SISG 2014

SISG Module 20: Plant and Animal Association Mapping

19th Summer Institute in Statistical Genetics

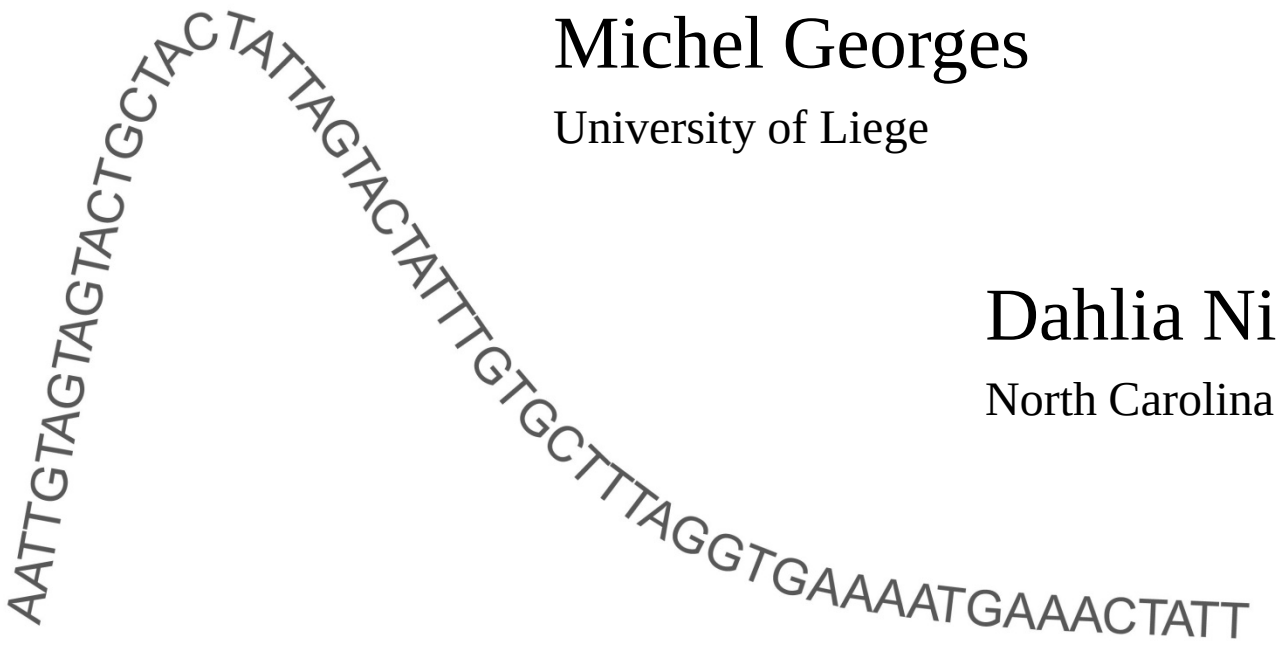
UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Plant and Animal Association Mapping

Michel Georges
University of Liege

Dahlia Nielsen
North Carolina State University



Introduction

Basics of mapping

Linkage disequilibrium

Simple tests of association

Gene mapping

- Identify regions of the genome that contain genes affecting the trait of interest.
- Best case scenario:

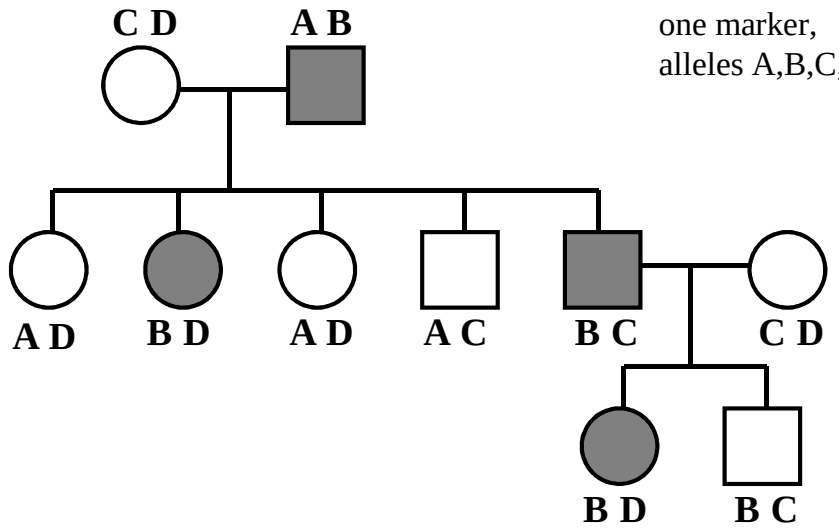
  Find **the** gene
- More likely with current technology:

  Find a region with 10s – 1000s of genes.

Linkage mapping

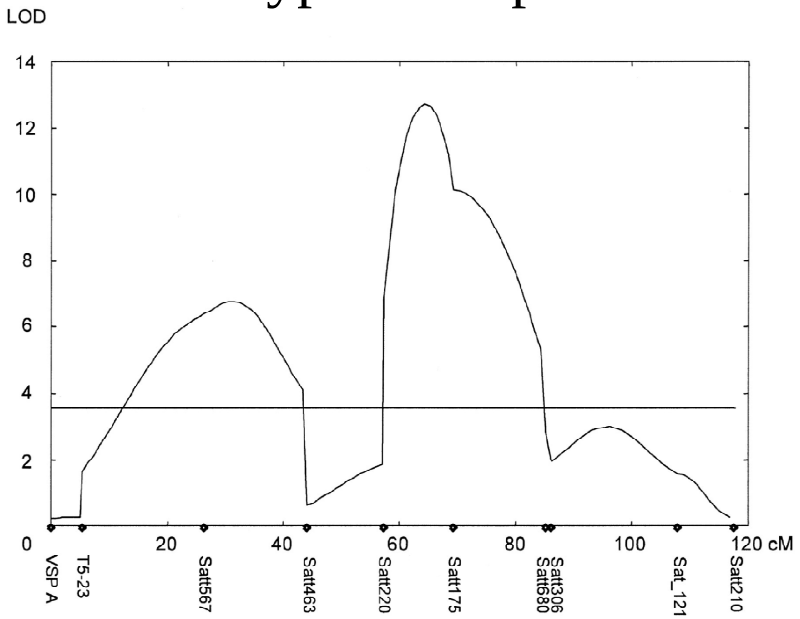
- Use family data to identify genomic regions.
- Utilizes recombination rates as measures of distances between loci.
- On average, the closer together two loci are on a chromosome, the smaller the recombination rate between them will be.
- With family data, can estimate recombination rates directly, and thus determine which markers are closest to the genes of interest.

Family data



- From first to second generation, observe zero recombination events (out of five possible events).
- From second generation to third generation, observe one recombination event (out of two possible).
- Estimate of recombination rate is 1/7.
- Can do this for many markers and find the ones with the smallest recombination rates
  - these markers should be the closest to the gene.

Typical output



https://www.crops.org/publications/cs/articles/45/5/2044

Linkage analysis

- Many successes.
- Identifies large genomic regions
  - 100s – 1000s of genes
  - resolution is an issue.

Association Mapping

- Uses unrelated individuals from a population.
- Capitalizes on historical recombination events.
- Has the ability to provide much more fine-scale resolution than linkage analysis.
  - much smaller regions
- Does not estimate genetic distances directly:
  - a positive result is just an indicator (not a measure of distance).

Association Mapping

- Uses unrelated individuals from a population.
- Capitalizes on historical recombination events.

How?

Disequilibrium

Hardy Weinberg Disequilibrium

- Describes genotype frequency properties:
- Pr (AA) = p_A^2
- Pr (AT) = 2 p_A p_T
- Pr (TT) = p_T^2

  (independence of alleles in a genotype)

- HW disequilibrium occurs when genotype frequencies deviate from these values.

Linkage Disequilibrium (LD)

- Alleles at one locus are correlated with alleles at a second locus on a population level.

A	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**a**	**a**	**a**	**a**
B	**B**	**B**	**B**	**B**	**B**	**b**	**b**	**b**	**b**	**b**	**b**

If you sample a haplotype at random from the population, does knowing the allele at the first locus of this haplotype give you information about the allele at the second locus?

Linkage Disequilibrium

§ One usual measure of LD is:

$$D_{AB} = P_{AB} - p_A p_B$$

A	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**a**	**a**	**a**	**a**
B	**B**	**B**	**B**	**B**	**B**	**b**	**b**	**b**	**b**	**b**	**b**

§ $P_A = 8/12$
§ $P_B = 6/12$
§ $P_{AB} = 6/12$
§ $D_{AB} = 1/6$

Linkage Disequilibrium

§ One usual measure of LD is:

$$D_{AB} = P_{AB} - p_A p_B$$

A	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**a**	**a**	**a**	**a**
B	**B**	**B**	**B**	**B**	**B**	**b**	**b**	**b**	**b**	**b**	**b**

§ **LD is a measure of extant haplotypes**
- Estimates do not rely on or measure inheritance
- LD does not measure how often alleles are *transmitted* together

LD can be created by

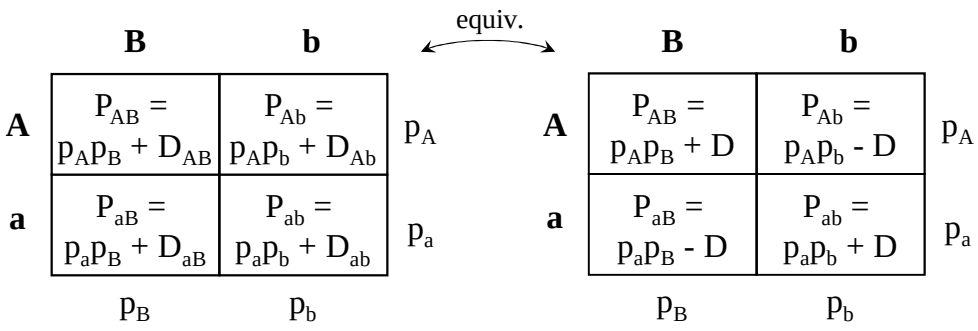
§ Mixing of populations.
§ Population substructure (non-random mating within populations).
§ Mutations creating new haplotypes.
§ Selection favoring certain alleles.
§ Founder effects.
§ Genetic drift.

Properties of LD

§ One usual measure of LD is:

$$D_{AB} = P_{AB} - p_A p_B$$

§ $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$



Bounds on LD

- $0 \le p_A p_B + D_{AB} \le \min(p_A, p_B)$
- $0 \le p_A p_b - D_{AB} \le \min(p_A, p_b)$
- $0 \le p_a p_B - D_{AB} \le \min(p_a, p_B)$
- $0 \le p_a p_b + D_{AB} \le \min(p_a, p_b)$

- $\max(-p_A p_B, -p_a p_b) \le D_{AB} \le \min(p_A p_b, p_a p_B)$

Other Measures of LD

- Can divide D_{AB} by the maximum value it can obtain:

$$D'_{AB} = \begin{cases} D_{AB} / [\max(-p_A p_B, -p_a p_b)] & \text{if } D_{AB} < 0 \\ D_{AB} / [\min(p_A p_b, p_a p_B)] & \text{if } D_{AB} > 0 \end{cases}$$

- The sampling properties of D'_{AB} are not well understood.

- $r^2_{AB} = \dfrac{D^2_{AB}}{p_A p_B p_a p_b}$

- $n \tilde{r}^2_{AB} \sim \chi^2_{(1)}$

Estimating LD

- $D_{AB} = P_{AB} - p_A p_B$

- $\tilde{D}_{AB} = \tilde{P}_{AB} - \tilde{p}_A \tilde{p}_B$

- $\tilde{P}_{AB} = n_{AB} / 2n$

- $\tilde{p}_A = n_A / 2n, \quad \tilde{p}_B = n_B / 2n$

Without Haplotype Data

- $\tilde{P}_{AB} = n_{AB} / 2n$

- Usually have genotype data, but not haplotype data

- AaBb $\quad \left|\begin{matrix}\mathbf{A} & \mathbf{a} \\ \mathbf{B} & \mathbf{b}\end{matrix}\right| \quad \mathbf{?} \quad \left|\begin{matrix}\mathbf{A} & \mathbf{a} \\ \mathbf{b} & \mathbf{B}\end{matrix}\right|$

Haplotyping

- Many tools exist
  - most assume HWE
- Examples (not a complete list):
  - Phase, fastPhase (stephenslab.uchicago.edu/software.html)
  - Beagle (faculty.washington.edu/browning/beagle/beagle.html)
  - HapSeq2 (www.ssg.uab.edu/hapseq)
  - MACH (www.sph.umich.edu/csg/abecasis/MACH)
  - Impute2 (mathgen.stats.ox.ac.uk/impute/impute_v2.html)

Composite Measures

- Can be used when haplotype data are not available.
- Does not rely on HWE.
- $\Delta_{AB} = D_{AB} + D_{A/B}$

Composite Measure

- Measures both D_{AB} and $D_{A/B}$:

Composite Measures

- $\Delta_{AB} = P_{AB} + P_{A/B} - 2\, p_A\, p_B$
- $\tilde{\Delta}_{AB} = (n_{AB} + n_{A/B}) / n \; - 2\, \tilde{p}_A\, \tilde{p}_B$
- $n_{AB} + n_{A/B} =$
  $2\, n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}\, n_{AaBb}$

Tests of disequilibrium: Z Test

Use of Fisher's approximate variance formula gives

$$\begin{aligned}\mathrm{Var}(\tilde{D}_{AB}) &= \frac{1}{n}[p_A(1-p_A)p_B(1-p_B) \\ &\quad + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2]\end{aligned}$$

For large sample sizes,

$$z = \frac{\tilde{D}_{AB} - \mathcal{E}(\tilde{D}_{AB})}{\sqrt{\mathrm{Var}(\tilde{D}_{AB})}} \sim N(0,1)$$

Therefore, under $H_0 : D_{AB} = 0$,

$$\begin{aligned}z &= \frac{\tilde{D}_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)/n}} \\ &\approx \frac{\tilde{D}_{AB}}{\sqrt{\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)/n}} \sim N(0,1) \\ z^2 &\sim \chi^2\end{aligned}$$

Tests of disequilibrium: composite disequilibrium

Substituting observed frequencies into the definition equations provides MLE's, and Fisher's formula gives approximate variances

$$\begin{aligned}n\mathrm{Var}(\hat{\Delta}_{AB}) &= (\pi_A + D_A)(\pi_B + D_B) + \frac{1}{2}\tau_A\tau_B\Delta_{AB} \\ &\quad + \tau_A D_{ABB} + \tau_B D_{AAB} + \Delta_{AABB}\end{aligned}$$

where n still refers to n individuals. If quadrigenic and trigenic coefficients can be ignored, a test statistic for composite digenic linkage disequilibrium is

$$X_{AB}^2 = \frac{n\hat{\Delta}_{AB}^2}{(\tilde{\pi}_A + \hat{D}_A)(\tilde{\pi}_B + \hat{D}_B)}$$

Note the explicit way in which departures from Hardy-Weinberg are included in this expression. Linkage disequilibrium can be tested for, even in non-random mating populations.

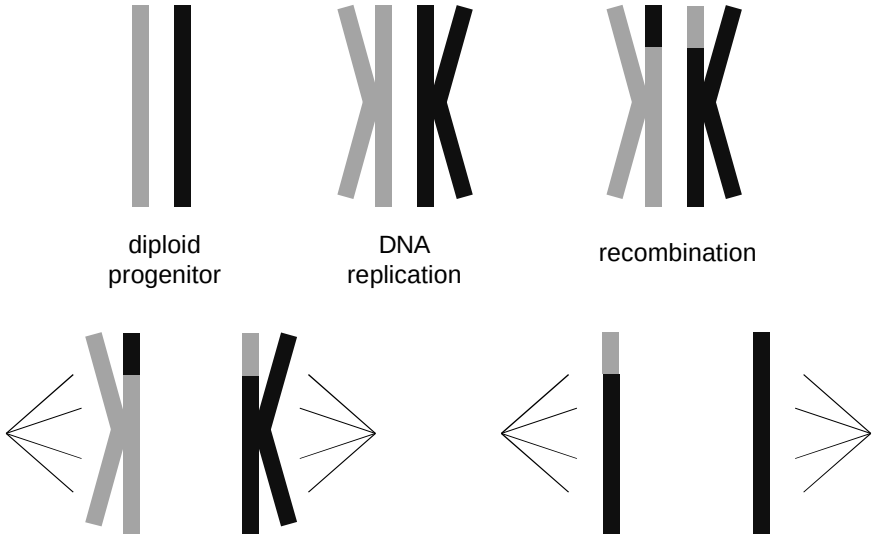
$\pi_A = p_A(1-p_A)$
$\tau_A = (1-2p_A)$

Using linkage disequilibrium to identify genes that may affect a trait …

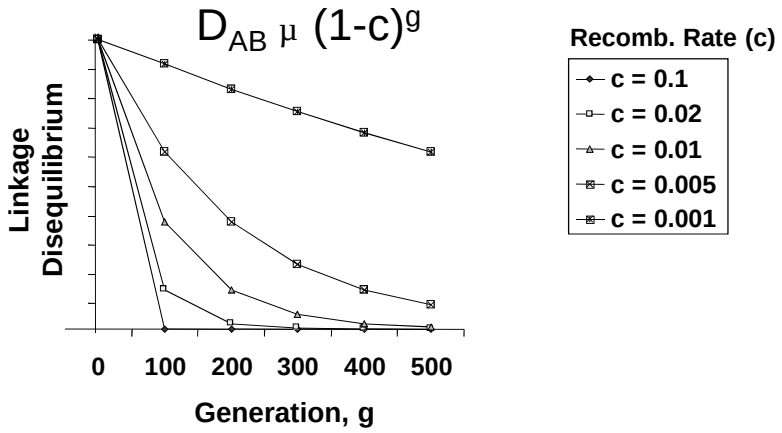
Predicted behavior of LD over time

- $D_{AB}(g) = (1 - c)^g \times D_{AB}(0)$
- Although this predicts the expected value of LD over time, there is a large variance around the mean.
- Populations with similar starting values can be quite different after time.

Reminder: meiosis & recombination (extremely simplified)

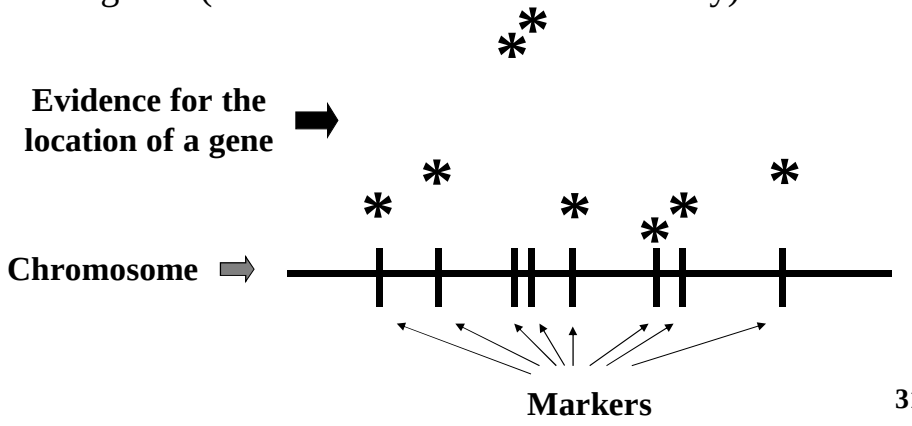


Linkage Disequilibrium versus Generations Since its Creation

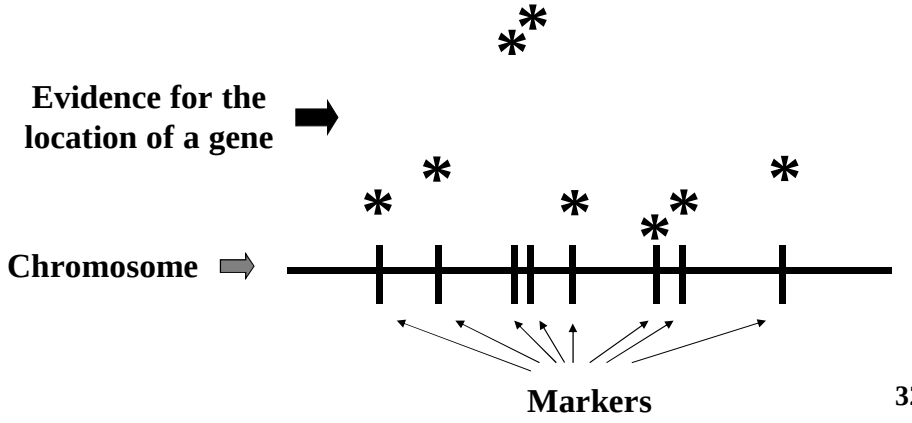


Locating genes of interest

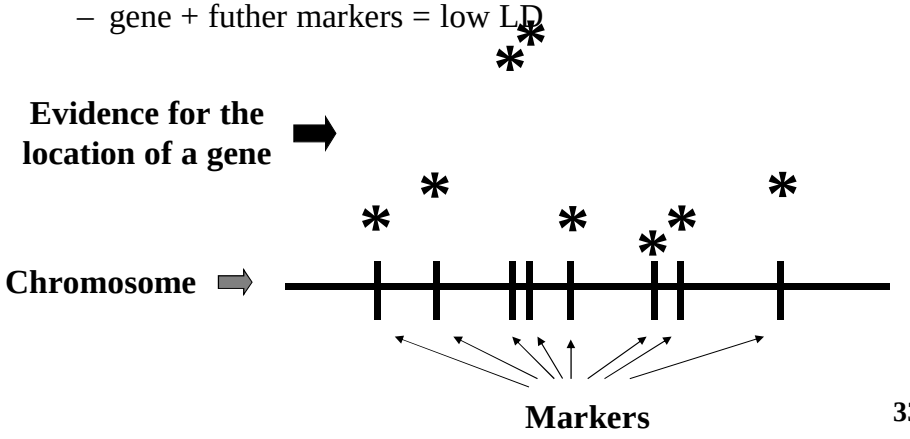
- We know where the markers are and we can get genotypes for them.
- This info can help us find the location of the genes (which can't be examined directly).



- We want to use *linkage disequilibrium* to provide evidence for the location of the unknown gene with respect to the markers.



§ If LD decays faster for larger recombination rates, it should be stronger for things that are close together and weaker for things further apart
  - gene + nearby markers = high LD
  - gene + futher markers = low LD



Consider Two Populations

§ Population 1 haplotypes:

A B	A B	A B	A B	A B	A B	A B

§ Population 2 haplotypes:

a b	a b	a b	a b	a b	a b	a b

If These Populations Mix

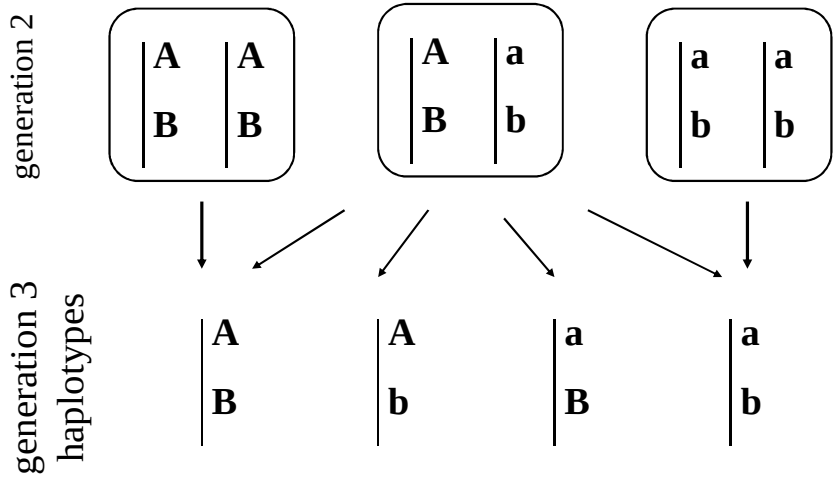
§ Individuals directly after mixing:

A B | A B	a b | a b	a b | a b

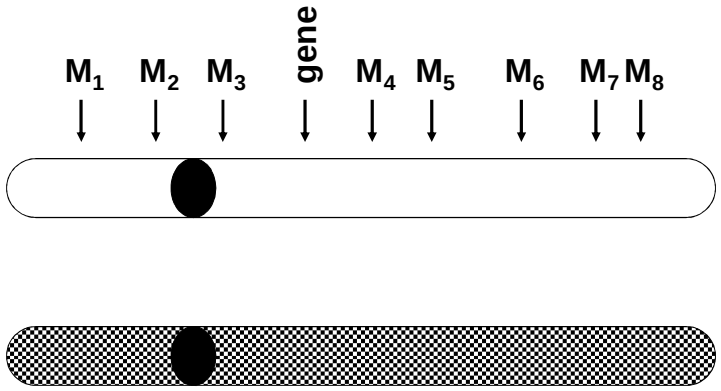
§ Their offspring:

A B | A B	A B | a b	a b | a b

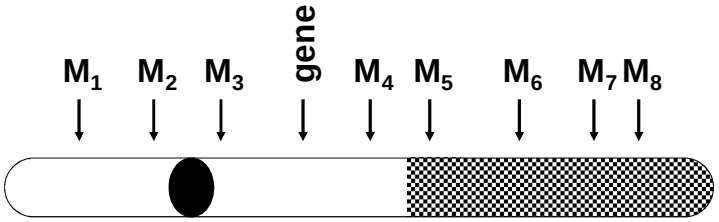
Haplotype Possibilities for the third generation



Consider this on a chromosome-wide basis

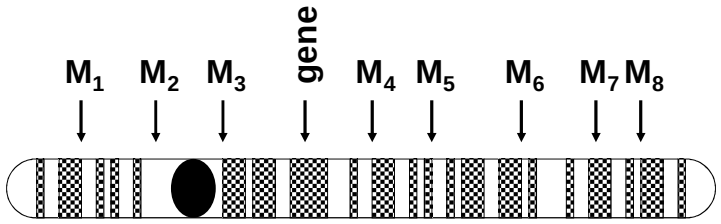


After a few generations ...

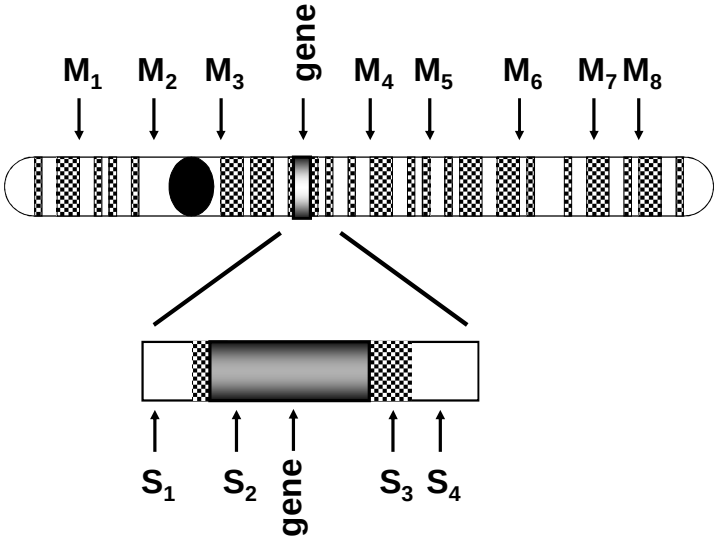


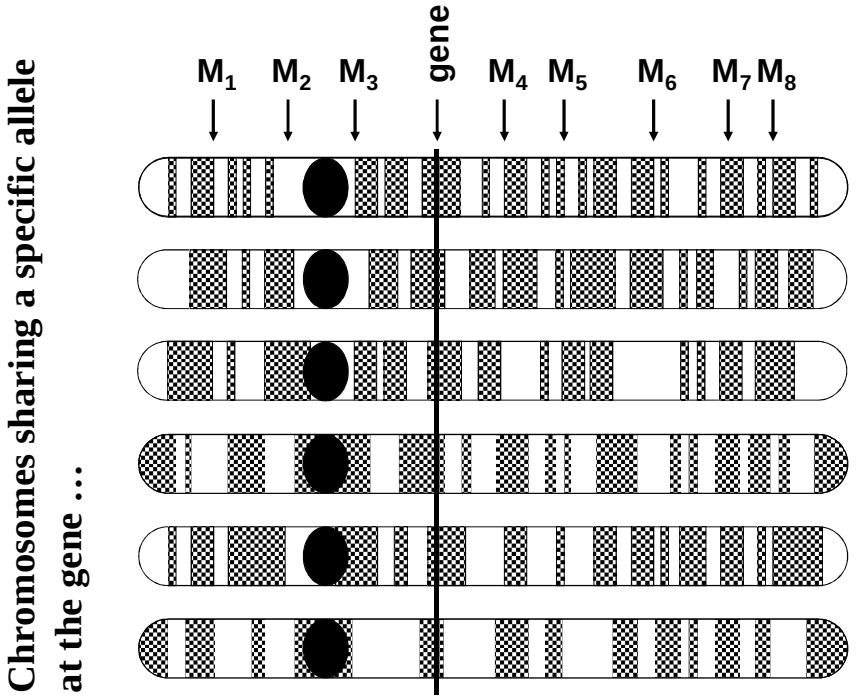
After many generations

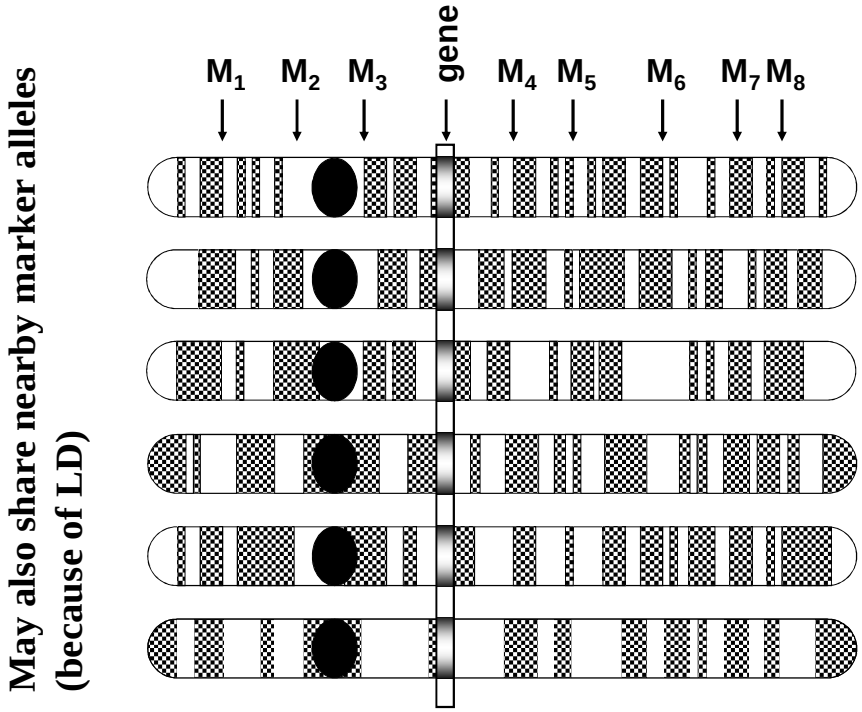
(many recombination events)



A close-up of this region







Identifying a genomic region

- Collect unrelated individuals who express the phenotype of interest.
- Examine many finely spaced markers for each of these individuals.
- Are there marker alleles these individuals seem to share?
- Compare to a control group (who do not express the phenotype of interest).

Example: case-control test

- Dichotomous traits
- Collect
  - individuals expressing the phenotype
  - individuals not expressing the phenotype
  - want two groups to "match" for all but the trait of interest
- Genotype everyone at the marker loci.
- Compare groups.

Case-Control Test

	marker geno 11	geno 12	geno 22	
Affected	$n_{11\|aff}$	$n_{12\|aff}$	$n_{22\|aff}$	n_{aff}
Unaffected	$n_{11\|unaff}$	$n_{12\|unaff}$	$n_{22\|unaff}$	n_{unaff}
	n_{11}	n_{12}	n_{22}	**N individuals**

if $n_{aff} = n_{unaff}$

$$X^2 = S_g \frac{(n_{g|aff} - n_{g|unaff})^2}{n_{g|aff} + n_{g|unaff}} \sim c^2_{(m-1)}$$

(m genotypes)

Logistic Model:

Let G_1 be 1 if Mm and G_2 be 1 if MM and 0 otherwise.

$$\ln\left(\frac{P[D|G_1,G_2]}{P[\overline{D}|G_1,G_2]}\right) = \mu + \alpha_1 G_1 + \alpha_2 G_2 + \beta X$$

D=aff

$\overline{D}$=unaff

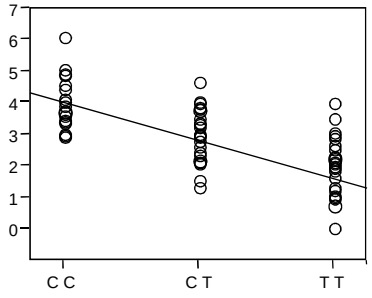
Standard GLM software (e.g. SAS) can fit this.

OR_1 is estimated by $\exp(\hat{\alpha}_1)$, OR_2 by $\exp(\hat{\alpha}_2)$

Can also include covariates

Quantitative trait

- Measure trait value for each individual, *i*: Y_i
- Test if marker genotypes explain a significant amount of the variation in Y_i
- $Y_i = m + g_i + e$
  - g_i is genotype of indiv *i*
  - assumes Y_i ~ normally distributed



Linkage versus Association

Linkage versus LD

Linkage is *defined by* recombination.

- Recombination occurs during meiosis and thus, in turn, is observed via inheritance.
- Loci whose recombination rates are < 0.5 are linked.
- Loci whose recombination rates are = 0.5 are unlinked.
- Linkage is measured via correlated *transmission* of alleles.

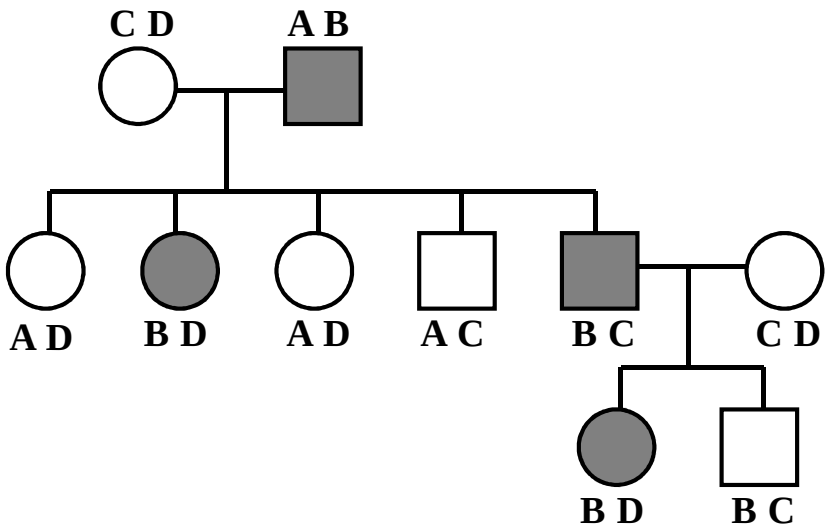
LD is *affected by* recombination (over time).

- LD measures correlation between alleles in a population.
- LD is based on extant haplotypes and *estimates do not rely on or measure inheritance*.
- LD breaks down over time via recombination.
- LD does ***not*** measure correlated *transmission* of alleles.

We can look for *evidence* of linkage …

- Look for marker alleles that are correlated with the phenotype *within a pedigree*.
- Different alleles can be connected with the trait in the different pedigrees.

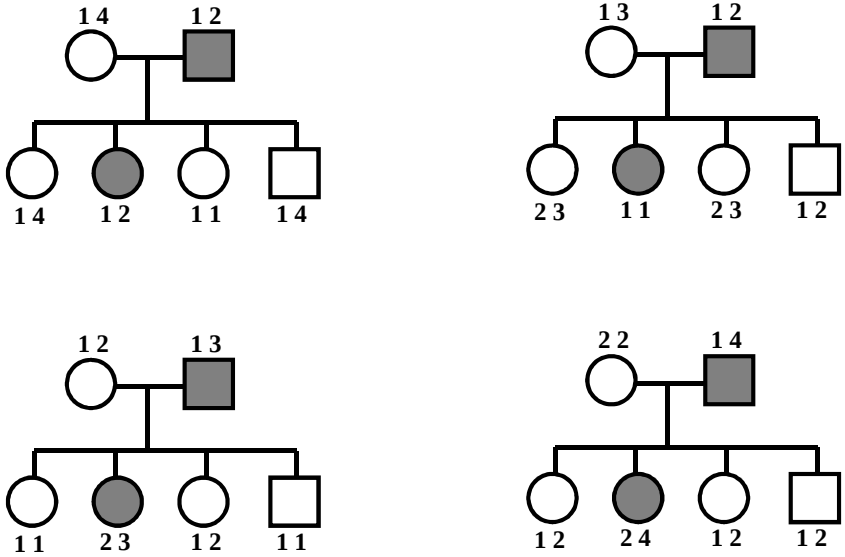
A Pedigree



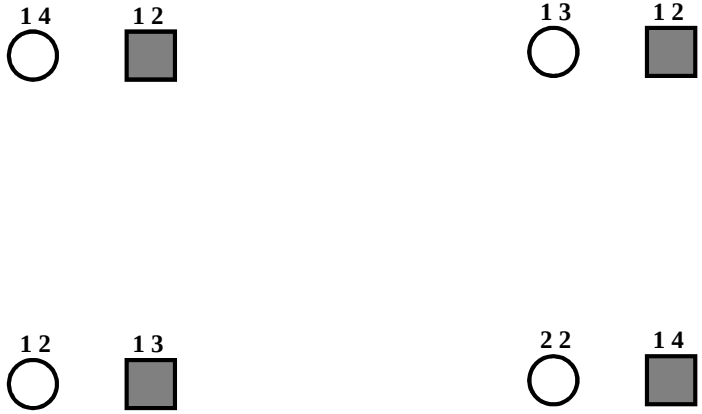
Association

- Marker alleles are correlated with a trait on a *population* level.
- Can detect association by looking at *unrelated* individuals from a population.
- Does not necessarily imply that markers are linked to (are close to) genes influencing the trait.

Linkage …



… but, no Association

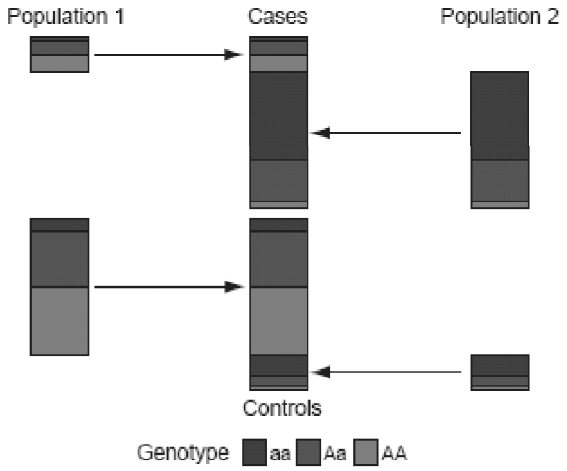


Alleles/genotypes don't appear to be connected to specific phenotypes among unrelated individuals

Case-Control Tests in Structured Populations

- Case-control approach: compare marker allele frequencies of cases to those of controls.
- Hidden population structure can cause "spurious associations"
  - associations due to structure alone

Population structure requires both allele frequency and disease prevalence differences



Case-control tests in structured populations

- Genomic Control (GC: Devlin & Roeder)
  - Calculate bias in distribution of test statistic, then apply correction to case/control association tests
- Structured Association (SA: Pritchard, et.al.)
  - Assign sub-samples to ethnic groups, then match them in case/control analyses to avoid stratification
- Mixed Model (QK: Yu, et.al.)
  - Fit sub-population id results from SA together with kinship coefficient estimates in mixed model analysis
- PCA (Eigenstrat: Price, et.al.)
  - Adjust phenotype and genotype measures using principle components (indicative of population structure)

Genomic Control

Rationale:

For any marker under H_o, $E(\chi^2) = df$

e.g., for SNPs, $E(\chi^2_1) = 1$

Population structure may cause inflation in test statistics:

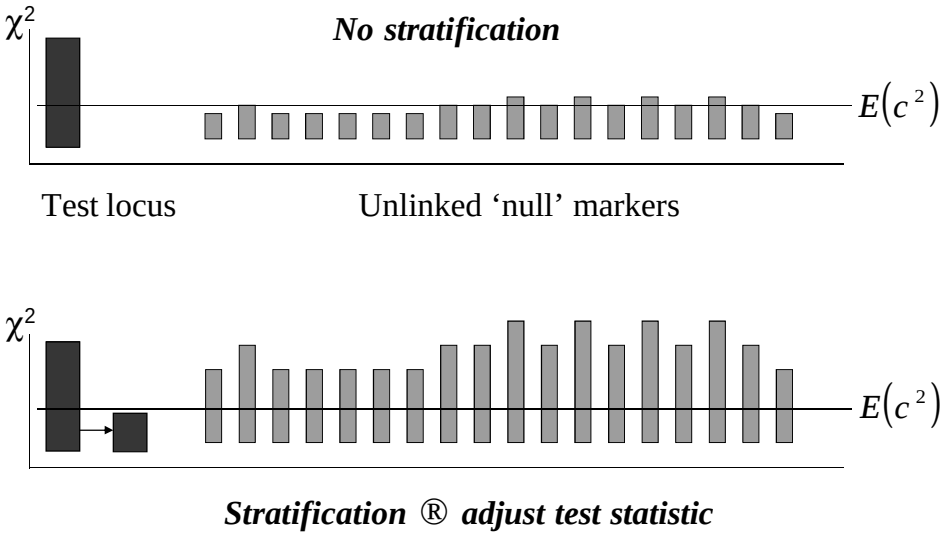
$E(\chi^2_1) > 1$

Procedure:

Perform many tests -- many SNPs

Use test statistics from these tests to adjust for inflation

Genomic control



Stratification ® adjust test statistic

Genomic control

§ Simple estimate of inflation factor

$$\hat{l} = median\{c_1^2, c_2^2, \mathbf{K}, c_N^2\} / 0.456$$

– using the median protects from outliers

§ i.e. if some of the null markers are also QTL

– bounded at minimum of 1

§ i.e. should never increase test statistic

– principle extended to multiple alleles, haplotypes, quantitative traits

§ Must formulate all tests as 1 df tests, however

Genomic control

§ λ Inflation factor

$$l \approx 1 + RF\sum_k (f_k - g_k)^2$$

R number of cases (controls)

F Wright's F_{ST} coefficient of inbreeding

g_k (f_k) Proportion of cases (controls) from subpopulation

§ Example

– 2 equifrequent subpopulations, F_{ST} = 0.01

– Disease twice as common in one subpopulation

– R = 1000

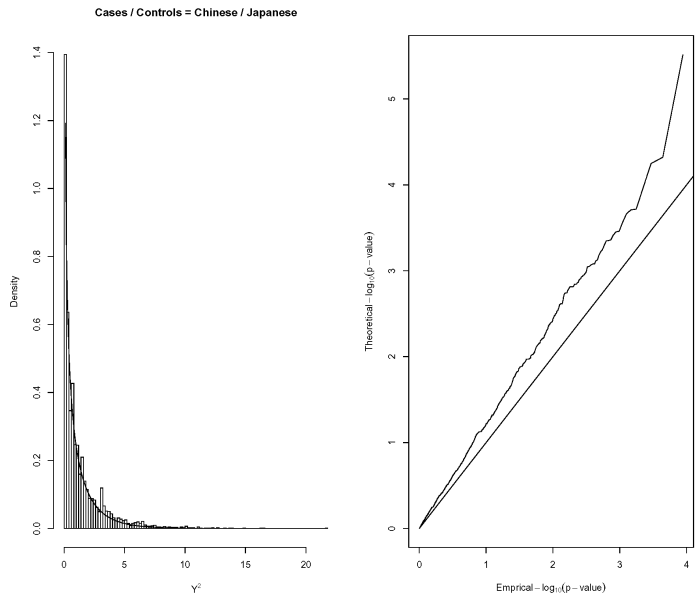
– l » **1.5**

TSC Dataset

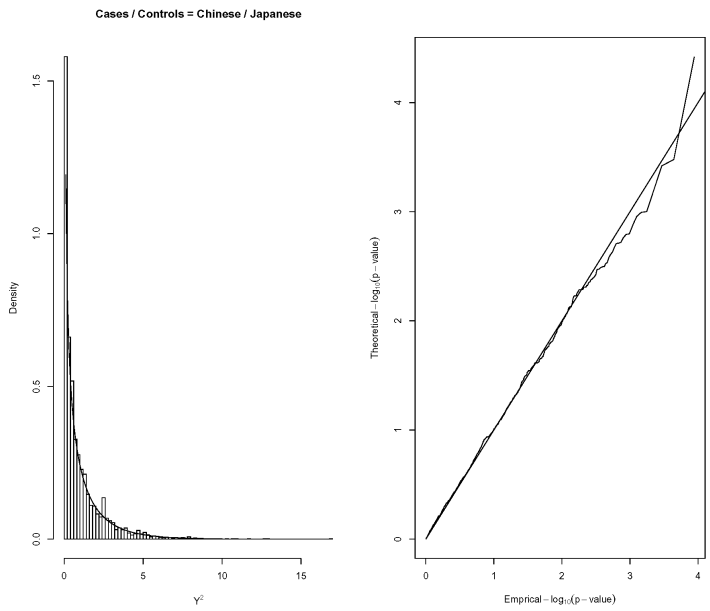
12,000 markers genotyped on 3 ethnic diversity panels

- 40 'Caucasians'
- 40 'African Americans'
- 40 'Asian Americans' **à** 30 Chinese/10 Japanese

Chinese/Japanese data: Before Genomic Control



Chinese/Japanese data: After Genomic Control



STRAT

- STRuctured population Association Test.
  - Pritchard *et al*. (2000).
- Collect cases and controls.
- Genotype everyone for a number of unlinked markers.
- Infer population structure.
- Use inferred structure in modified case-control test.

Inferring Population Structure

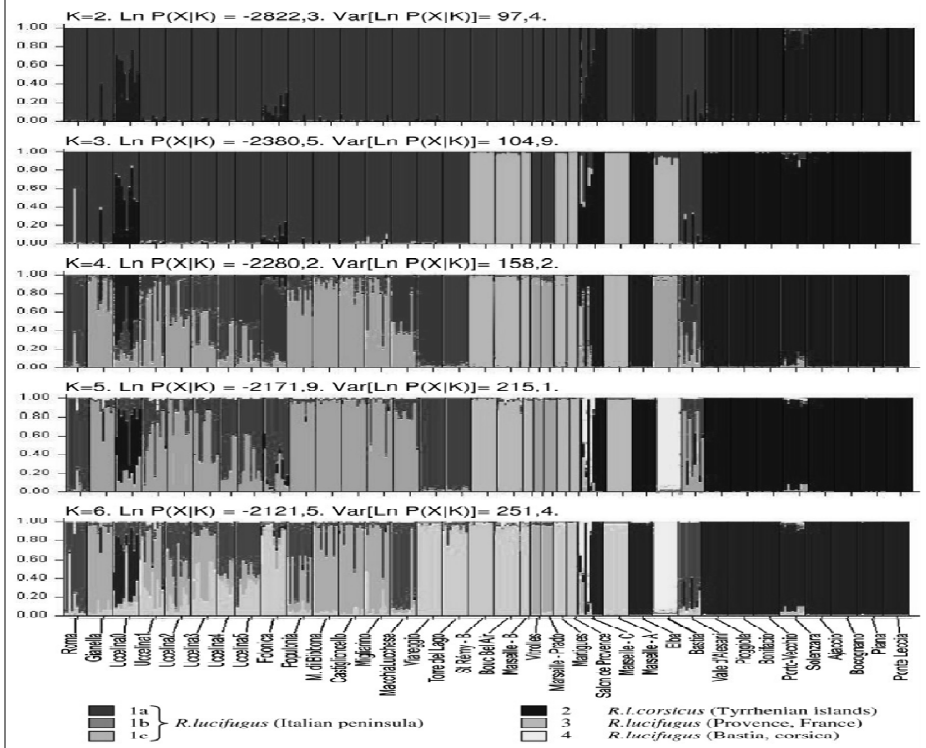
- Assume admixed population with K contributing founding populations.
- Each of the K founding populations was in equilibrium (HWE, no LD).
- In this generation, each allele copy originated in one of the founding populations.
- Want to figure out the probability alleles in each individual originated in population *k*: **Q** vector.

Variables

- $q^{(i)}_k$ = proportion of individual *i*'s genome that originated in population *k*.
  - **Q**
- $x^{(i,a)}_l$ = allele copy for individual *i* at locus *l* (a = 1,2)
  - **X**
- $z^{(i,a)}_l$ = population of origin of allele copy $x^{(i,a)}_l$
  - **Z**
- p_{klj} = allele frequency for allele *j* at locus *l* in population *k*
  - **P**

Inferring Population Structure

- Bayesian approach
- Likelihood:
  - Pr (**X** | **Z**, **P**, **Q**)
- Priors:
  - Pr (**Z**), Pr (**P**), Pr (**Q**)
- MCMC to sample
  - $P^{(m)}$, $Q^{(m)}$ from Pr (**P**, **Q** | **X**, $\mathbf{Z}^{(m-1)}$)
  - $Z^{(m)}$ from Pr (**Z** | **X**, $\mathbf{P}^{(m-1)}$, $\mathbf{Q}^{(m-1)}$)



Lefebvre et al.
BMC Evolutionary Biology 2008 8:38
doi:10.1186/1471-2148-8-38

Using this in a case-control test

- Collect cases, controls.
  - genotype for a number of unlinked markers.
- For these individuals, infer **Q.**
  - Proportion of each individuals *i*'s genome originating in population *k.*
- Calculate likelihood of genotypes at the candidate gene under the null hypothesis (no association between candidate gene and phenotype).
- Calculate likelihood of candidate gene genotypes under alternate hypothesis (association).
- Perform a likelihood ratio test.

Modified Case-Control Test

§ $Pr_{H_0}(c^{(i,a)} = j \mid Q, P, \phi) = \Sigma_k \, q^{(i)}_k \, p_{kj}$

§ $Pr_{H_1}(c^{(i,a)} = j \mid Q, P, \phi) = \Sigma_k \, q^{(i)}_k \, p^{[\phi(i)]}_{kj}$

Modified Case-Control Test

$$\Lambda = \frac{Pr_{H_1}(C \mid \hat{P}_1, \hat{Q})}{Pr_{H_0}(C \mid \hat{P}_0, \hat{Q})}$$

$$C(t) \sim Pr_{H_0}(\bullet \mid \hat{Q}, \hat{P}_0, \phi)$$

$$\alpha = \#\{m : \Lambda(C^{(m)}) > \Lambda(C)\} / M$$

"QK": A unified mixed-model method for multiple levels of relatedness

$$\mathbf{y} = \mathbf{X}b + \mathbf{S}a + \mathbf{Q}v + \mathbf{Z}u + \mathbf{e}$$

Trait values

Environmental covariates, etc.

Candidate SNP effects

Subpopulation effects (fixed effects)

Background Genetic effects (random) Var(u) = $2KV_A$

Q from SA, prop. of individual's genome from each sub-population (broad effects due to population)

K kinship coefficient matrix (adjust for different levels of relatedness)

Yu JM, et al. Nat. Genet. 38:203-208

PCA: Eigenstrat

§ Goal: use genome-wide markers to estimate population structure correction factors
  – on an individual (rather than population) basis

§ Does not assign individuals to subpopulations

§ Use correction factors to adjust genotypes and trait values to perform association analysis
  – correction method equivalent to performing regression with adjustment values as covariates

§ Correction factors based on principal components of the genotype matrix

Eigenstrat

- Genotype matrix:
  - individuals in rows
  - SNPs in columns
  - cells contain 0, 1 or 2

    = number of an arbitrarily chosen allele for that individual's genotype at that SNP
- Calculate column averages (for each SNP)
- Adjust each cell by subtracting its column ave and dividing by a "normalizing" factor: $\sqrt{p_j(1-p_j)}$
- Call this adjusted genotype matrix: **M**

Eigenstrat

- Calculate estimate of var/covar matrix:

  $\mathbf{X} = (1/n)\,\mathbf{M}\mathbf{M}'$ (n = #SNPs)

  - cells of **X** are covariances for pairs of individuals (diagonals are variances)
- Calculate eigenvectors, eigenvalues, PCs of **X**
- Main idea: if there is population structure evident in the data, this structure will be captured in the first k-1 PCs of **X**
  - k = number of subpopulations

General note about PCs

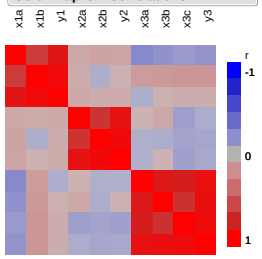
- The PCs are linear combinations of the original variables.
- These combinations are chosen so that they capture maximum available variation

  and so that they are independent of all other combinations.
- If there is a block structure to the covar matrix, the PCs should capture this.

PCs and blocky var/covar matrix

Multivariate

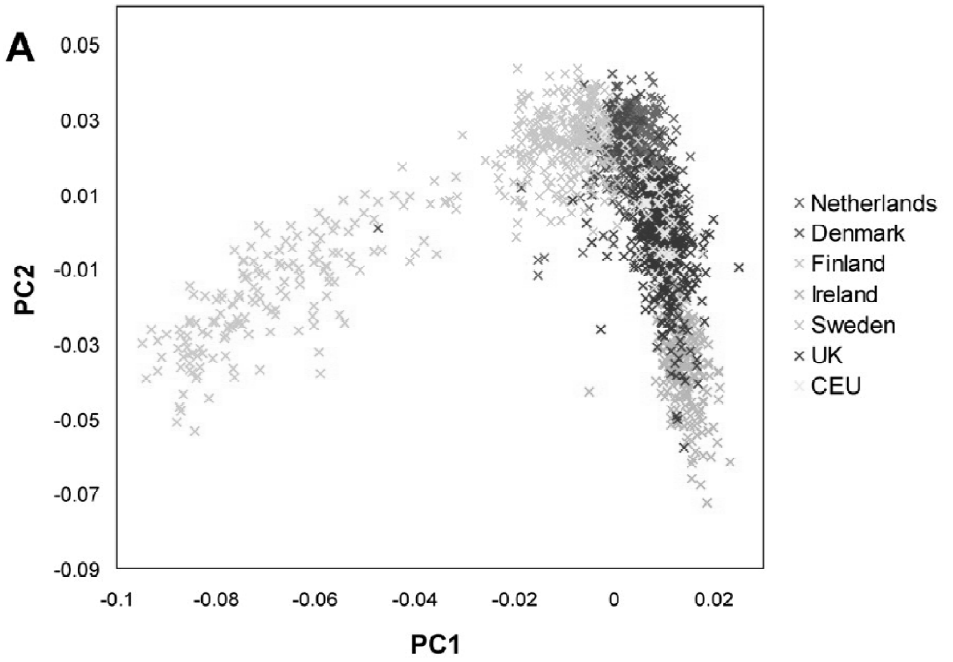
Color Map On Correlations



Principal Components / Factor Analysis

Principal Components: on Covariances

Eigenvalue	23.1511	9.2978	7.8604	1.0864	0.5471	0.4778	0.3123	0.0000	0.0000	0.0000
Percent	54.1763	21.7579	18.3944	2.5423	1.2803	1.1180	0.7308	0.0000	0.0000	0.0000
Cum Percent	54.1763	75.9342	94.3286	96.8708	98.1511	99.2692	100.0000	100.0000	100.0000	100.0000
Eigenvectors										
x1a	-0.03767	0.08146	0.30931	-0.08082	-0.30573	-0.29559	0.61283	0.09894	0.55828	-0.10896
x1b	0.05123	0.09904	0.48009	0.08402	0.28512	0.24439	-0.52505	0.09894	0.55828	-0.10896
y1	0.01356	0.18050	0.78939	0.00319	-0.02062	-0.05121	0.08778	-0.09894	-0.55828	0.10896
x2a	-0.01429	0.31639	-0.06333	0.15062	-0.09678	0.67054	0.28404	0.56858	-0.10018	0.00300
x2b	-0.03020	0.47255	-0.11209	-0.17193	0.07793	-0.56868	-0.26619	0.56858	-0.10018	0.00300
y2	-0.04449	0.78893	-0.17542	-0.02131	-0.01885	0.10186	0.01785	-0.56858	0.10018	-0.00300
x3a	0.36630	0.03877	-0.04784	0.16659	0.80724	-0.11076	0.41277	0.00000	0.00000	0.00000
x3b	0.35368	0.04845	-0.02645	0.65485	-0.27136	-0.16230	-0.09864	0.01601	0.10782	0.56697
x3c	0.40233	-0.01600	0.00961	-0.68985	-0.01751	0.16805	0.00335	0.01601	0.10782	0.56697
y3	0.75601	0.03245	-0.01684	-0.03500	-0.28887	0.00575	-0.09530	-0.01601	-0.10782	-0.56697



http://genome.cshlp.org/content/19/5/804.full.html

Eigenstrat

- Statistical test to determine significant eigenvalues
  - # significant results = # of subpops (k) - 1
  - Eigenvectors (corresponding to significant eigenvalues) termed “axes of variation”
- Adjust original genotype & phenotype scores using these “axes of variation” (eigenvectors)
- Use adjusted genotype and phenotype values in case-control test for association

Eigenstrat

- Adjusted score, SNP_i, $indiv_j$
  - adjust the genotype ($c_{i,j}$) and trait ($t_{i,j}$) by the coordinates along a given PC (a_i):

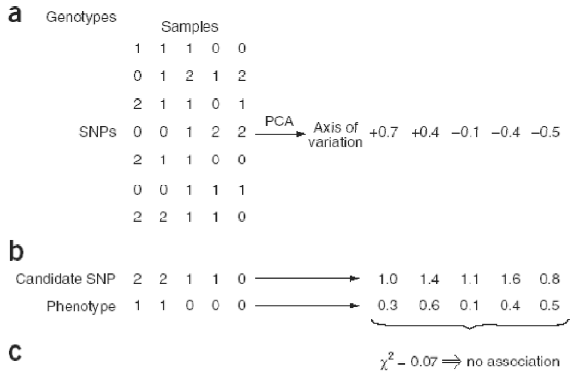
$$c_{i,j}^{adj} = c_{i,j} - \gamma_j^{(c)} a_i \quad \text{and} \quad t_{i,j}^{adj} = t_{i,j} - \gamma_j^{(t)} a_i$$

where $\gamma_j^{(\cdot)}$ is from regression of genotype/phenotype on PC a_i:

$$\gamma_j^{(c)} = \frac{\sum_i a_i c_{i,j}}{\sum_i a_i^2} \quad \text{and} \quad \gamma_j^{(t)} = \frac{\sum_i a_i t_{i,j}}{\sum_i a_i^2}$$

Eigenstrat

- Adjusted score, SNP(i), indiv(j)



Eigenstrat

- Test statistic for adjusted genotype/phenotype values: adjusted Armitage trend test
- Test stat = $(N - K - 1) \times corr^2 (geno_{adj}, pheno_{adj})$
- Compare to $\chi^2_{(1)}$ distribution to test H_0

OPEN ACCESS Freely available online PLoS GENETICS

An *Arabidopsis* Example of Association Mapping in Structured Samples

Keyan Zhao[1], María José Aranzana[1], Sung Kim[1], Clare Lister[2], Chikako Shindo[2], Chunlao Tang[1], Christopher Toomajian[1], Honggang Zheng[1], Caroline Dean[2], Paul Marjoram[3], Magnus Nordborg[1]*

- Compared SA, QK and PCA using real data
- Best results appeared to be using either of:
  - QK as published (using Q matrix from SA)
  - technique of QK but substituting PCs for Q

Multiple Testing Issues

Background

p values

p values

- *p* value: Probability of seeing your data or more extreme data *IF the null hypothesis is TRUE*
- To perform a hypothesis test, calculate the appropriate test statistic.
- *This test statistic is a function of your data.*
- Case-control statistic (equal # of cases/controls):

$$T = \frac{(n_{A|aff} - n_{A|unaff})^2}{n_{A|aff} + n_{A|unaff}}$$

⇐ each $n_{A|pheno}$ is a count from your data

p values

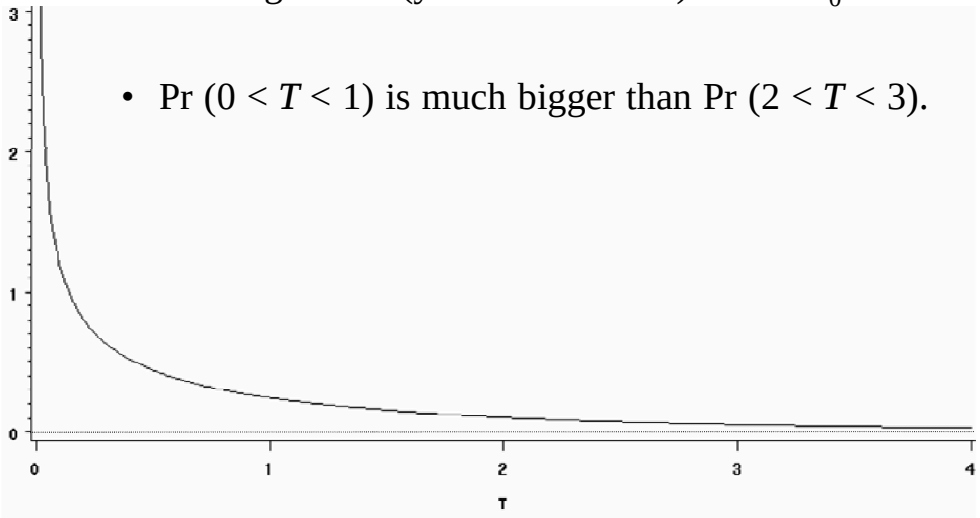
- *p* value: Probability of seeing your data or more extreme data *IF the null hypothesis is TRUE*.
- Case-control statistic under H_0:

$$T = \frac{(n_{A|aff} - n_{A|unaff})^2}{n_{A|aff} + n_{A|unaff}} \sim \chi^2_{(1)}$$

- If H_0 is true (no association), expect T (your data) approximately to follow a chi-square distribution.

c^2 distribution, 1 df (H_0)

- Area under the curve represents the probability for ranges of *T* (your test statistic) under H_0.



- Pr ($0 < T < 1$) is much bigger than Pr ($2 < T < 3$).

c^2 distribution, 1 df (H_0)

- If the test statistic for our data is (for example) $T = 8.42$, the probability of seeing this value or a bigger one is very small under H_0.
- The probability is not zero, so it could happen!!
- But, perhaps this is evidence that H_0 is not true.

- If our rule is to say that H_0 is false every time we see a value *T* this big or bigger, we will be making a mistake occasionally (since these values will occasionally occur even if H_0 is true).
- These mistakes are the ***false positives***.
- We usually try to control the rate at which these occur.
- (while still finding at least some true positives)

One moral of the story …

- The *p* value is ***not*** the probability that the null hypothesis is true.
- It represents the probability *of the data* under the null hypothesis.

Multiple testing problem

- Performing one test at an alpha level of 0.05 implies 5% chance of rejecting a true null hypothesis (false positive).
- Performing 100 tests at $\alpha = 0.05$ when all 100 null hypotheses are true, we expect 5 of the tests to give false positive results.
- Pr (at least one false positive) = 1 - Pr (no false positives) = $1 - (0.95)^{100} = 0.994$
  - (if the tests are independent)

Multiple testing: Controlling the false positive rate

Family-wise Error Rate (FWER)

- FWER = probability of rejecting at least one true H_0 (among all the hypotheses being tested).
  - Examining each H_0 individually.
- Controlling FWER:
  - Pr (reject at least one true H_0) = α
  - (*e.g.* $\alpha = 0.05$).
- Bonferroni adjustment controls FWER
  - rejection level for each of M tests is $\alpha_i = \alpha / M$
  - $\alpha = 1 - (1 - \alpha_i)^M \approx \alpha_i M$ (α = FWER) (α_i = indiv test rejection level)

Problems with Bonferroni

- Highly conservative
  - May need very small p values to reject H_0
  - May lose power to detect true effects
- Other ways to control FWER …

Stepwise procedure (Hochberg*)

- Perform each test and calculate a p-value, p_j
  - $p_1 = 0.75$, $p_2 = 0.007$, $p_3 = 0.58$, $p_4 = 0.22$, $p_5 = 0.06$
- Order the p-values, $p_{(i)}$
  - $p_{(1)} = 0.007$, $p_{(2)} = 0.06$, $p_{(3)} = 0.22$, $p_{(4)} = 0.58$, $p_{(5)} = 0.75$
- Find largest $p_{(i)}$ such that $p_{(i)} \leq \alpha / (M\text{-}(i)+1)$
- Reject all H_{0j} where $p_j \leq p_{(i)}$
  - reject for tests with $p_j \leq 0.007$ →
- Potential increase in power

(modification of Holms' "step down" procedure#)

(i)	$p_{(i)}$	$\alpha/(M\text{-}(i)+1)$
1	0.007	0.01
2	0.06	0.0125
3	0.22	0.0167
4	0.58	0.025
5	0.75	0.05

*Biometrika 75:800-802 (1988)
#Scand.J.Stat 6:65-70(1979)

Permutation tests

- P-values are the probability of seeing the data or more extreme data if the null hypothesis is true
  - Pr (marker test result when marker is not associated)
- Sometimes know that test statistic follows a given distribution under H_0 (e.g. chi-squared)
  - calculate p-values using this distribution
- Sometimes do not know what the distribution is
  - don't know what are considered "big" test statistics
  - don't know how to get a p-value for a test statistic

Permutation tests

- Allow you to calculate a p-value when you don't know the distribution of the test statistic
  - P-values are the probability of seeing the data or more extreme data if the null hypothesis is true
  - Pr (≥ marker test result when marker is not associated)
- Empirically calculate what kinds of results to expect when the null hypothesis is true:
  - if there is no association between a marker and a trait, how big might our test statistics be?

Permutation tests

indiv	pheno	SNP1	SNP2	SNP3	…	SNPj
1	aff	AA	AB	AB		BB
2	unaff	AA	AA	BB		AB
3	unaff	BB	AB	AA		AB
4	aff	AA	AB	BB		AA
5	unaff	AB	AB	AA		BB
…						
N	aff	AB	BB	AB		AB

Separate phenotypes from genotypes

indiv	pheno
1	aff
2	unaff
3	unaff
4	aff
5	unaff
…	
N	aff

indiv	SNP1	SNP2	SNP3	…	SNPj
1	AA	AB	AB		BB
2	AA	AA	BB		AB
3	BB	AB	AA		AB
4	AA	AB	BB		AA
5	AB	AB	AA		BB
…					
N	AB	BB	AB		AB

Randomly assign genotypes to individuals

indiv	pheno
1	aff
2	unaff
3	unaff
4	aff
5	unaff
…	
N	aff

indiv	SNP1	SNP2	SNP3	…	SNPj
1003	AA	AB	AB		BB
204	AA	AA	BB		AB
5008	BB	AB	AA		AB
4251	AA	AB	BB		AA
12	AB	AB	AA		BB
…					
246	AB	BB	AB		AB

Randomly assign genotypes to individuals

indiv	pheno
1	aff
2	unaff
3	unaff
4	aff
5	unaff
…	
N	aff

indiv	SNP1	SNP2	SNP3	…	SNPj
1003	AA	AB	AB		BB
204	AA	AA	BB		AB
5008	BB	AB	AA		AB
4251	AA	AB	BB		AA
12	AB	AB	AA		BB
…					
246	AB	BB	AB		AB

- Calculate a test statistic using the permuted data
  - this gives us a test statistic value under the assumption that the marker is not associated with the phenotype

Repeat many times

- K permutations
- yields K values of the test statistic under “no association”
- P-value: probability of seeing your data or more extreme data if the null hypothesis is true
- How many times was your permuted test statistic bigger (more extreme) than your real test statistic?

  28% of the time, then p=0.28

  1% of the time, then p=0.01

Controlling FWER with permutation

- FWER = Pr (at least one false positive)
  - control FWER at, say, 0.05
  - Pr (at least one false positive among all the tests) = 0.05
- Want to set the significance threshold to a value such that, if you do M tests,
  - 5% of the time you expect you will have ***at least one*** permuted (null hypothesis) test result that is ≥ this threshold value.

Controlling FWER with permutation

- FWER = Pr (at least one false positive) = α (e.g. 0.05)
  - Pr (at least one false positive test among all tests) = 0.05
- Perform permutation procedure
  - make H_0 true for all the markers
- Test all the markers
- Keep largest test statistic ("top" marker)
  - if you call this test statistic significant, then you have created one false positive.

Controlling FWER with permutation

- Perform permutation procedure K times
  - for each permutation keep test statistic value for top marker
- After K permutations, have K "top" test statistic values.
  - If you said the top test statistic was significant for a given permutation, then would have one FP for that permutation.
  - Pr (at least one FP test among all tests) = 0.05
- Keep upper 5% of the K largest test statistics.
  - Any *real* test statistic in this range (or higher) is statistically significant.

Controlling FWER with permutation

Marker	Perm1	Perm2	Perm3	Perm4	…	PermK	largest
1	0.28	0.04	4.87	0.09		0.75	17.69
2	2.25	0.20	0.55	0.47		1.68	20.28
3	1.09	1.19	1.40	0.35		9.20	16.27
4	0.04	4.61	0.01	1.27		0.13	17.11
5	0.91	0.77	0.82	0.97		1.10	16.54
…							…
M	1.18	1.01	2.86	0.54	…	3.83	18.66

Take top α = 5% of these

any *real* test statistic in this top 5% range (or higher) is considered statistically significant

False Positive Rates

versus

False Discovery Rates

	Reject H_0	Do not reject H_0	
H_0 true	**F** (# false positives)	m_0 - F	m_0
H_0 false	**T**	m_1 - T	m_1
	S	M - S	M

- False positive rate = **F / m_0**
  - What proportion of true null hypotheses are rejected?
  - Of all the times H_0 is true, how often do you reject it?
- False discovery rate (FDR) = **F / S**
  - What proportion of rejections are when H_0 is true?
  - Of all the times you reject H_0 how often is H_0 true?

Benjamini & Hochberg*

- Proposed a method to control FDR.
- Provides weak control of FWER.
  - strong control: FWER is controlled for all combinations of true and false null hypotheses.
  - weak control: FWER is only controlled when all null hypotheses are true.
- Error rate (FDR) is equivalent to FWER when all H_0 are true but is smaller otherwise.
- Does not provide an estimate the FDR itself.

*J.Royal Stat Soc B-methodol. 57:289-300 (1995)

Benjamini & Hochberg: procedure

- Order all p_i values ($p_{(1)}, p_{(2)}, \ldots$)
- Find largest $p_{(i)}$ such that $p_{(i)} \leq i\alpha / M$.
- Reject H_0 for all tests with p-values $\leq p_{(i)}$

Storey, Taylor, Siegmund, Tibshirani*

- Proposed a method of controlling FDR.
- Provides an estimate of FDR.
- Less conservative that Benjamini & Hochberg's method.
- Introduced: ***q* values**

*e.g. PNAS (2003) 100:9440-9445.

q values

- Calculate test statistic for i^{th} hypothesis test:
  $T_i = 8.42$ (for example).
- The *q* value for this T_i is the false discovery rate you expect in your experiment if you call this value "significant" (*i.e.* if you reject H_0 when you see this or a more extreme value).

For example …

- You've performed many tests, and here are five of them (sorted by *p* value):
- Call the top one significant … everything with a smaller *p* value will also be significant.
- What proportion of those called significant will be false positives?
- This is the *q* value for that test.

	T	*p* value
1	12.26	0.000463
2	8.42	0.003711
3	6.71	0.009587
4	5.52	0.018800
5	3.18	0.074545
…		

For example …

- If you call the next one significant (#2), the number of significant tests increases …
- (everything with a smaller *p* value will also be significant).
- Its *q* value will be bigger.
- Being less strict about which *p* values are considered significant causes the false discovery rate to increase.

	T	*p* value
1	12.26	0.000463
2	8.42	0.003711
3	6.71	0.009587
4	5.52	0.018800
5	3.18	0.074545
…		

Controlling the FDR

- Being less strict about which *p* values are considered significant will cause the false discovery rate to increase.

1. Decide on an acceptable rate beforehand (*e.g.* 0.05).
2. Perform all the tests and calculate all *p* values and *q* values.
3. Determine the significance threshhold based on the *q* value cutoff that satisifies (1) for these tests.

- This provides a data-dependent significance threshhold that controls the FDR.

Estimating the FDR

- Defined FDR to be F / S

  # false positives / number of significant results

- F and S are random variables.

- Technically, S can take on a value of zero, and the proportion F / S would be undefined.

FDR: Technicalities

- FDR = E $[F / S] \approx$ E $[F]$ / E $[S]$

- Pr $(S = 0) > 0$ (when $S = 0$ F/S is undefined).
  - FDR = E $[F / S \mid S > 0] \times$ Pr $(S > 0)$
  - Positive (or conditional) FDR = E $[F / S \mid S > 0]$ (pFDR)

- For large M (lots of tests), Pr $(S > 0) \approx 1$
  - not much distinction between these definitions
  - stick with E $[F]$ / E $[S]$.

Estimating FDR for threshhold *t*

- Must define a threshhold (based on a *p* value) to calculate FDR.
- For instance, if we say that a *p* value of ≤ 0.30 is significant, we'll have a very high FDR.
- If we say that we must have a *p* value of ≤ 0.0001 to reject H_0, then we should have a much smaller FDR
  - (assuming the same number of tests are performed)

- FDR for threshhold t is FDR(t)

Estimating FDR(t)

- FDR(t) ≈ E[$F(t)$] / E[$S(t)$]
  - Expected number of false positives / expected number of positives.

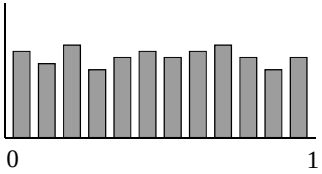
- Estimate E[$S(t)$] with the observed $S(t)$, the number of your tests you reject when using t as the threshhold.
- To estimate E[$F(t)$], exploit the fact that for the null hypothesis, p values are uniformly distributed:
  - Pr (p value ≤ t | H_0) = t
  - So, E[$F(t)$] = $m_0 \times t$.

Estimating E[$F(t)$] (cont)

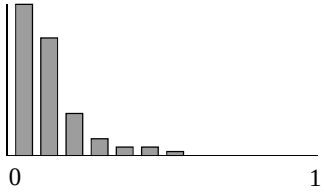
- E[$F(t)$] = $m_0 \times t$.
- m_0 is the number of true null hypotheses.
  - Don't know this.
- Estimate $m_0 = \pi_0 \times M$.
  - π_0 is the proportion of true null hypotheses.
  - (M is the total number of tests)
- So … finally … need to estimate π_0.

Estimating π_0

- Of all the tests performed, this is the proportion where the null hypothesis is true.
- Again, use the property that p values are uniformly distributed when H_0 is true:

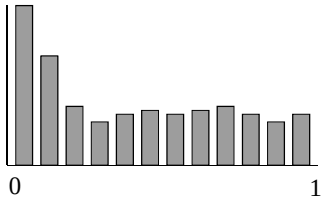


- p values are piled up around zero when H_0 is false:



Estimating π_0

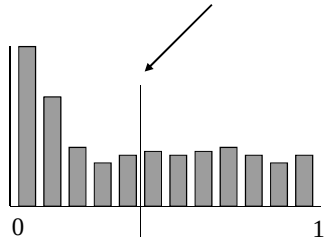
- We have a mixture of results in reality …



- The p values towards the right part of this distribution should mostly arise when H_0 is true.
- Use this part of the distribution to estimate π_0.

Estimating π_0

- Choose a tuning parameter λ $(0 < \lambda < 1)$.



- Now, estimate π_0 as:

$$\frac{\#\text{ of all p values} > \lambda}{M\,(1 - \lambda)}$$

Estimating π_0

- $M\,(1 - \lambda)$ tells us how many p values would be in this region if all the hypotheses were null.
- The numerator is how many we actually found (some of the total number of tests will be missing from this region, since for them, the null hypothesis was false, and most of their *p* values were in the region we avoided examining).
- The proportion is then $\hat{\pi}_0$.

Estimating FDR(t)

- $\text{FDR}(t) \approx \text{E}[F(t)] / \text{E}[S(t)]$

- estimate $\text{E}[F(t)] = \hat{\pi}_0 \times M \times t$
- estimate $\text{E}[S(t)] = S\,(t)$

$$\hat{FDR}(t) = \frac{\hat{p}_0 \cdot M \cdot t}{S(t)}$$

- $\hat{q}(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t)$

Some things about FDR control …

Methods tend to work best when:

- the proportion of true effects is known to be large;
- a moderate amount of false-positive results can be tolerated;
- or, a follow-up study is anticipated.

Some things in general …

- Most techniques for multiple testing adjustment built on the assumption that individual tests are independent.
- Association tests are probably not independent if the markers are in LD.
- Modifications for many of these methods have been proposed to handle correlated tests.

Relevant Research in the area of genetic analysis

(though generally applicable)

Linkage mapping of QTL

- Cheverud (Heredity,2001,87:52-58)
- May be testing ***M*** markers, but probably don't have ***M*** independent tests.
- Instead, calculate m_{eff} = number of "effective" tests:
  - calculated from the variance of the eigenvalues of the observed marker correlation matrix.
- Use this value rather than ***M*** in the multiple test correction procedure of choice.
- Since $m_{eff} \leq M$, adjustments less conservative.

Using effective number of markers (m_{eff})

- "simple" Bonferroni correction:

  $\alpha_i = \alpha / m_{eff}$

- Šidák correction:

  $\alpha_i = 1 - (1- \alpha)^{1/ m_{eff}}$

- m_{eff} -based step-up procedure: reject all tests with p-values smaller (≤) than

$$\max (i): \; p(i) \leq \frac{a}{m_{eff}} + \frac{i-1}{M-1}\left(a - \frac{a}{m_{eff}}\right)$$

(modified from B-H)

Association analysis

- Nyholt (Am.J.Hum.Genet.,2004,74:765-769)
- Adapted theory of Cheverud.
- Explicitly incorporate estimates of LD into calculation of m_{eff}.

Association analysis

- Li & Ji (Heredity,2005,95:221-227)
- Updated methodology of Cheverud, Nyholt.
- Alternative procedure for the calculation of m_{eff}.
  - L-J's $m_{eff} \leq$ C's m_{eff}

Mapping populations and tools

Maize Nested Association Mapping (NAM)

Edward Buckler

USDA-ARS

Cornell University

http://www.maizegenetics.net

The Maize Diversity Project

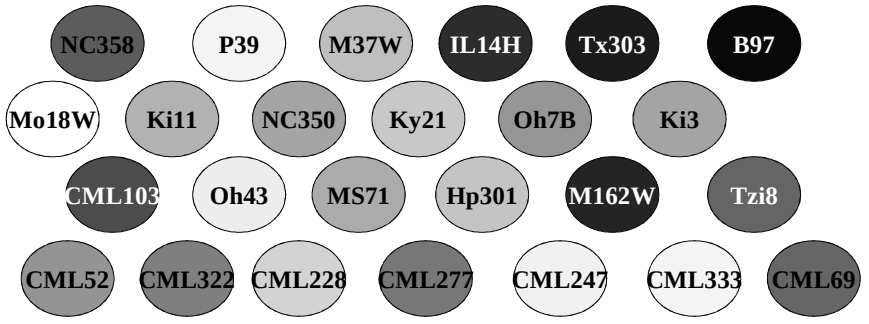
- Ware, Cold Spring Harbor Lab.
- Kresovich, Cornell University
- Holland, North Carolina State Univ.
- McMullen & Flint-Garcia, University of Missouri
- Doebley, University of Wisconsin
- USDA-ARS

- www.panzea.org

(The following slides are compliments of Dr. Ed Buckler, Cornell University)

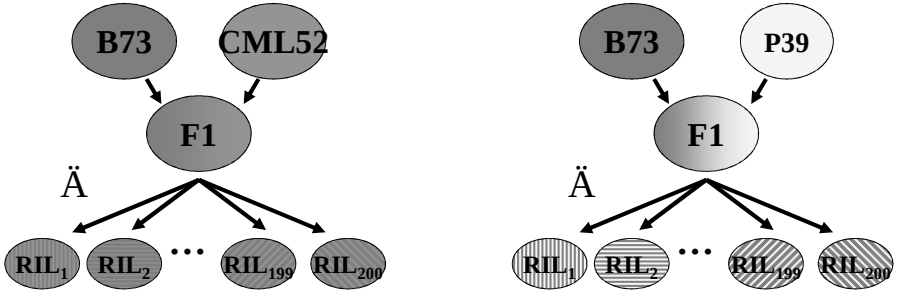
Maize NAM

- 25 diverse lines were chosen to maximize diversity based on SSRs
- Crossed to B73 for a reference design
- Project joint efforts of Holland, McMullen, Kresovich, and Buckler groups
- 60% tropical origin
- 12% maize oddities (popcorn and sweet corn)

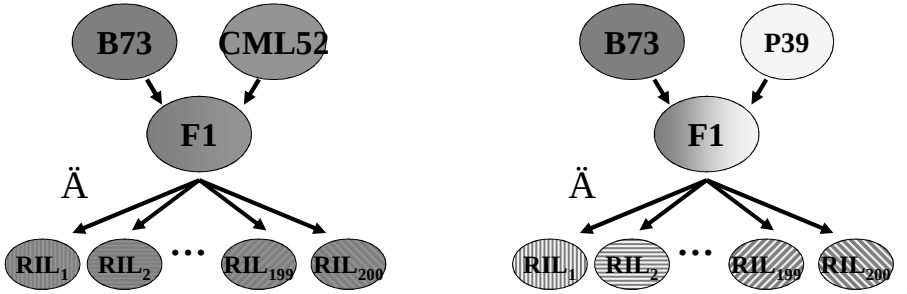


Maize NAM

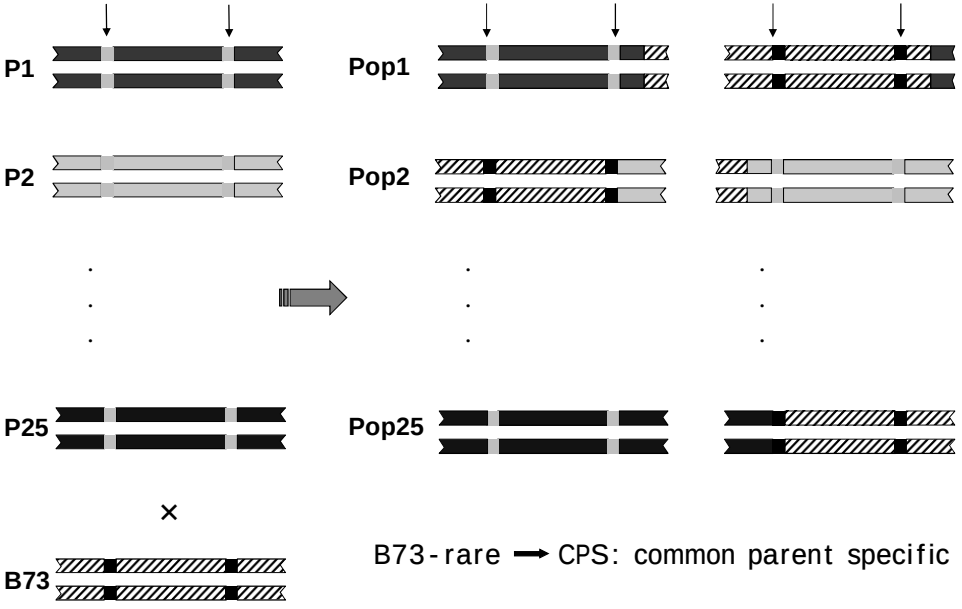
- 25 diverse lines were chosen to maximize diversity based on SSRs
- Crossed to B73 for a reference design
- Project joint efforts of Holland, McMullen, Kresovich, and Buckler groups
- 60% tropical origin
- 12% maize oddities (popcorn and sweet corn)



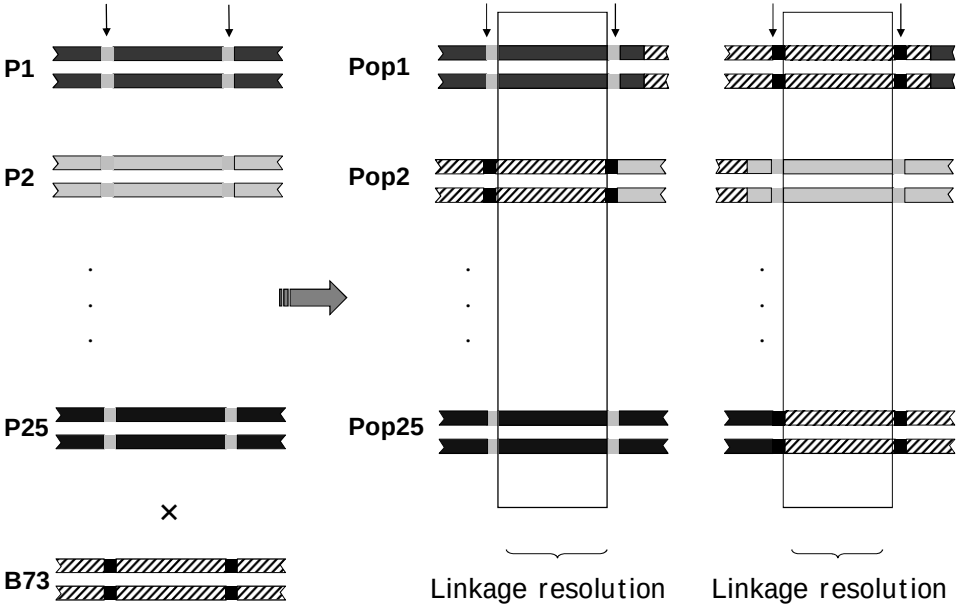
Genotyping B73-rare SNPs to track the recent recombination



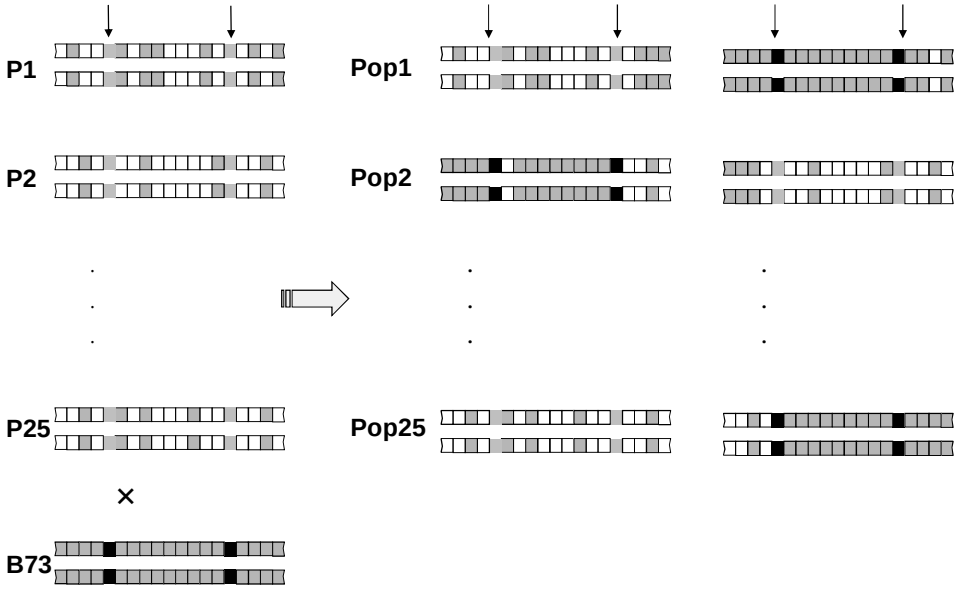
Genotyping B73-rare SNPs to track the recent recombination



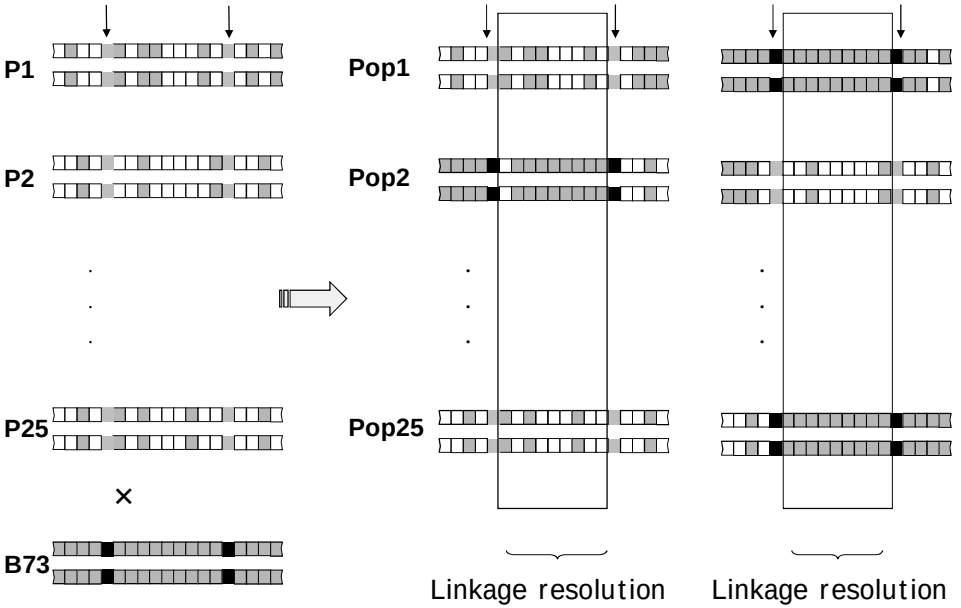
Genotyping B73-rare SNPs to track the recent recombination



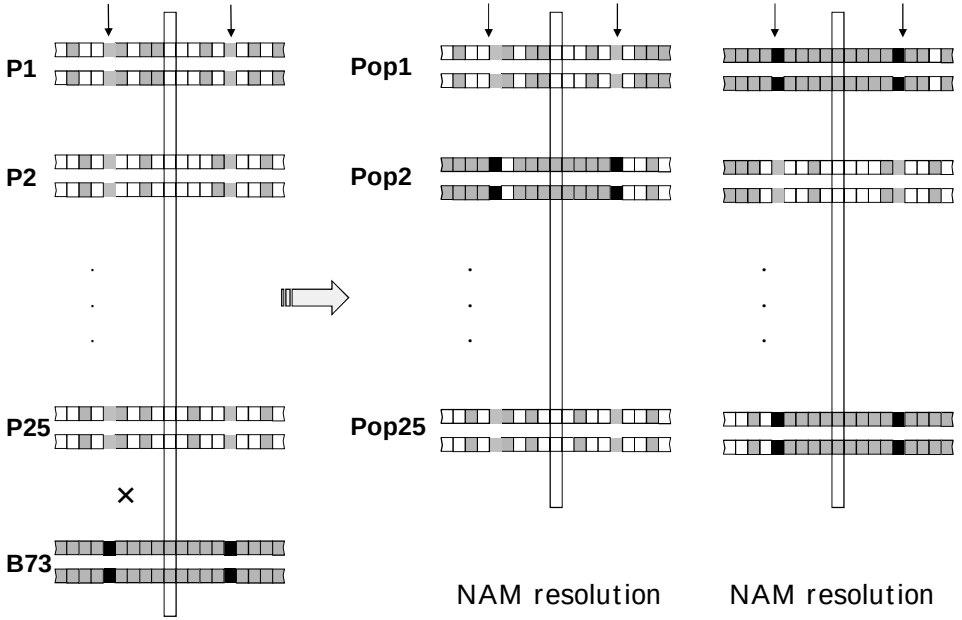
Genotyping parents by sequencing to exploit both recent and ancient recombination



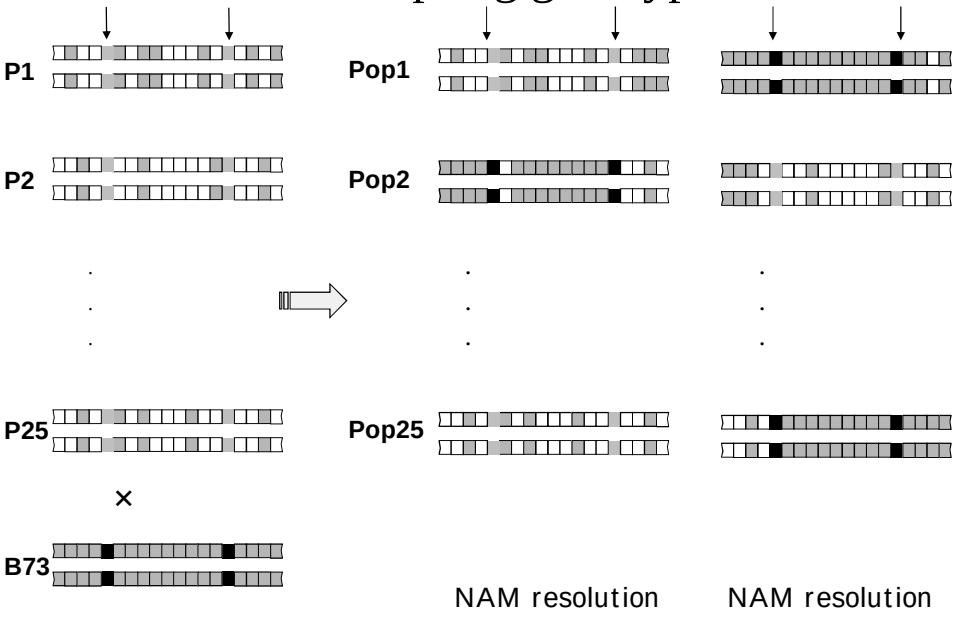
Genotyping parents by sequencing to exploit both recent and ancient recombination



Genotyping parents by sequencing to exploit both recent and ancient recombination



Genotyping parents by sequencing – infer offspring genotypes



Analyses

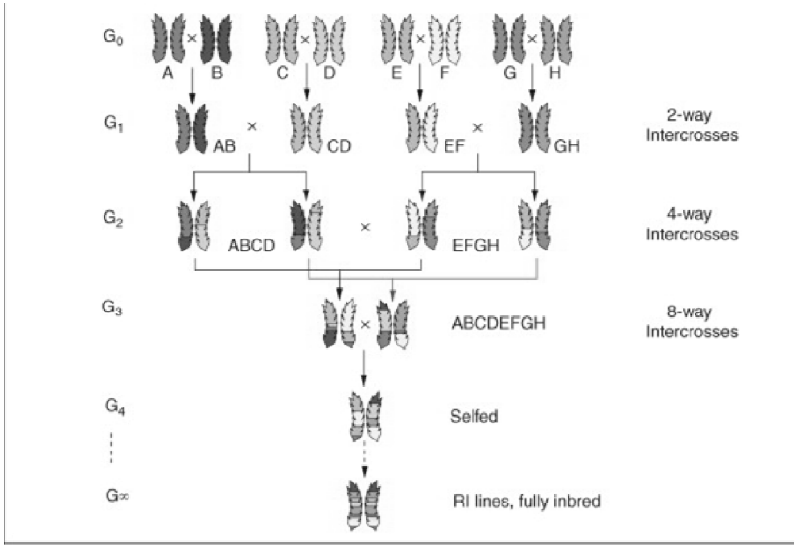
- Within-cross RILs (linkage analysis): standard approaches
  - ANOVA
- Cross-population analyses (including data from multiple crosses), need to take relationships between founders into consideration.
  - The original NAM paper (Yu, et. al. 2007, Genetics 178:539–551) fit a model including the mean value of each population:

$$y = b_0 + a_f m_f + Xb + e$$

  - More complex models possible
    - Blanc, et al. 2006, Theor Appl Genet (2006) 113:206–224: "Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize"

Magic

- Multiparent Advanced Generation Inter-Cross lines



Cavanaugh, et al., Curr Opin Plant Biol 11: 215–221.

Analyses

- Simple single-marker association tests possible
  - ANOVA on genotypes (for SNPs two genos expected)
- Haplotype-based tests
  - Parent-of-origin specific haplotypes (individuals with genotypes identical by state are not necessarily identical by descent)
  - Overview of approach plus extensions:
    - Haplotype Probabilities for Multiple-Strain Recombinant Inbred Lines

      Teuscher and Broman 2007, Genetics 175: 1267–1274.

Analysis tools/resources

TASSEL

- Buckler Lab, Cornell University
  - www.maizegenetics.net/tassel/
- Implements many types of analysis: e.g.
  - Missing phenotype imputation, genotype imputation
  - Principal component analysis
  - Estimation of kinship using genetic markers
  - Association analysis using GLM
- Well developed with a high-end user interface.
- Detailed, well written documentation.
- Tutorials, including video tutorials.

iPlant

- www.iplantcollaborative.org
- Online portal for many genomics tools
  - including for analysis of next-gen data
- Menu-driven user-interfaces
- High-performance compute nodes

Developing Genome Resources for Association Mapping

With a focus on next-gen sequencing

General topics

- Genotyping by sequencing (GBS)
  - RAD-tag sequencing
  - skim sequencing
- Allele/haplotype calling
- *de novo* assembly of genome sequences
  - some background information, not an extensive treatment of the topic
    - library types
    - data format
    - k-mers
- RNA-Seq

What next-gen sequence data looks like

Standard: fastq format

- text file
- four lines in the file for each sequence read
  1. "@" followed by sequence read ID
  2. DNA sequence of the read
  3. "+" (can be followed by a repeat of the read ID)
  4. base call quality scores, coded in ASCII format (one quality score character per nucleotide)

```
@1:1:11253:1209:Y
CTTGGATAATTCAATCATATACTCTAAACAAGGCACTTCACTTGAGATTGCCAGGTTAAAAAGGAGAGAGAA
+
IIIIIF)FFFGGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIFIIIIIIHIIIIIIIIIIIII@IHB
@1:1:11364:1207:Y
TATTCCTCTCCACCACCACCACATTACTCATCTCCTCCACCTCCACCGGTTCACACTTACCCTAAACCAGTC
+
HHHHHF.FFFFFFFFHGHBHGG<GGDEGGGHHHHHHHHHHBHHHGHHHH>B@GGGGGHHHGHHHFHHEFFFEG
@1:1:11631:1206:N
CCGGGCTCAAGTCCCCTGGAAAGGGGCGCCGGAGAGGGTGAGAGCCCCGTTGTGCTCGGACCCTGTCGCACC
+
BBBB@>&@=A,1>78C=B8B=BBB8=8;=BB<6>>==A06C2>;8444?#######################
@1:1:11744:1205:Y
CAGATATCTATCATCTCTGACAGACATGCCATACATTTGCTGCTCAATGTATGGAACTCCCTGTTGCTGTTG
+
DD2DDD&BBD>BC>ADDDDDDDDDDDDDDDDDDDDBDDDDDDDDCDDDDBDBCBBDDDCDD@D@)@@@0;
@1:1:12089:1204:Y
TTTAGAACAAATGTATTGCCACAGGTAACAGCCACGGGAAACATCCACAATGGAATCATTGCAGGAAAGTTG
+
IIIIIF'FFF@@FEFIIIIHIIIIIHIIIIIIIIIIIHHIIIHIHIIIGIFIHFIIIIIIGFIIIGIGGGEB
@1:1:12276:1210:Y
TATAGGGCAGCCTTGCACACTGTCGAACCATATTTGACATAGTCAAAGTCCTTAACTGTGATGGCATACTTT
+
IGIIIG-GGGGGGGGIIGIIHIIHIIIIIIIBIIHIIIGIIIHIIIDIGIIIIHIIHIGGDFIBIHIBFIID
@1:1:12325:1208:Y
TGTTATTCTAATTATTCTCTTTTGATTGGTTCCAAAAACGTGTCTCAAAGTCACAAAGTGTGTTTTTTCTGT
+
HHHHHG5GGGEGGGGHHHHHHHGHBHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHEFHHHHHHHHHHHHHFHH
```

```
@1:1:11253:1209:Y
CTTGGATAATTCAATCATATACTCTAAACAAGGCACTTCACTTGAGATTGCCAGGTTAAAAAGGAGAGAGAA
+
IIIIIF)FFFGGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIFIIIIIIHIIIIIIIIIIIII@IHB
@1:1:11364:1207:Y
TATTCCTCTCCACCACCACCACATTACTCATCTCCTCCACCTCCACCGGTTCACACTTACCCTAAACCAGTC
+
HHHHHF.FFFFFFFFHGHBHGG<GGDEGGGHHHHHHHHHHBHHHGHHHH>B@GGGGGHHHGHHHFHHEFFFEG
@1:1:11631:1206:N
CCGGGCTCAAGTCCCCTGGAAAGGGGCGCCGGAGAGGGTGAGAGCCCCGTTGTGCTCGGACCCTGTCGCACC
+
BBBB@>&@=A,1>78C=B8B=BBB8=8;=BB<6>>==A06C2>;8444?######################
@1:1:11744:1205:Y
CAGATATCTATCATCTCTGACAGACATGCCATACATTTGCTGCTCAATGTATGGAACTCCCTGTTGCTGTTG
+
DD2DDD&BBD>BC>ADDDDDDDDDDDDDDDDDDDDDBDDDDDDDDCDDDDBDBCBBDDDCDD@D@)@@@0;
@1:1:12089:1204:Y
TTTAGAACAAATGTATTGCCACAGGTAACAGCCACGGGAAACATCCACAATGGAATCATTGCAGGAAAGTTG
+
IIIIIF'FFF@@FEFIIIIHIIIIIHIIIIIIIIIIIHHIIIHIHIIIGIFIHFIIIIIGFIIIGIGGGEB
@1:1:12276:1210:Y
TATAGGGCAGCCTTGCACACTGTCGAACCATATTTGACATAGTCAAAGTCCTTAACTGTGATGGCATACTTT
+
IGIIIG-GGGGGGGGIIGIIHIIHIIIIIIIBIIHIIIGIIIHIIIDIGIIIIHIIHIGGDFIBIHIBFIID
@1:1:12325:1208:Y
TGTTATTCTAATTATTCTCTTTTGATTGGTTCCAAAAACGTGTCTCAAAGTCACAAAGTGTGTTTTTTCTGT
+
HHHHHG5GGGEGGGGHHHHHHHGHBHHHHHHHHHHHHHHHHHHHHHHHHHGHHHEFHHHHHHHHHHHHHFHH
```


```
@1:1:11253:1209:Y
CTTGGATAATTCAATCATATACTCTAAACAAGGCACTTCACTTGAGATTGCCAGGTTAAAAAGGAGAGAGAA
+
IIIIIF)FFFGGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIFIIIIIIHIIIIIIIIIIIII@IHB
@1:1:11364:1207:Y
TATTCCTCTCCACCACCACCACATTACTCATCTCCTCCACCTCCACCGGTTCACACTTACCCTAAACCAGTC
+
HHHHHF.FFFFFFFFHGHBHGG<GGDEGGGHHHHHHHHHHBHHHGHHHH>B@GGGGGHHHGHHHFHHEFFFEG
@1:1:11631:1206:N
CCGGGCTCAAGTCCCC
+
BBBB@>&@=A,1>78C
@1:1:11744:1205:
CAGATATCTATCATCT
+
DD2DDD&BBD>BC>AD
@1:1:12089:1204:
TTTAGAACAAATGTAT
+
IIIIIF'FFF@@FEFI
@1:1:12276:1210:Y
TATAGGGCAGCCTTGCACACTGTCGAACCATATTTGACATAGTCAAAGTCCTTAACTGTGATGGCATACTTT
+
IGIIIG-GGGGGGGGIIGIIHIIHIIIIIIIBIIHIIIGIIIHIIIDIGIIIIHIIHIGGDFIBIHIBFIID
@1:1:12325:1208:Y
TGTTATTCTAATTATTCTCTTTTGATTGGTTCCAAAAACGTGTCTCAAAGTCACAAAGTGTGTTTTTTCTGT
+
HHHHHG5GGGEGGGGHHHHHHHGHBHHHHHHHHHHHHHHHHHHHHHHHHHGHHHEFHHHHHHHHHHHHHFHH
```


Each nucleotide call has a numeric quality value.

Coded in ASCII (see next slide).

The higher the value, the higher the quality score for the nucleotide call.

ASCII

- Standardized translation of numeric values to single characters
- provides single "digit" character codes for numeric values ranging from 32 to 126

examples:

numeric value	symbol	numeric value	symbol
33	!	60	<
34	"	61	=
35	#	62	>
36	$	63	?
37	%	64	@
38	&	65	A
39	'	66	B
40	(	67	C
41	)	68	D
42	*	69	E
43	+	70	F
44	,	71	G
45	-	72	H
46	.	73	I
47	/	74	J
48	0	75	K
49	1	76	L

Sanger scores vs. Solexa scores

- Quality scores
- Both use ascii coding
  - different scales
- Lowest Solexa score is ascii character "B"
  - also used for "unknown" quality
- Lowest Sanger score is ascii character "#"
- Many assembly/alignment programs require you to specify the correct quality score type for your input data.

Genotyping by sequencing (GBS)

- The idea: sequence your samples' genomes and compare sequence variation across samples
  - identify variable sites
  - call genotypes at these sites
- Coverage & accuracy vs. cost
  - the deeper the coverage, the more reliable the genotype calls, and the higher the per sample cost.
- The higher the heterozygosity, the lower the accuracy
  - need good coverage to reliably distinguish heterozygotes from sequencing error.

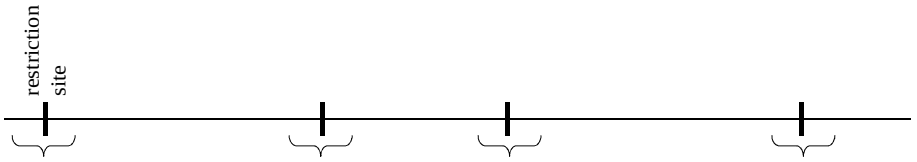
Genotyping by sequencing (GBS)

- Full genome sequencing
  - may be reasonable if the genome is very small or a good reference genome is available.
  - (currently) prohibitively expensive if the genome size is moderate or large and no reference is available.
- Instead of sequencing the entire genome, focus on particular regions
  - e.g. exome
    - exon capture
    - mRNA
  - or random sections of the genome
    - e.g. RAD-tag sequencing

Genotyping by sequencing (GBS)

RAD-tag sequencing

- Focus on high-depth sequencing of a small fraction of the genome:
  - short sections of DNA directly adjacent to specific restriction enzyme recognition sites
- Restriction-site Associated DNA (RAD)



Baird et al, Plos ONE (2009) 3(10): e3376

GBS: single-digest RADseq

General idea:

- Extract genomic DNA, cut with restriction enzyme.
- Ligate adapter to ends
  - (includes Illumina sequencing primer)
- Shear DNA & ligate second adapter
- Amplify fragments that contain adapter bound to restriction site
- Sequence

Baird et al, Plos ONE (2009) 3(10): e3376

GBS: double-digest RADseq

- ddRADseq
- Only one side of restriction site is sequenced
  - reduces redundancy of marker identification; increased overall marker coverage for same amount of sequencing.
- Two restriction enzymes: one common, one rare.
- Size select fragments
  - one end containing rare restriction site, one with common restriction site.
- Sequence from end of fragment with the rare restriction site.

Peterson et al, PLoS ONE (2012) 7(5): e37135

GBS: RADseq downstream analyses

- If a reference genome sequence is available, reads are aligned to the reference.
- If no reference genome is available, assembly-like algorithms are used.
  - e.g. Stacks (creskolab.uoregon.edu/stacks), rtd (github.com/brantp/rtd)
  - These take advantage of the fact that only a small portion of the genome has been sequenced (at high coverage)
  - Sequencing is expected to start at the same nucleotide location for each region of the genome that was targeted.
    - (reads largely overlapping, not tiled)
  - Autopolyploids (no reference), feasibility unclear.

GBS: marker ID & geno calls

- Sites where a sufficient number of aligned or assembled reads contain sequence differences are determined to be polymorphic.
- The proportion of reads containing each allelic sequence determines genotype status:
  - 100% (or close to) indicates a homozygote
  - proportions somewhere around 50% one type/50% the other indicates a heterozygote in a diploid species.
  - For polyploids, various ratios are possible.
    - some methods exist (e.g. Garcia, et al., 2013, Sci. Rep, 3:3399)
    - software underdeveloped
    - pipelines described (e.g. Saintenac, et.al., 2013, G3 3:1105-1114)

GBS: Skim sequencing

- Generally relies on having a reference genome
  - possibly also already known marker sites.
- Sequence genomic DNA
- Align reads to genome
- Marker/genotype calling software

Marker/genotype calling

diploids

- Outbred individuals (heterozygosity expected)
  - see Nature Reviews Genetics (2011) 12:443-451.
- Inbred lines (low heterozygosity)
  - JGIL (Stone, Genome Res. 2012 May;22(5):966-74)

polyploids

- tools underdeveloped.
- allopolyploids:
  - use of tools designed for diploids is feasible
- complex genomes: high coverage essential to determine different allelic combinations

Haplotyping (post-genotyping)

- Diploids (examples, not a complete list)
  - Phase, fastPhase (stephenslab.uchicago.edu/software.html)
  - Beagle (faculty.washington.edu/browning/beagle/beagle.html)
  - HapSeq2 (www.ssg.uab.edu/hapseq)
  - MACH (www.sph.umich.edu/csg/abecasis/MACH)
  - Impute2 (mathgen.stats.ox.ac.uk/impute/impute_v2.html)
- Polyploids
  - PolyHap (Su, et al., 2008, *BMC Bioinformatics 2008, 9:513)*
    - designed for use with SNP genotype data, where the marker order is known.

Creating a reference genome

- *De-novo* genome assembly
- Fractionate genomic DNA
- Sequence DNA fragments
- Assemble shorter sequences into larger contigs in silico

- "Next-gen" sequencing
  - Short(ish) read lengths
  - Tens of millions reads per sample
- Different library types (hence data types) possible

Library types: Single end*

- Fractionate DNA
- Sequence one end of fragments

- Result:
  - Data file containing millions of DNA sequences
  - each sequence generated independently of the others.

*Single-end data can be generated from paired-end libraries (next slide).

Library types: Paired end

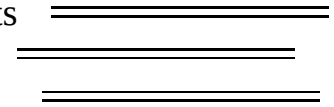
- Size select DNA fragments (e.g. ~400 base pairs)
- Sequence both ends of a fragment

- Result:
  - Two output files, paired.
  - Each file contains sequences for one end of a fragement

- Insert size:
  - Ave. number of nucleotides between sequenced ends
  - If DNA fragment size is ~400nt and sequenced 100nt on each end, the insert size is ~200nt.

Library types: Mate pair

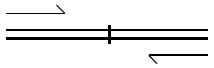
- Size-select longer DNA fragments
  - e.g. 2kb, 10kb, 20kb



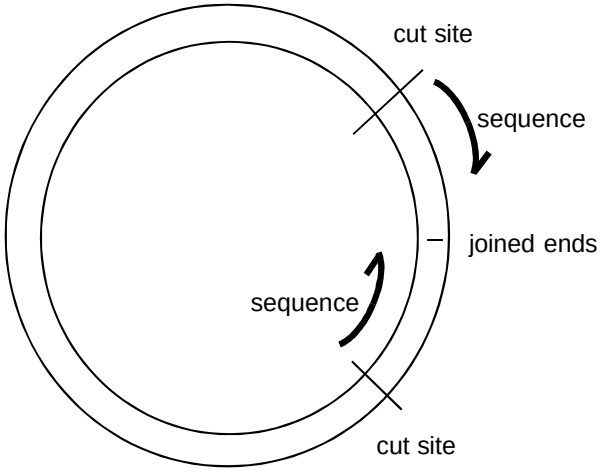
- Circularize fragments

- Cut surrounding joined ends

- Sequence ends of resulting fragment

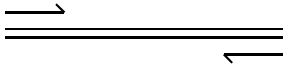


Library type: Mate pair

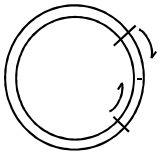


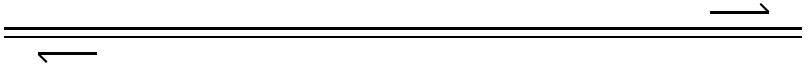
Paired end vs. Mate pair

- Paired ends:
  - Sequence fragments inward



- Mate pair:
  - because of circularization, direction of sequence pairs is outward





Library types: Fosmid

- Long insert sizes (40 kb)
- Mate-pair cloning
  - Circularization
- (Data as standard mate-pair with longer insert sizes.)

Paired end/mate pair files

- Usual format is two files for each sample, one file for one end of the reads, one file for the other:
  - name_of_file_1.fastq
  - name_of_file_2.fastq

```
@FCD23ETACXX:2:1101:1653:2088#ATGTCAGA/1
GTTTATGAATATTAAAACTCGAAAATTACAAACAAATTACTTAATAGTTT
+
aaaeeeeeggggggghhiiiiihhiggghiiiihiiiiighghiiiiihhi
@FCD23ETACXX:2:1101:1812:2099#ATGTCAGA/1
ATACATATGAAATTGAAGAGATTAATTAAATAATTAATTTTTTCCCAGTC
+
bbbeeeeeggggiiiiihihhiigiiihiiiiiiiiiiiiiiiihi
```

⇦ _1 file

_2 file ⇨

```
@FCD23ETACXX:2:1101:1653:2088#ATGTCAGA/2
ACAAGAAAAAAAGAAAATTTGTTAAAATTAACTTTATAGTAGATTTATTT
+
bbbeeeeeggggghiiiihiiiiiiiiiiiiiighhgffeehhhcghi
@FCD23ETACXX:2:1101:1812:2099#ATGTCAGA/2
ATTAAAAATAAATAAATTAAAAGAAGAAAATAATAAAATGAAAGAAGAAT
+
bbbeeeeeggggiiiiiiiiiiiiiihfhiiiiiiiihiihhiiii
```

K-mers

- Useful to understand when working with next-gen sequence data for assembly (genome and transcript)
- Used by *de novo* assembly programs
  - often length of k-mer is specified by user
- Also used for cleaning data
  - sequencing errors
  - determining sequencing coverage
- Assessing data
  - e.g. determining heterozygosity

K-mers

- all sequences of nucleotides that are k characters long found in the sequence read data file
- for example, the sequence read

  GTTTATGAATATTAAAACTCGA
- contains the following 15-mers (k=15)

GTTTATGAATATTAA
TTTATGAATATTAAA
TTATGAATATTAAAA
TATGAATATTAAAAC
ATGAATATTAAAACT
TGAATATTAAAACTC
GAATATTAAAACTCG
AATATTAAAACTCGA

Tracking K-mer counts

- Keep track of how often a given k-mer is observed in the data (k-mer counts or k-mer frequencies)
- For uniform sequencing coverage, expect to see each k-mers at similar frequencies.
- Rare k-mers are an indication that the reads that contain them contain sequencing errors.
- Can also identify heterozygote SNPs
  - k-mers containing alternative alleles are at ~50% frequencies (relative to the mean or mode of frequencies of all k-mers)

de novo assembly with short sequence reads

Many software packages exist. Examples:

- Soap de novo
  - transcriptome: SOAPdenovo-Trans
- Velvet
  - transcriptome: Oasis
- All-paths
- ABySS
  - transcriptome: trans-ABySS

Caveats

A number of factors increase the difficulty of creating a correct assembly

- High heterozygosity
- Repetitive regions
- Genome duplications
- Polyploidy

- If possible, use an accession that is diploid and inbred (low heterozygosity) to create the reference
  - Can then use this to aid genomics/transcriptomics of more complex accessions/species

Integrating expression data into association mapping

focus on RNA-Seq data

Gene expression

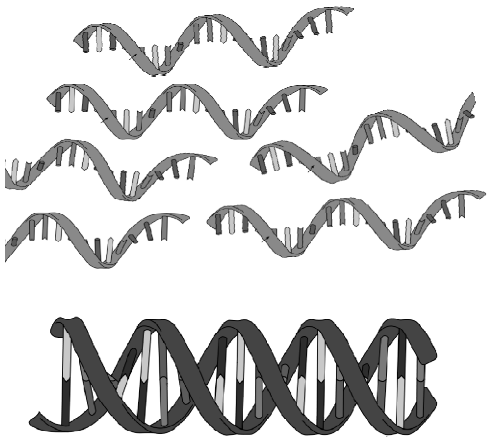
- Measured through transcript (mRNA) abundance

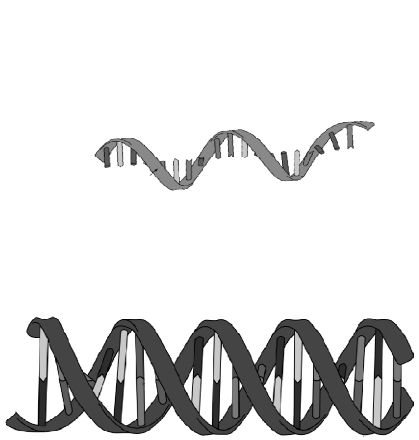
Condition A

Condition B

Gene expression

- Measured through transcript (mRNA) abundance





Gene expression: RNA-Seq

- Collect biological sample
- extract mRNA
- ultra-high throughput sequencing
  - each mRNA molecule sampled for sequencing produces one sequence read
- if a gene was highly expressed in the sample
  - transcript abundance is high
- many sequence reads will be generated for that gene (relative to other genes)

Sequence reads

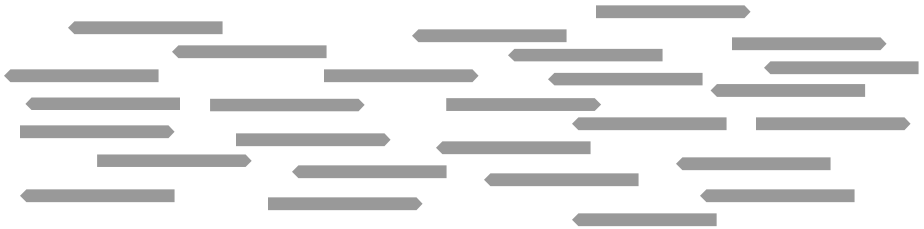
Sequence reads

```
GTTAAGGCTGCCATCAAGGACAGGGTTGTCAATGTTGCTCAAGTTACCAGCAACACACTCGCTTT
CAACAAGAGAAAACAAGGTGCAAGTATTGCCTTGGAACTGGTTACTTGGCTTGCGCTCGGTGTTC
CGGGAAACCAAATCAAGAAGCAGGCAATCCTTAGGATTGCTTTTCGTGGGTAGAGCGAGGGGTTT
ATTTTTCAGTCTTCTCTCGTGGCATTTATTGTCGGTTGGTTTTCTATATATTGCTCGTGCAACTC
CGTCCCTACCATATCTCATCATCATTATCAATAATATAAGAAACATAATTATCATAATAGAGGAA
CTCTTGCCGGCATTGTGGGCAAAGAGAGAATTGTTGTGTCCACTTCTTGCTCACTTCTTCACACT
TGTCATAATAACACTCTCTGCTGGTAGAGGTGCAGAATGCTGTAACATACCATCCCCTTCTTTTA
AAAATATATTTCTGGGGATCAATTGACAAAGGATGATATCAAAGTGTACGGATATGTTTCTGAGA
```

```
@1:1:11253:1209:Y
CTTGGATAATTCAATCATATACTCTAAACAAGGCACTTCACTTGAGATTGCCAGGTTAAAAAGGAGAGAGAA
+
IIIIIF)FFFGGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIFIIIIIIHIIIIIIIIIIIII@IHB
@1:1:11364:1207:Y
TATTCCTCTCCACCACCACCACATTACTCATCTCCTCCACCTCCACCGGTTCACACTTACCCTAAACCAGTC
+
HHHHHF.FFFFFFFFHGHBHGG<GGDEGGGHHHHHHHHHHBHHHGHHHH>B@GGGGGHHHGHHHFHHEFFFEG
@1:1:11631:1206:N
CCGGGCTCAAGTCCCCTGGAAAGGGGCGCCGGAGAGGGTGAGAGCCCCGTTGTGCTCGGACCCTGTCGCACC
+
BBBB@>&@=A,1>78C=B8B=BBB8=8;=BB<6>>==A06C2>;8444?########################
@1:1:11744:1205:Y
CAGATATCTATCATCTCTGACAGACATGCCATACATTTGCTGCTCAATGTATGGAACTCCCTGTTGCTGTTG
+
DD2DDD&BBD>BC>ADDDDDDDDDDDDDDDDDDDDDBDDDDDDDDCDDDDBDBCBBDDDCDD@D@)@@@0;
@1:1:12089:1204:Y
TTTAGAACAAATGTATTGCCACAGGTAACAGCCACGGGAAACATCCACAATGGAATCATTGCAGGAAAGTTG
+
IIIIIF'FFF@@FEFIIIIHIIIIIHIIIIIIIIIIIHHIIIHIHIIIGIFIHFIIIIIIGFIIIGIGGGEB
@1:1:12276:1210:Y
TATAGGGCAGCCTTGCACACTGTCGAACCATATTTGACATAGTCAAAGTCCTTAACTGTGATGGCATACTTT
+
IGIIIG-GGGGGGGGIIGIIHIIHIIIIIIIIBIIHIIIGIIIHIIIDIGIIIIHIIHIGGDFIBIHIBFIID
@1:1:12325:1208:Y
TGTTATTCTAATTATTCTCTTTTGATTGGTTCCAAAAACGTGTCTCAAAGTCACAAAGTGTGTTTTTTCTGT
+
HHHHHG5GGGEGGGGHHHHHHHHGHBHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHHEFHHHHHHHHHHHHHFHH
```

Millions of short sequence reads

Millions of short sequence reads



AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

Align short reads to genome

AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

If no reference genome

- *de novo* assembly of transcriptome
  - Can be advantageous to collect tissue from multiple organs, developmental time points
    - better representation of full transcriptome.
  - Use this assembly to align RNA-Seq reads to.
- Use reference genome of a closely related organism.

Align short reads to genome

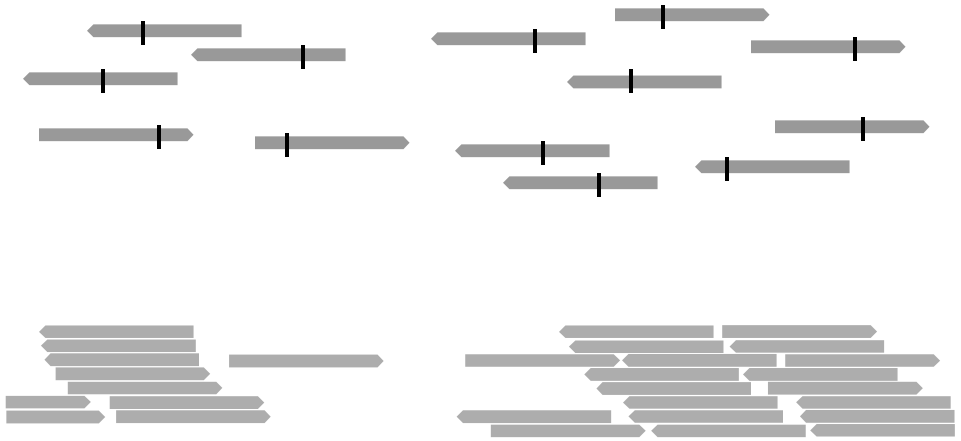
AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

Reads that don't align in first pass …

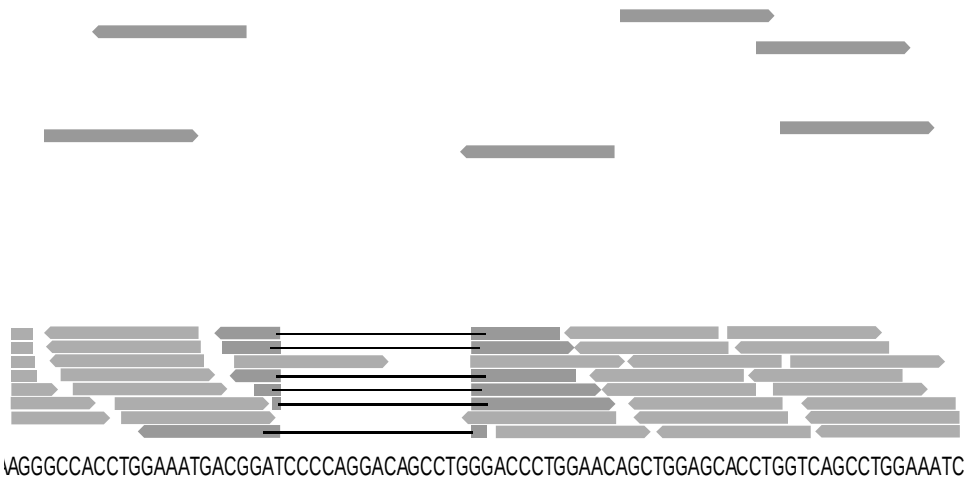
genome sequence

Break into pieces



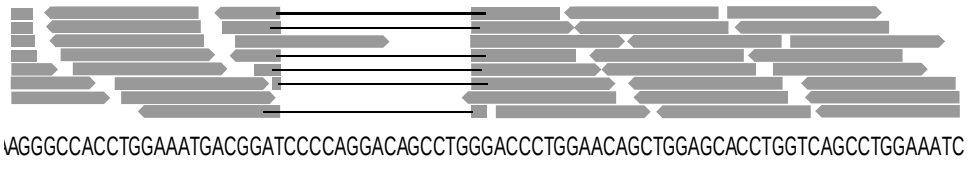
genome sequence

Align allowing for gaps: introns



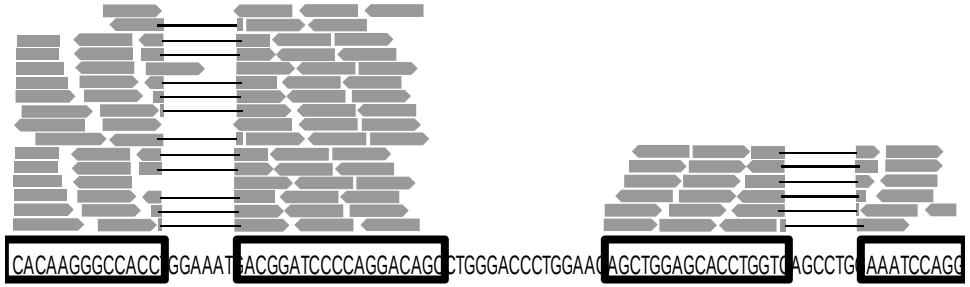
genome sequence

Use alignments to determine which genes contributed which sequenced transcripts

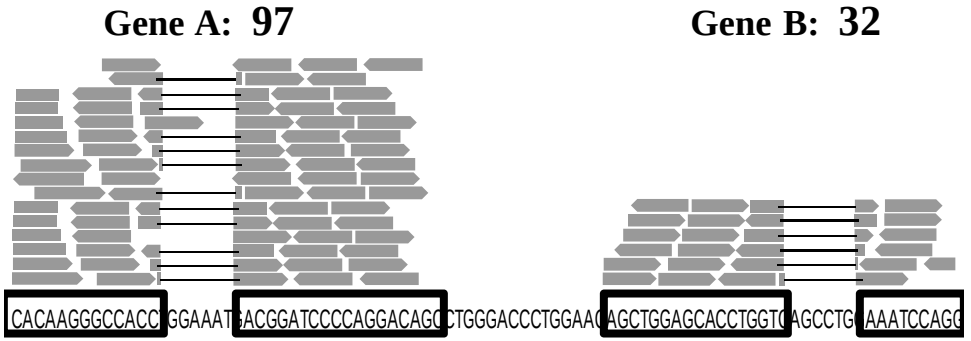


genome sequence

Use alignments to determine which genes contributed which sequenced transcripts

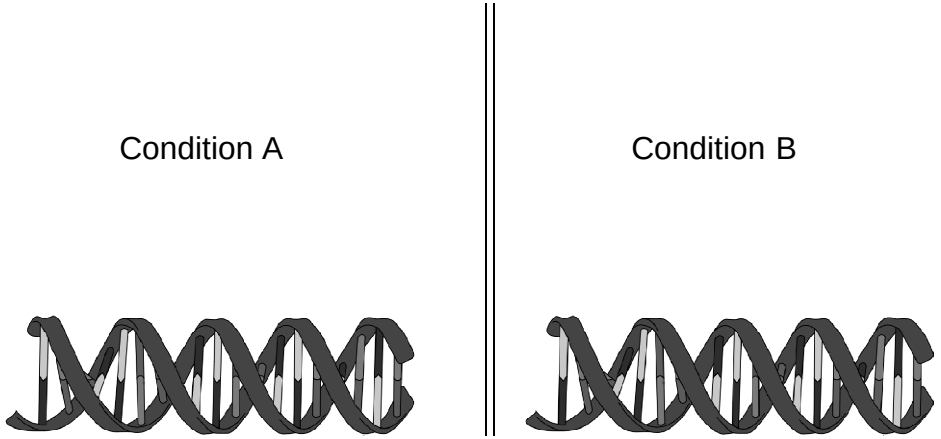


And for quantification of gene expression (counts of reads per gene)

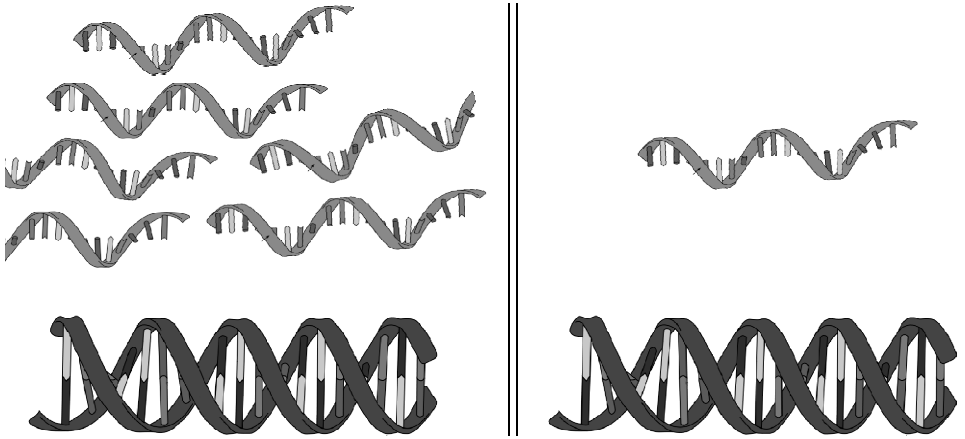


Differences in gene expression

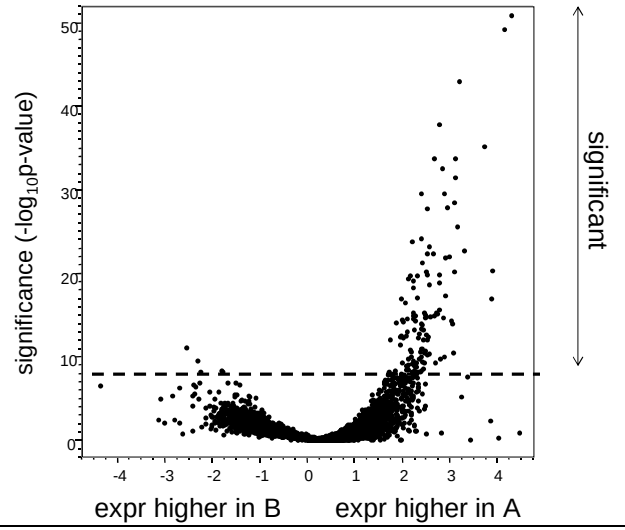
- Measured through transcript (mRNA) abundance



Differences in gene expression



Volcano plot: summary of results



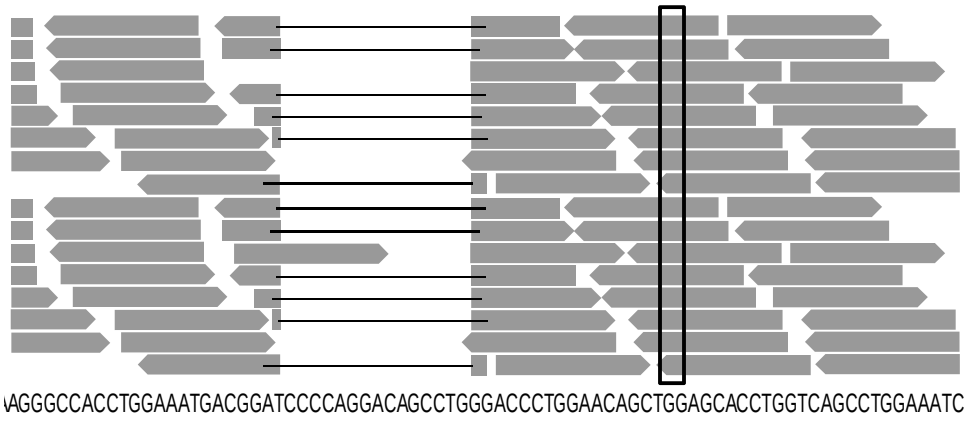
Integrated data

- Gene mapping to identify loci associated with a phenotype
- Expression analysis on individuals expressing different phenotypes
- Do any of the genes within the genomic regions identified by gene mapping show differential expression across phenotypes?

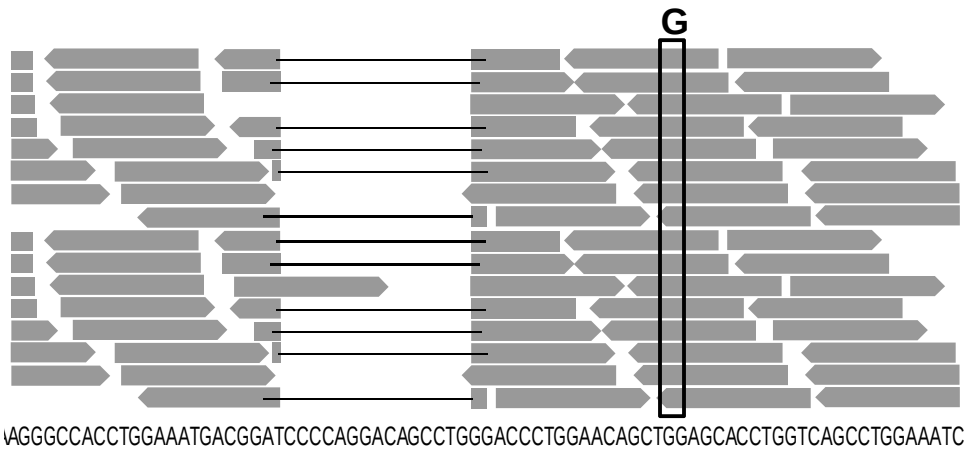
Marker detection & genotyping

- RNA-Seq data is sequence data that was derived from the genome
  - expressed regions
- With sequence data can identify polymorphic sites
- Genotype individuals

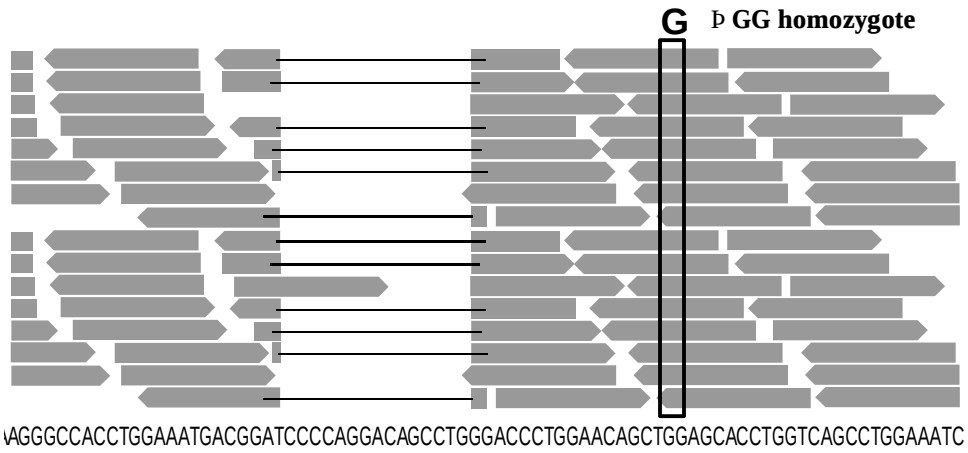
Identifying sequence polymorphisms



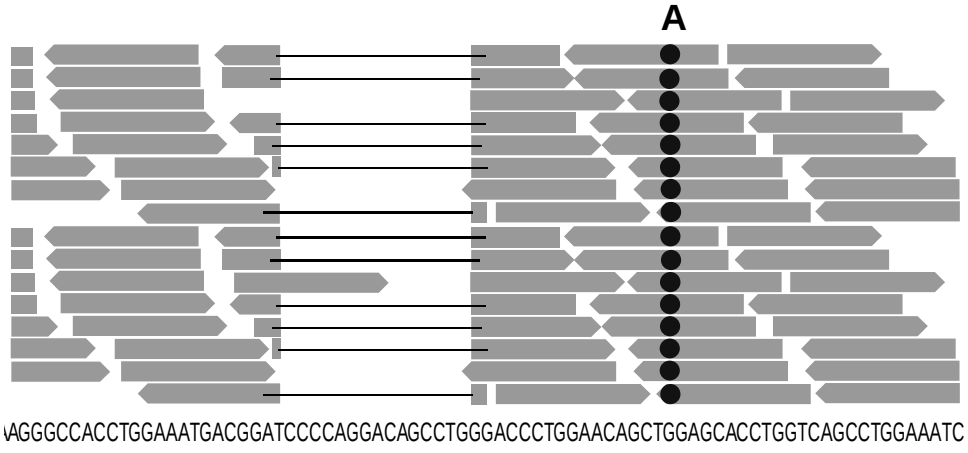
Identifying sequence polymorphisms



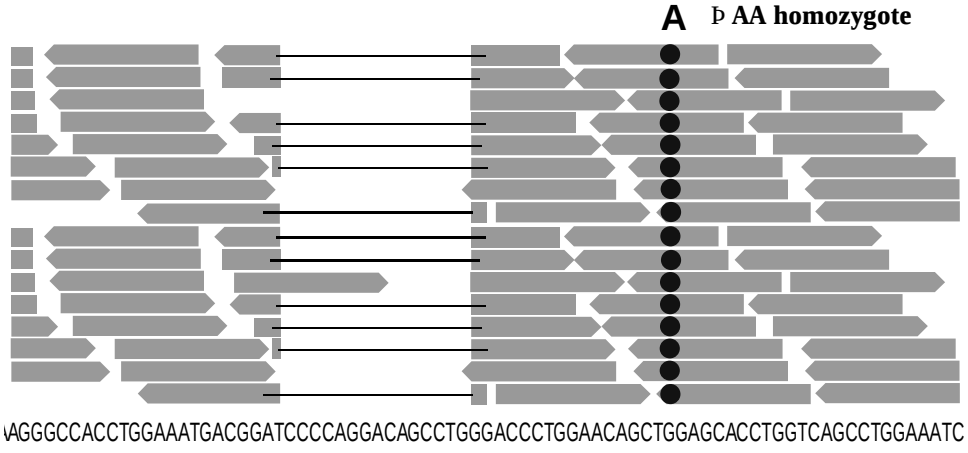
Identifying sequence polymorphisms



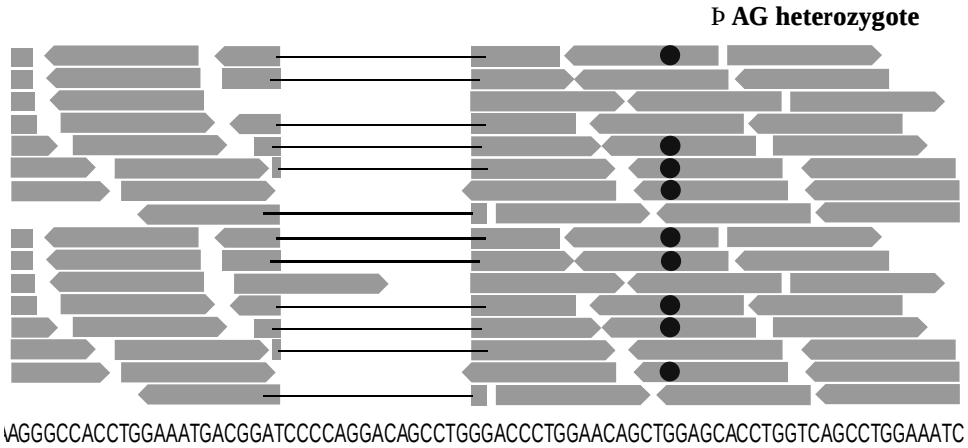
Identifying sequence polymorphisms



Identifying sequence polymorphisms



Heterozygote (~50% of each allele)
(diploids)



Polyploids

- Diploids: expect proportions of 100%, 50/50%, or 0% (for a given allele)
- Polyploids: expect more complex proportions.
- High coverage (highly expressed genes or deeper sequencing) needed for accurate determination of proportions
  - lower coverage possible for presence/absence calls.

Marker identification & genotyping

- Same principles as with genomic sequencing
- Software examples as before.

Allele-specific expression

- Biologically, one allelic variant of the gene is expressed at higher levels than the other.
- Will cause distortions of the expected proportions for homozygous/heterozygous calls.
  - potentially causes errors in genotype calls

Information from RNA-Seq data

Can provide both:

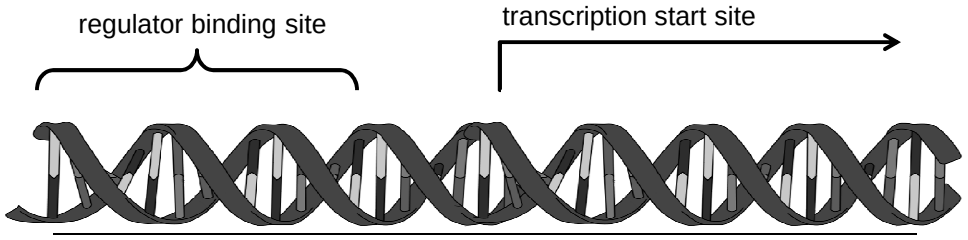
- information on gene expression levels across different conditions,
- information on the presence of sequence variation of the individual
  - (within expressed regions of the genome)

Genetic basis of gene regulation

expression QTL (eQTL) mapping

eQTL mapping

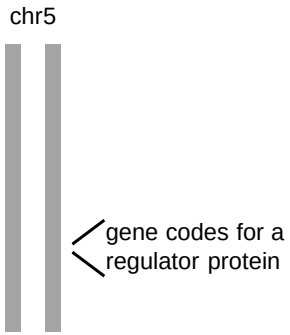
- Can identify genes containing polymorphisms that cause changes in expression for that gene
  - cis-eQTL



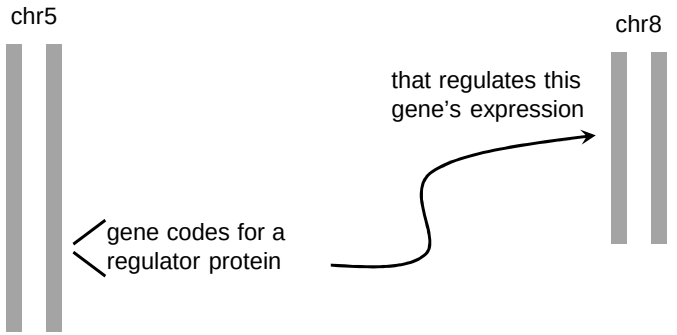
eQTL mapping

- Can identify polymorphisms in the genome correlated with expression changes at distal sites:
- Identify the location of regulator genes in the genome
  - trans-eQTL

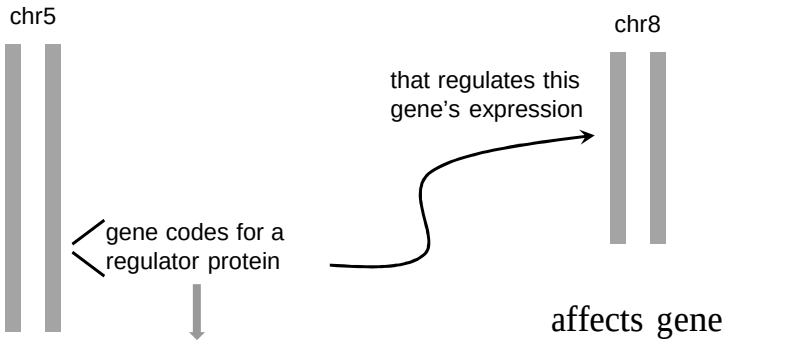
Location of regulator genes



Location of regulator genes



Location of regulator genes

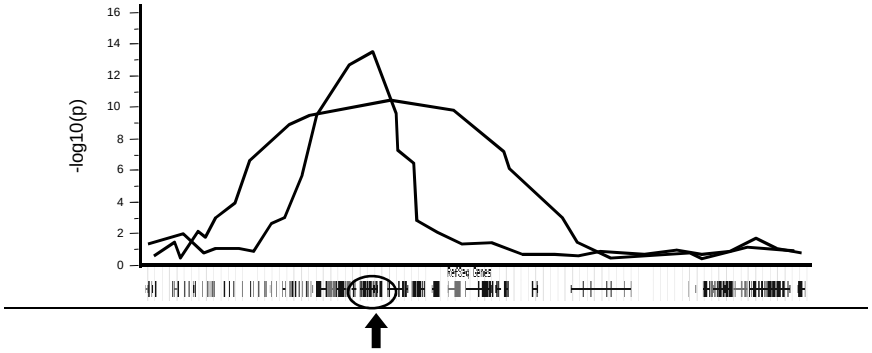


eQTLs & trait mapping

- Of interest may be to identify loci associated with a phenotype of interest …
  - standard trait gene mapping
- … co-localized with loci associated with gene expression changes
- Possible detection of candidate genes:
  - genes within the genomic region associated with the trait, and with expression correlated with the phenotype.

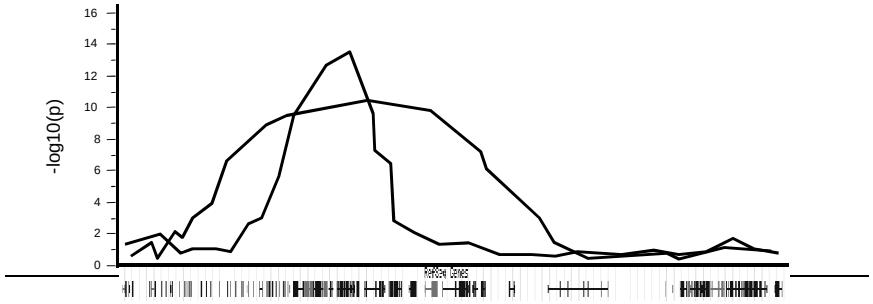
Example

- blue: QTL for tolerance
- red: eQTL for the gene indicated by the arrow.
  - possible candidate gene.



Better dissection of genetic architecture

- Co-localization of QTL and eQTL peaks
- eQTL associated with a distal gene's expression (trans-eQTL)
  - Potential action of QTL: expression (mis)regulation



Heritability

Broad sense

Narrow sense

"Missing"

Misconception

- Heritability describes the degree to which a trait is genetic in nature
  - Incorrect.
- Genes may have a profound effect on a trait, and yet heritability for this trait can still be low.

Misconception

- Heritability describes the degree to which a trait is genetic in nature
  - Incorrect.
- Genes may have a profound effect on a trait, and yet heritability for this trait can still be low.

- What heritability does tell you:

  How much of the **variation** seen in the trait can be explained by genetic variation in the population.
  - do individuals look different from one another because they have different genotypes?

Heritability

- Do individuals look different from one another because they have different genotypes?
  - or because they were/are exposed to different environmental conditions?
- How much of the **variance** in phenotype can be explained by genetic variation between individuals?

Example

- You observe a population of birds with blue plumage
  - The blue color in the feathers comes from a genetically encoded pigment molecule.
- Is feather color genetic?
- Is heritability high or low?

Example

- You observe a population of birds with blue plumage
  - The blue color in the feathers comes from a genetically encoded pigment molecule.
- Is feather color genetic?
- Is heritability high or low?
  - If all birds in the population have the same color feathers, there is no variation in the phenotype and heritability is not defined.

Example

- You observe a population of birds with blue plumage
  - The blue color in the feathers comes from a genetically encoded pigment molecule.
- Is feather color genetic?
- Is heritability high or low?
  - What if there is sequence variation in the population within the gene that encodes the pigment molecule …
  - different variants cause some birds to be darker blue and others to be lighter blue? Or knockouts exist that make some birds white?

Example

- You observe a population of birds with blue plumage
  - The blue color in the feathers comes from a genetically encoded pigment molecule.
- Is feather color genetic?
- Is heritability high or low?
  - There is a fine white ash produced in the location th where the population of birds lives; when birds take “ash baths”, it discourages mites. How would this information affect your answers?

Example

- Clearly, feather color is genetic: caused by a locus that codes for a pigment molecule.
- ***Variation*** in color (individuals looking different from one another) can be caused by
  - allelic variation among individuals,
  - or by the amount of ash currently appearing on different individuals' feathers,
  - or by a combination of the two.

Feather color is genetic

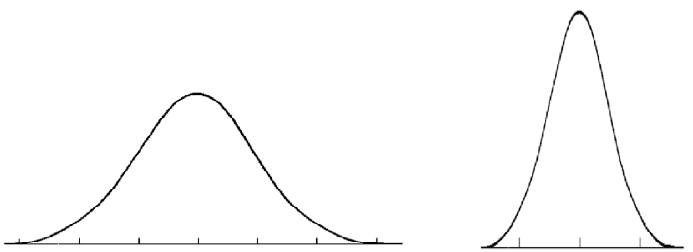
- However,
  - whether a population is segregating for relevant polymorphisms,
  - what the frequencies of these polymorphisms are,
  - how the birds behave around ash,
  - and what the availability of ash is for different individuals
- all contribute to the heritability of the trait
- in any given population.

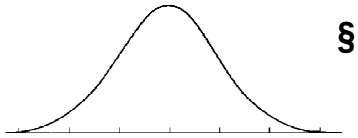
Heritability

- The degree to which *variation* in phenotype is caused by genetic differences between individuals in a population.
  - heritability of a trait varies from population to population.
    - different allele frequencies.
    - different environmental conditions.
  - heritability of a trait within a population varies over time.
    - changing allele frequencies.
    - changing environmental conditions.

Measuring variation in phenotype

- Measure variation in phenotype using ***variance***
  - assumes numeric quantification of phenotype values
- Measures how far observations are expected to be spread from the mean.





§ Variance of a phenotype that follows ← this distribution

§ What happens to overall variance if males and females differ slightly?

§ And to overall variance if they differ even more?

§ Note that it's not enough just to have both males and females in the population – they need to differ.

§ And note that there's no effect on variance if males and females differ, but there are only males or only females in the population.

Effect of genotypes

§ If there are multiple genotypes present in a population

§ and different genotypes affect phenotype differently

§ Variance will be increased relative to a monomorphic population

– or relative to what the population would be like if all genotypes have exactly the same effect on phenotype.

§ The question is: how much of the variance in phenotype can be explained by the presence of different genotypes in the population?

Variance in phenotype

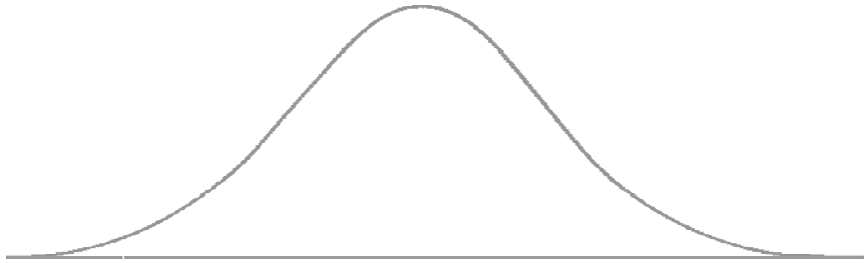
§ How much of the variance in phenotype can be explained by the fact that individuals in the population have different genotypes?

– Quantify *how* different individuals are because they have different genotypes.

§ Total phenotype variance: s^2_P

§ Phenotype variance that can be explained by individuals having different genotypes: s^2_G

Total phenotype variance, σ^2_P



Total phenotypic variance, σ^2_P

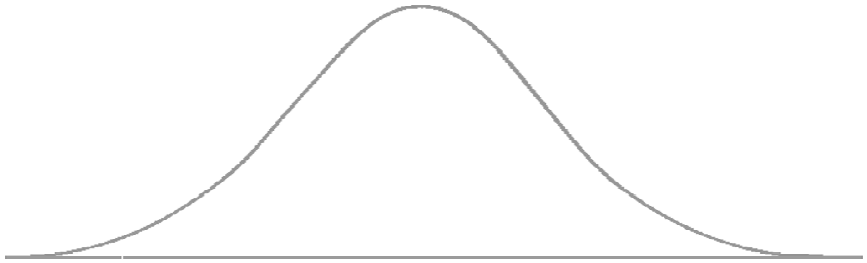
- $s^2_P = E[(X - \mu)^2]$
  - **X** represents what individual instances of phenotypes could look like.
  - **X -** μ is the difference between an instance of X (an individual pheno) and the population mean.
  - **E [ ]** is the expected value: the theoretical /population mean.
    - $\mu = E[X]$ (μ is the mean population phenotype)

Total phenotypic variance, σ^2_P

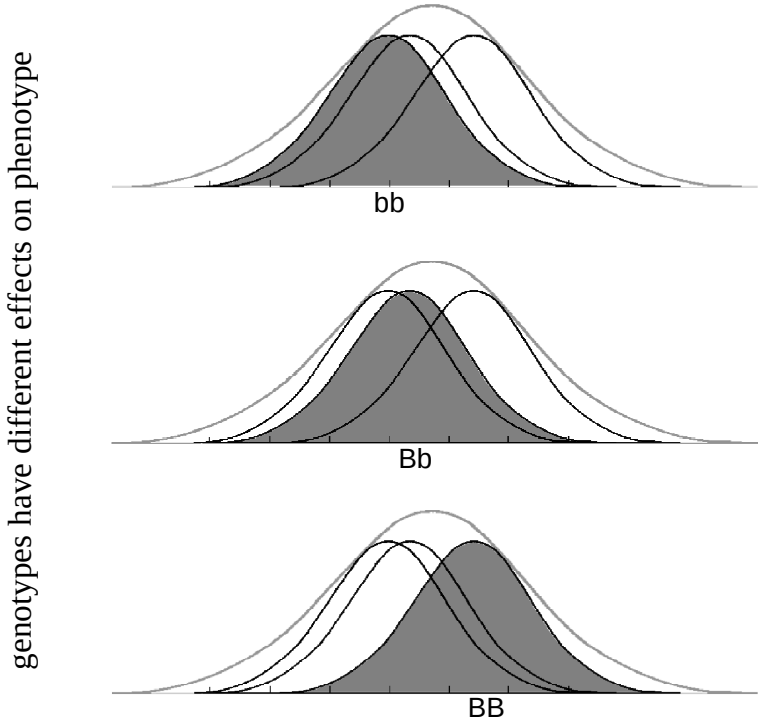
- $s^2_P = E[(X - \mu)^2]$
  - **X** represents what individual instances of phenotypes could look like.
  - **X -** μ is the difference between an instance of X (an individual pheno) and the population mean.
  - **E [ ]** is the expected value: the theoretical /population mean.
    - $\mu = E[X]$ (μ is the mean population phenotype)

- If $(X - \mu)^2$ is, <u>on average</u>, large, then variance is high.
  - Instances of X (individual phenos) tend to be far from the mean.
- If $(X - \mu)^2$ is, <u>on average</u>, small, then variance is low.
  - Instances of X (individual phenos) tend to be close to the mean.

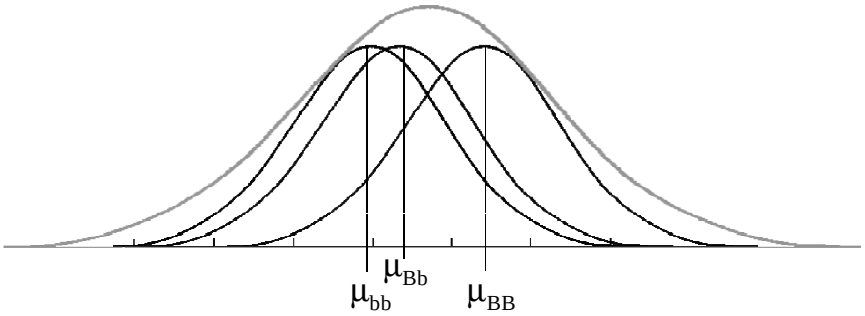
Consider a single locus affecting phenotype



- Single locus with two alleles: B, b
  - genotypes: BB, Bb, bb
  - genotypes have different effects on phenotype



Genos have different effects on pheno



- μ_{bb} μ_{Bb} μ_{BB}
- Genotype value: mean phenotype for all individuals with that genotype

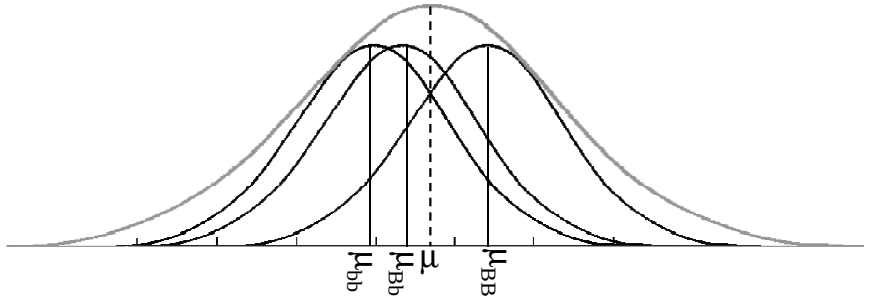
Variance in phenotype caused by genetic variation: σ^2_G

- Variance = $E[(X - \mu)^2]$
  - μ is still the mean population phenotype
- Key question: when defining σ^2_G, what is X ?

Variance in phenotype caused by genetic variation: σ^2_G

- Variance = $E[(X - \mu)^2]$
  - μ is the mean population phenotype
- Key question: when defining σ^2_G, what is X ?
  - mean pheno for all individuals with a given genotype
  - μ_{bb} μ_{Bb} μ_{BB}
  - genotype values
- s^2_G = E[(difference between these genotype values and the population mean)^2]

Genos have different effects on pheno



- s^2_G = E[(difference between the genotype values and the population mean)^2]
  - genotype values: μ_{bb} μ_{Bb} μ_{BB}
    (mean phenotype for all individuals with that genotype)

Variance in phenotype caused by genetic variation: σ^2_G

- s^2_G = E[(difference between the genotype values and the population mean)^2]

- s^2_G = E [$(X - \mu)^2$] =

  $p_{BB} (\mu_{BB} - \mu)^2 + 2p_{Bb} (\mu_{Bb} - \mu)^2 + p_{bb} (\mu_{bb} - \mu)^2$

Variance in phenotype caused by genetic variation: σ^2_G

- s^2_G = E[(difference between the genotype values and the population mean)^2]

- s^2_G = E [$(X - \mu)^2$] =

  $p_{BB} (\mu_{BB} - \mu)^2 + 2p_{Bb} (\mu_{Bb} - \mu)^2 + p_{bb} (\mu_{bb} - \mu)^2$

  - E [ ] is a mean
  - means depend on how often any given value appears (its frequency)
    - if 95% of a population is 12' tall, and 5% is 8' tall, the mean will be close to (but not quite) 12'.

When σ^2_G is small

- s^2_G =

  $p_{BB} (\mu_{BB} - \mu)^2 + 2p_{Bb} (\mu_{Bb} - \mu)^2 + p_{bb} (\mu_{bb} - \mu)^2$

- Small when differences between genotype effects and overall means are small
  - or aren't small, but the genotype frequencies *are* small.
- A genotype that produces a very different phenotype than the mean will cause an increase in variance due to genotype
  - but the effect will be negligible if that genotype is rare.

It's all relative!

- A genotype that produces a very different phenotype than the mean will cause an increase in variance due to genotype
  - but the effect will be negligible if that genotype is rare.

- However, if this is the *only* factor that contributes to variation of the phenotype, then it could still be important!

Broad sense heritability, H^2

- σ^2_P : variance in phenotype
  - measures how different individuals are from one another …
  - for any reason.
- σ^2_G : variance in phenotype
  - measures how different individuals are from one another …
  - *because* they have different genotypes.
- $\sigma^2_G \leq \sigma^2_P$
  - because σ^2_G is probably only one part of σ^2_P.

$$H^2 = \sigma^2_G / \sigma^2_P$$

Broad sense heritability, H^2

- Of all the phenotypic variance in the population, what proportion of it is because individuals have different genotypes.

Narrow sense heritability, h^2

- To understand h^2, you need to understand additive genetic variance
  - σ^2_A ("A" for additive)
  - phenotype variance that can be inherited from one generation to the next.
    - Can be predicted in offspring based on parents.
- To understand σ^2_A, it's helpful to understand additive effects of alleles
  - α_i ("i" for allele i)

Additive effects of alleles: α_i

- α_i =

  E [phenotype for all individuals who carry allele i] – μ
- Consider a locus with two alleles, B and b
  - Three genotypes: BB, Bb, bb

  - α_B = E [pheno for all BB and Bb indivs] – μ
    - (BB and Bb individuals as a single group)
  - α_b = E [pheno for all Bb and bb indivs] – μ

Additive effects of alleles

§ α_B =

E [phenotype for all BB and Bb indivs] – μ .

– How much do individuals who carry at least one B allele differ from the overall population mean?

§ $\alpha_B = p_B \mu_{BB} + p_b \mu_{Bb} - \mu$.

– depends on allele frequencies

– If the b allele is rare, then there will be a lot more BB individuals than Bb

§ which affects the mean phenotype of the group of BB and Bb individuals.

Genotype values & additive effects

§ What if all you knew about a phenotype were the additive effects of alleles?

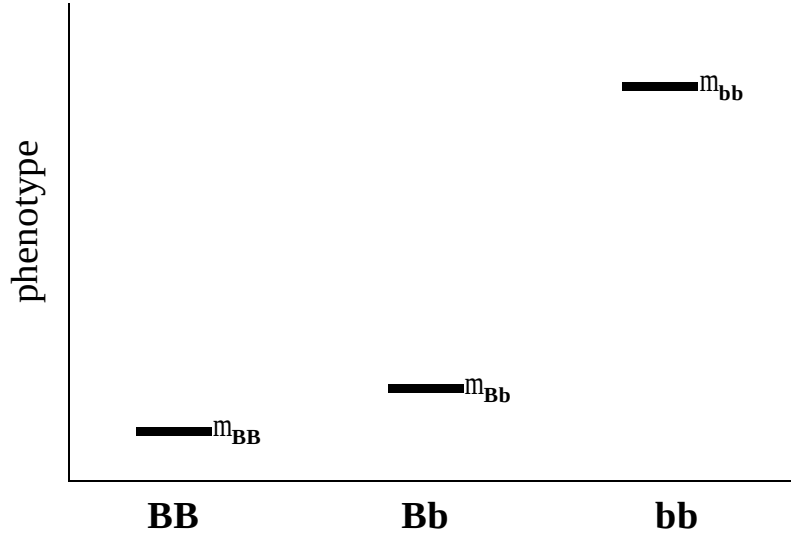
– (for all alleles affecting the trait; let's assume one locus, two alleles for now)

§ Could you describe the genotype values in terms of the additive effects of alleles?

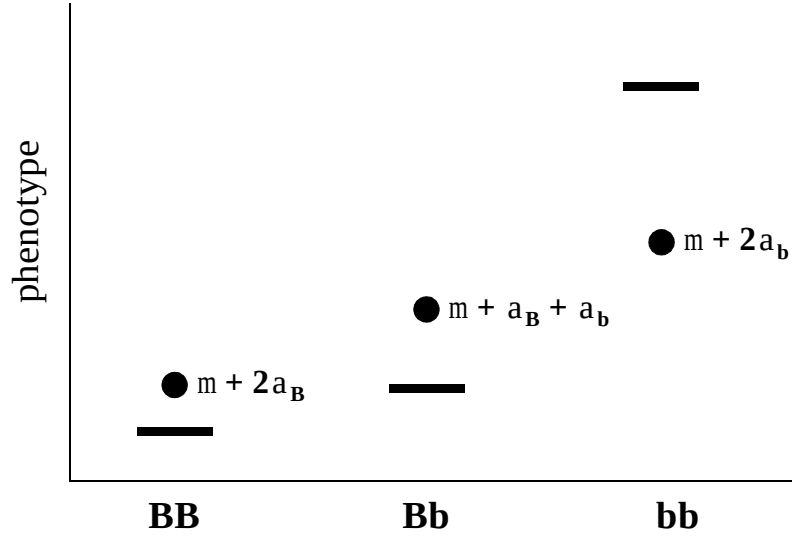
- $\mu_{BB} = \mu + \alpha_B + \alpha_B = \mu + 2\alpha_B$
- $\mu_{Bb} = \mu + \alpha_B + \alpha_b$
- $\mu_{bb} = \mu + \alpha_b + \alpha_b = \mu + 2\alpha_b$

§ Is this reasonable? Would it be perfect?

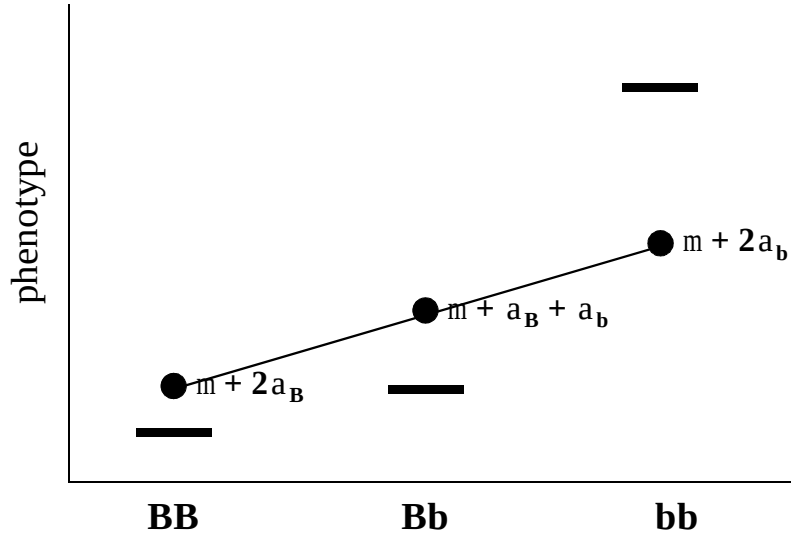
genotype values (μ_{BB}, μ_{Bb}, μ_{bb})



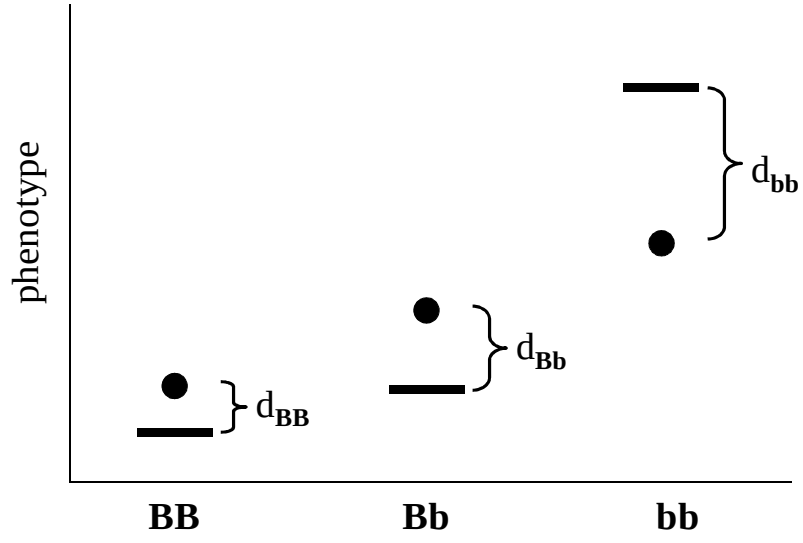
genotype values predicted from additive effects



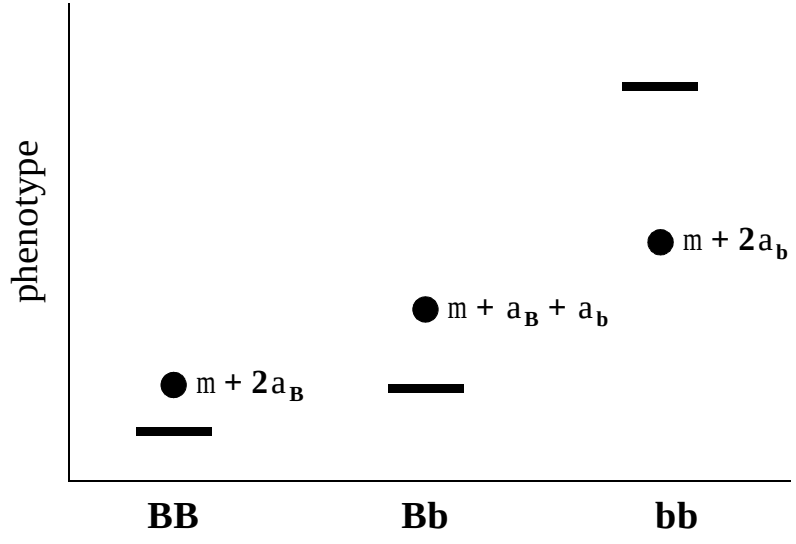
genotype values predicted from additive effects



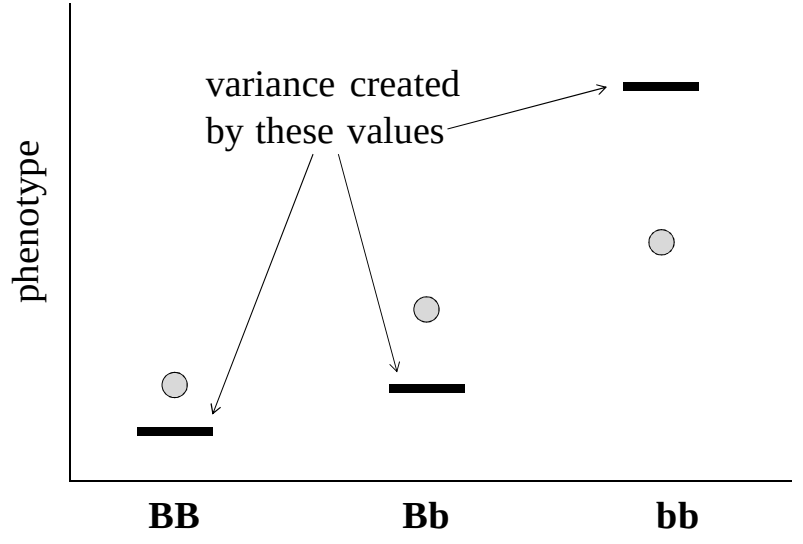
Additive effects & dominance deviations



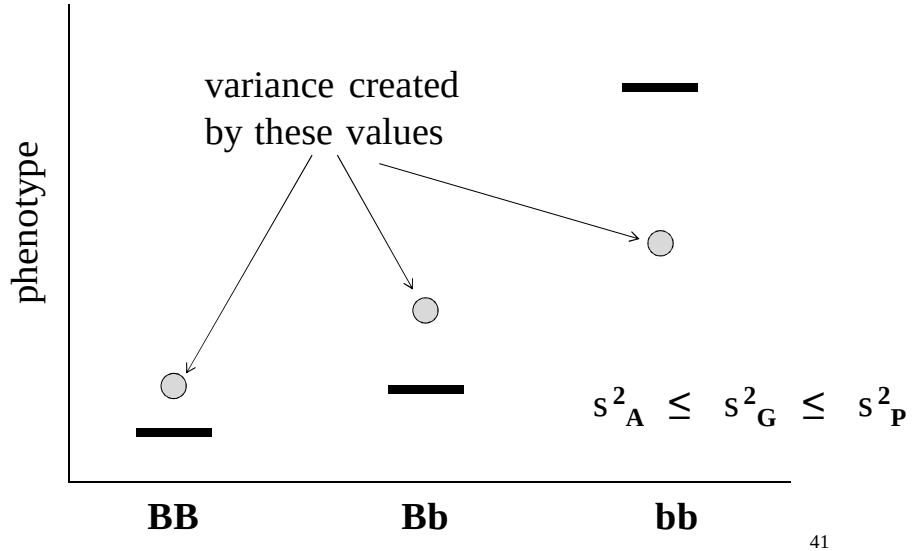
genotype values predicted from additive effects



σ^2_G : variance in phenotype due to variation in genotype



σ^2_A : Additive genetic variance



Narrow-sense heritability: h^2

- $h^2 = \sigma^2_A / \sigma^2_P$
- Of all the variance in phenotype in the population, what proportion of it is explained by the additive effects of the different alleles that are segregating.
  - what proportion of it can be inherited from one generation to the next
    - can be predicted for offspring by knowing the parents.
- Because *alleles* are transmitted from one generation to the next, not genotypes.

H^2 and h^2

- Described here for a single locus, but …
- Same principles apply for multiple loci.
  - Total heritability is the sum across loci
    - no epistasis
  - With epistatis:
    - sum across loci + sum across epistatic effects.

Both H^2 and h^2

- Depend on allele frequencies.

Ü Both H^2 and h^2 vary from population to population.
- different allele frequencies.
- different environmental conditions.

Ü Both H^2 and h^2 vary over time within a population.
- changing allele frequencies.
- changing environmental conditions.

“Missing” heritability

- Heritability is generally estimated using family data (not using genotype data)
  - estimate total heritability of a trait (not loci individually)
- Gene mapping provides loci identified to be involved with a trait of interest.
- When you take all the loci identified through mapping and estimate σ^2_G for these loci,
  - cumulatively they only account for a much smaller amount of the variance in phenotype than estimates based on family data predict.

“Missing” heritability

- Heritability accounted for by all known loci is smaller than estimates of total heritability based on family data.
  - The rest is “missing.”
- Why?
  - Many genes with small effects
    - need enormous sample sizes to find them.
  - Interactions between loci
    - heritability explained by gene x gene interactions (not well understood)
  - Epigenetic effects
  - Possible overestimates of heritability from family data
  - …