

So for example:

$$P(Data \mid H_1) = P(25, 90, 85; 200, 0.6, 0.275, 0.125) = \frac{200!}{25!90!85!} 0.6^{25} 0.275^{90} 0.125^{85}$$

Likelihood Ratio Tests

- Can use a similar LRT to determine whether the data are better explained when treated as two subpopulations, like cases and controls:
  - H_0: p_{AA}, p_{Aa} and p_{aa} are sufficient to explain the data
  - H_1: we do better by considering two subpopulations:
    - p^1_{AA}, p^1_{Aa} and p^1_{aa} for subpopulation 1 (D^1)
    - p^2_{AA}, p^2_{Aa} and p^2_{aa} for subpopulation 2 (D^2)
- Then our test statistic T is:

$$T = -2\ln \frac{P(D \mid p_{AA}, p_{Aa}, p_{aa})}{P(D^1 \mid p^1_{AA}, p^1_{Aa}, p^1_{aa}) P(D^2 \mid p^2_{AA}, p^2_{Aa}, p^2_{aa})}$$

  - approx. Chi-square distributed with $df = 4 - 2 = 2$

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014