~~Sample Size and Power Estimation for QTL Experiment~~ in Plant and Animal Breeding

What is “selection”?

Idea: Reproduction is a relay race. We want the baton of *good traits* to pass to the next generation.

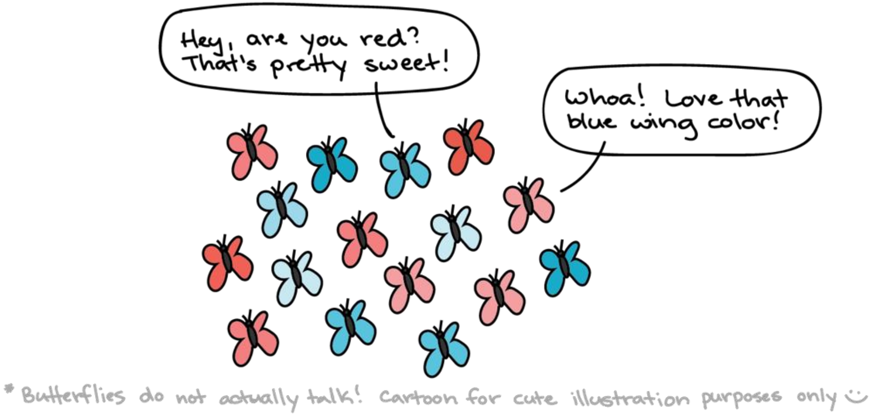
- **Selection** = choose the best individuals to be parents so useful traits become more common.
- **Natural selection:** the environment “scores” survival and reproduction.
- **Artificial selection:** people (breeders) “score” and choose the parents.

Traditional ways to select

- **Phenotypic selection:** “pick what you can see.”
- **Progeny testing:** “judge by the children.”
- **Backcross breeding:** “move a good gene into a good variety.”

We’ll look at each and why they can be slow or imprecise.

Phenotypic selection


Hey, are you red? That's pretty sweet!
Whoa! Love that blue wing color!
* Butterflies do not actually talk! Cartoon for cute illustration purposes only

How it works - Choose plants/animals with better appearance or performance.

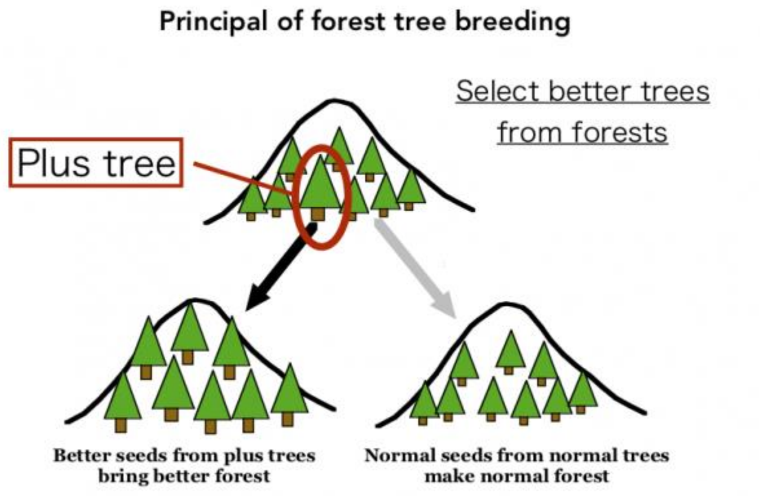
Limitations - Need many locations & seasons to separate genetics from weather and soil.

(Khan Academy Darwin-evolution-natural-selection)

Progeny testing

How it works - Some traits can’t be judged in the parent. Therefore, we look at offspring performance.

Limitations - Slow, expensive, resource-hungry.

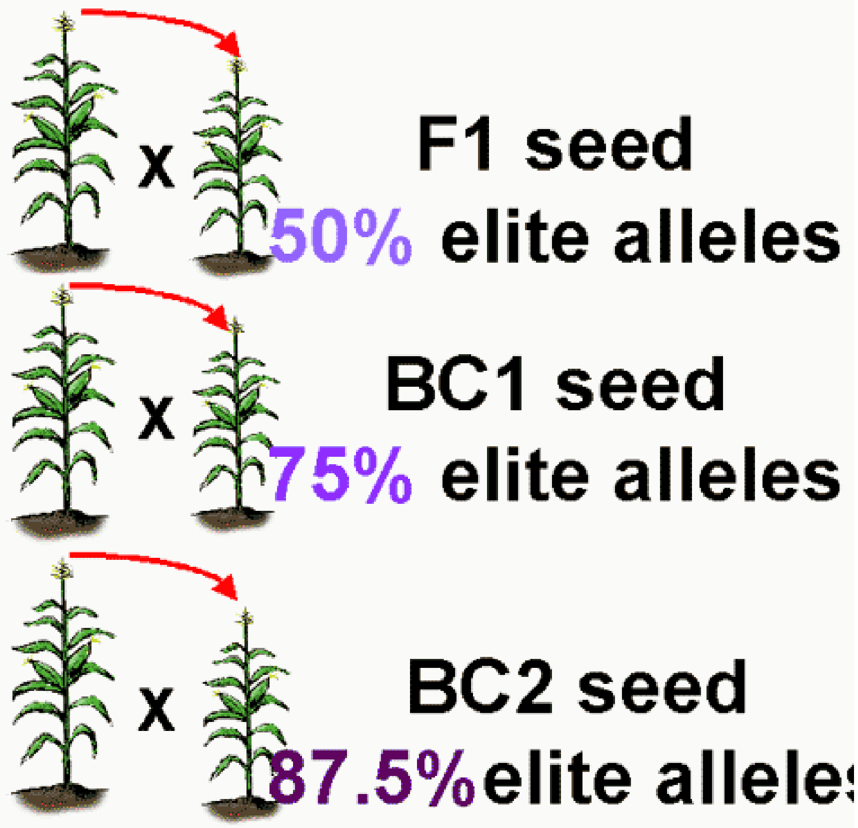

Principal of forest tree breeding
Select better trees from forests
Plus tree
Better seeds from plus trees bring better forest
Normal seeds from normal trees make normal forest

(https://www.ffpri.go.jp/hokuiku/en/research/plustree_progeny.html)

Backcross breeding

How it works - Bring a useful gene from a donor into an elite variety you already like.

Limitations - 6–8 backcross generations



(https://passel2.unl.edu/view/lesson/c55bd4193305/3)

Why traditional methods are often inefficient

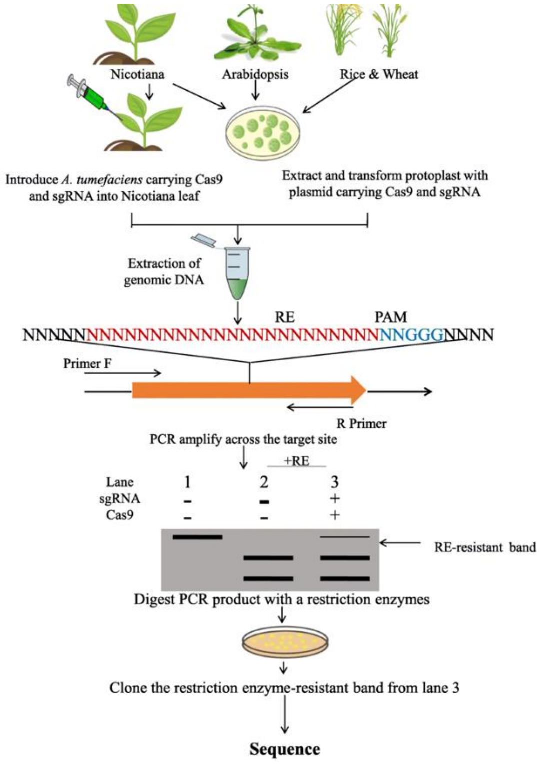
- Environment gets in the way
- Time and money
- Late or hard-to-measure traits
- Low-heritability traits
- Linkage drag

So we need faster, more precise tools !

Marker-assisted selection

Core idea - Don’t wait for traits to show up. Look at **markers !**

If a marker sits very close to a useful gene/QTL and is rarely separated by reshuffling, seeing the marker ≈ having the useful piece.


Nicotiana
Arabidopsis
Rice & Wheat
Introduce *A. tumefaciens* carrying Cas9 and sgRNA into Nicotiana leaf
Extract and transform protoplast with plasmid carrying Cas9 and sgRNA
Extraction of genomic DNA
RE
PAM
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGGNNNN
Primer F
R Primer
PCR amplify across the target site
+RE
Lane 1 2 3
sgRNA - - +
Cas9 - - +
RE-resistant band
Digest PCR product with a restriction enzymes
Clone the restriction enzyme-resistant band from lane 3
Sequence

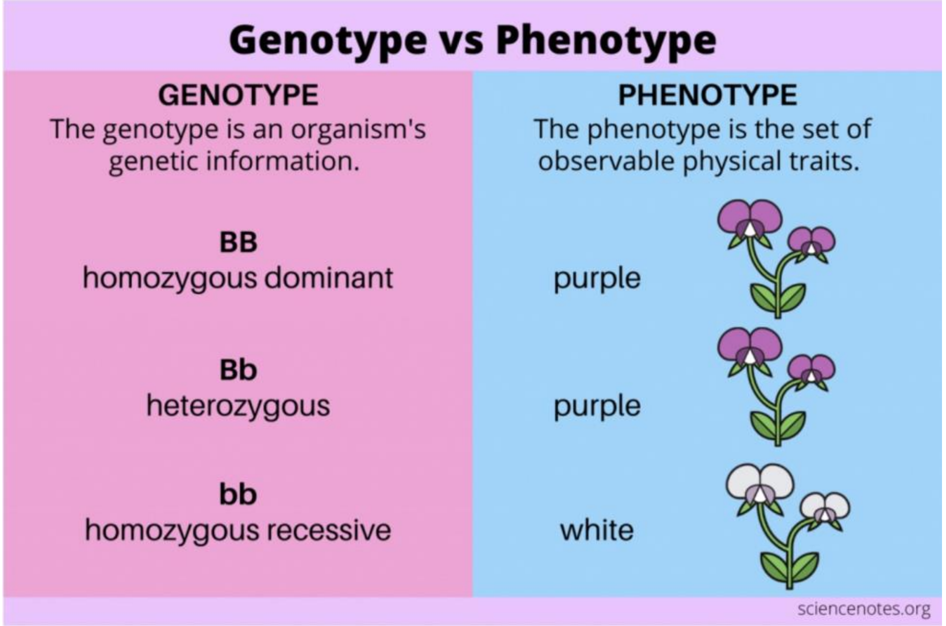
https://jgeb.springeropen.com/articles/10.1186/s43141-021-00231-1

Why MAS helps?

- Earlier screening
- More stable than visible performance
- Stack multiple good genes
- Cheaper for certain traits

Connect phenotype and genotype

- How the letters make the colour in this one-gene example:
- BB: the flower is purple.
- Bb: the flower is purple too, because B is dominant—one copy is enough to give purple and it masks b.
- bb: the flower is white, because there’s no B to make it purple.
- So, the genotype sets the instructions, and the phenotype is the outcome.



Genotype vs Phenotype

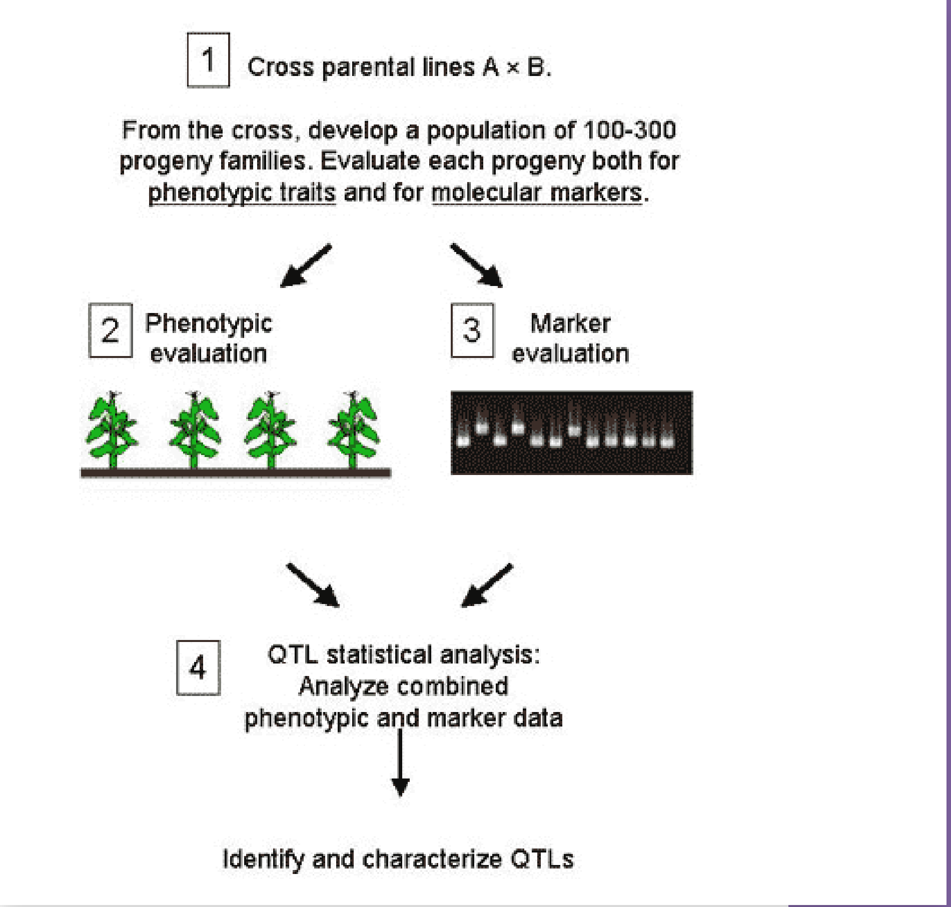
GENOTYPE The genotype is an organism's genetic information.	PHENOTYPE The phenotype is the set of observable physical traits.
BB homozygous dominant	purple
Bb heterozygous	purple
bb homozygous recessive	white

sciencenotes.org

(https://sciencenotes.org/genotype-vs-phenotype-definitions-and-examples/)

What is a QTL?

A QTL (Quantitative Trait Locus) is not a single precise gene address. It’s more like a neighbourhood—a stretch of DNA statistically linked with variation in a quantitative trait.



1 Cross parental lines A × B.

From the cross, develop a population of 100-300 progeny families. Evaluate each progeny both for phenotypic traits and for molecular markers.

2 Phenotypic evaluation

3 Marker evaluation

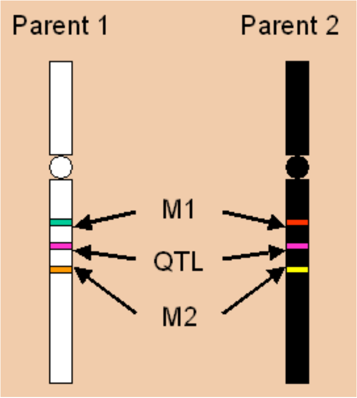
4 QTL statistical analysis: Analyze combined phenotypic and marker data

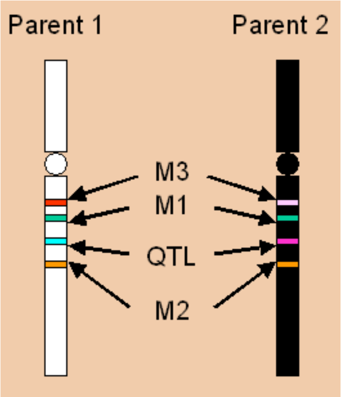
Identify and characterize QTLs

(https://passel2.unl.edu/view/lesson/aa527da941e0/2)

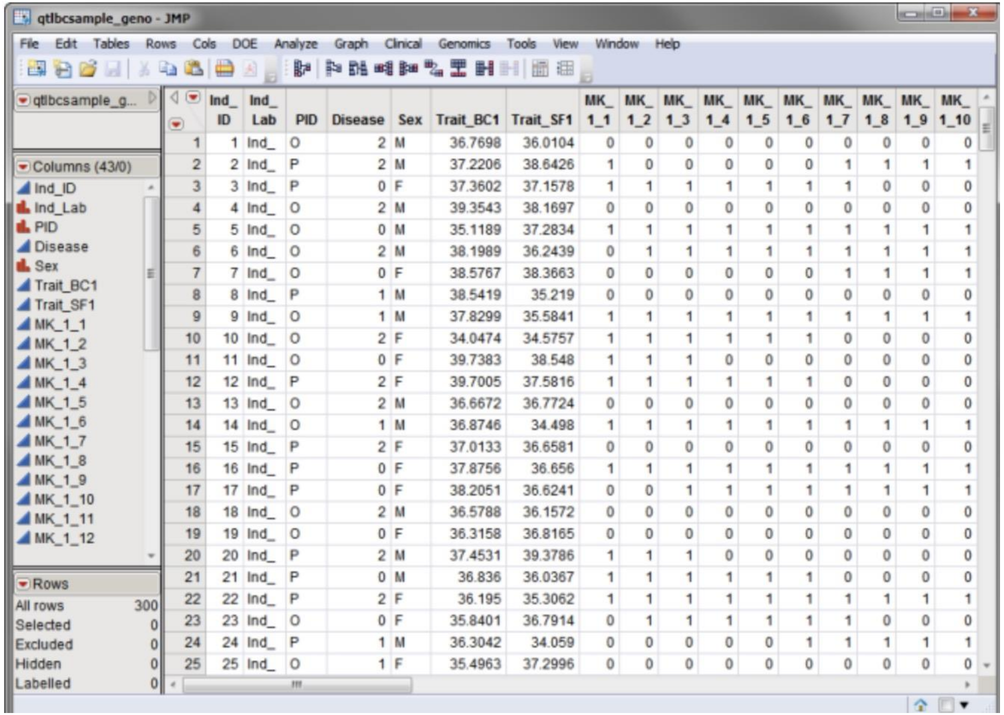
How to find QTL

- From the first figure, the QTL cannot be detected because the QTL is not polymorphic.
- From the second figure, the QTL is different, therefore, it can be detected. Besides, we can continue screening additional markers in the vicinity to find one that was polymorphic, as shown by M3





Data structure of QTL


qtlbcsample_geno - JMP
File Edit Tables Rows Cols DOE Analyze Graph Clinical Genomics Tools View Window Help
qtlbcsample_g...

Columns (43/0)

- Ind_ID
- Ind_Lab
- PID
- Disease
- Sex
- Trait_BC1
- Trait_SF1
- MK_1_1
- MK_1_2
- MK_1_3
- MK_1_4
- MK_1_5
- MK_1_6
- MK_1_7
- MK_1_8
- MK_1_9
- MK_1_10
- MK_1_11
- MK_1_12

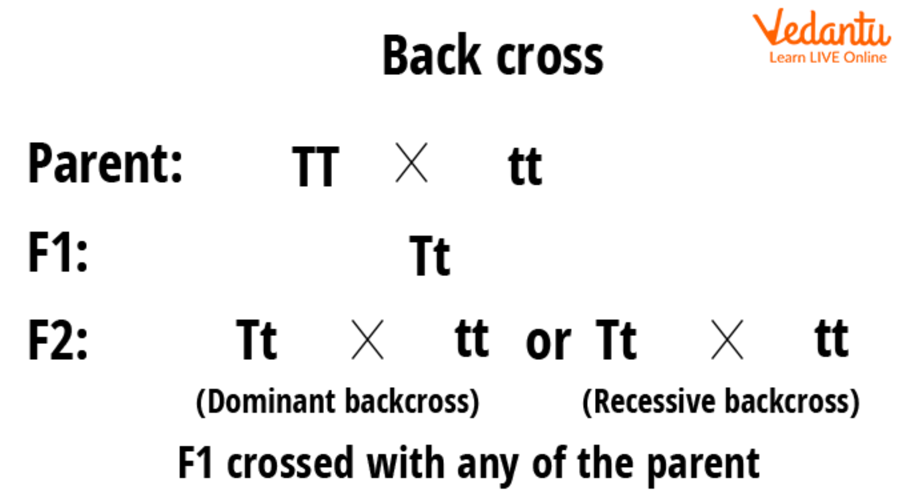
Rows	
All rows	300
Selected	0
Excluded	0
Hidden	0
Labelled	0

	Ind_ ID	Ind_ Lab	PID	Disease	Sex	Trait_BC1	Trait_SF1	MK_ 1_1	MK_ 1_2	MK_ 1_3	MK_ 1_4	MK_ 1_5	MK_ 1_6	MK_ 1_7	MK_ 1_8	MK_ 1_9	MK_ 1_10
1	1	Ind_	O	2	M	36.7698	36.0104	0	0	0	0	0	0	0	0	0	0
2	2	Ind_	P	2	M	37.2206	38.6426	1	0	0	0	0	0	1	1	1	1
3	3	Ind_	P	0	F	37.3602	37.1578	1	1	1	1	1	1	1	0	0	0
4	4	Ind_	O	2	M	39.3543	38.1697	0	0	0	0	0	0	0	0	0	0
5	5	Ind_	O	0	M	35.1189	37.2834	1	1	1	1	1	1	1	1	1	1
6	6	Ind_	O	2	M	38.1989	36.2439	0	1	1	1	1	1	1	1	1	1
7	7	Ind_	O	0	F	38.5767	38.3663	0	0	0	0	0	0	1	1	1	1
8	8	Ind_	P	1	M	38.5419	35.219	0	0	0	0	0	0	0	0	0	0
9	9	Ind_	O	1	M	37.8299	35.5841	1	1	1	1	1	1	1	1	1	1
10	10	Ind_	O	2	F	34.0474	34.5757	1	1	1	1	1	1	0	0	0	0
11	11	Ind_	O	0	F	39.7383	38.548	1	1	1	0	0	0	0	0	0	0
12	12	Ind_	P	2	F	39.7005	37.5816	1	1	1	1	1	1	0	0	0	0
13	13	Ind_	O	2	M	36.6672	36.7724	0	0	0	0	0	0	0	0	0	0
14	14	Ind_	O	1	M	36.8746	34.498	1	1	1	1	1	1	1	1	1	1
15	15	Ind_	P	2	F	37.0133	36.6581	0	0	0	0	0	0	0	0	0	0
16	16	Ind_	P	0	F	37.8756	36.656	1	1	1	1	1	1	1	1	1	1
17	17	Ind_	P	0	F	38.2051	36.6241	0	0	1	1	1	1	1	1	1	1
18	18	Ind_	O	2	M	36.5788	36.1572	0	0	0	0	0	0	0	0	0	0
19	19	Ind_	O	0	F	36.3158	36.8165	0	0	0	0	0	0	0	0	0	0
20	20	Ind_	P	2	M	37.4531	39.3786	1	1	1	0	0	0	0	0	0	0
21	21	Ind_	P	0	M	36.836	36.0367	1	1	1	1	1	1	0	0	0	0
22	22	Ind_	P	2	F	36.195	35.3062	1	1	1	1	1	1	1	1	1	1
23	23	Ind_	O	0	F	35.8401	36.7914	0	1	1	1	1	1	1	0	0	0
24	24	Ind_	P	1	M	36.3042	34.059	0	0	0	0	0	1	1	1	1	1
25	25	Ind_	O	1	F	35.4963	37.2996	0	0	0	0	0	0	0	0	0	0

Datasets often encode these two classes as 0/1 to simplify regression and plotting. A common convention is g = 0 for AA and g = 1 for AB.

(https://www.jmp.com/support/downloads/JMPG101_documentation/Content/JMPGUserGuide/PR_G_GN_0048.htm)

Backcross & Intercross



Back cross

Vedantu
Learn LIVE Online

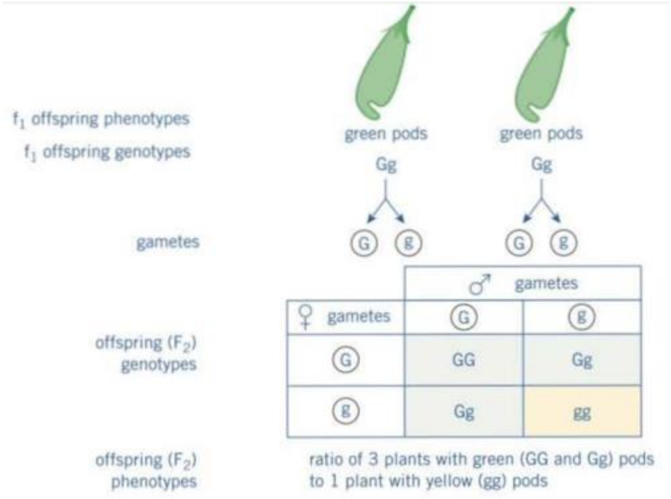
Parent: TT × tt

F1: Tt

F2: Tt × tt or Tt × tt

(Dominant backcross) (Recessive backcross)

F1 crossed with any of the parent



f_1 offspring phenotypes
f_1 offspring genotypes
green pods
green pods
Gg
Gg
gametes
G
g
G
g
♂ gametes
♀ gametes
G
g
offspring (F_2) genotypes
G
GG
Gg
g
Gg
gg
offspring (F_2) phenotypes
ratio of 3 plants with green (GG and Gg) pods to 1 plant with yellow (gg) pods

▲ **Figure 2** *F_1 intercross between pea plants that are heterozygous for green pods*

Backcross: The target locus is known, markers are well defined, and the aim is to rapidly **import and retain** the recipient parent background.
Advantages: fast controllable genetic background.
Disadvantages: not suitable for accumulating many small-effect loci.

(https://www.vedantu.com/biology/test-cross)

Inter-cross: The trait is controlled by multiple loci and requires accumulating/recombining **favourable** alleles.
Advantages: advance improvement at multiple loci simultaneously.
Disadvantages: longer breeding cycle; more demanding in terms of population size and marker design.

(https://quizlet.com/gb/625534518/inheritance-flash-cards/)

Phenotype factor decomposition

- Full model: $P = G + E + G \times E + residual$

Phenotype

Genetic effect

Environment effect

Co-effect of genome and environment

- Simplified model:

$$P = G + residual$$

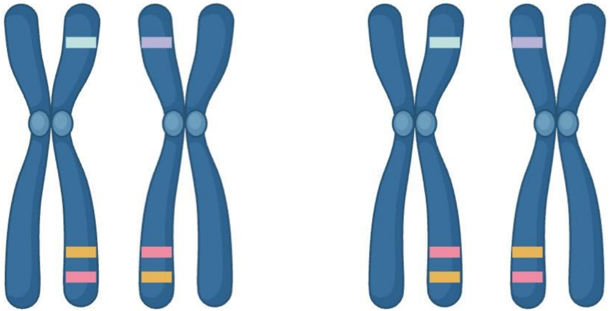
Phenotype

Genetic effect

Everything else

Linkage & Recombination frequency

- Linkage = two loci on the same chromosome tend to be inherited together because crossovers between them are infrequent.

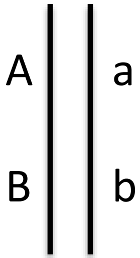


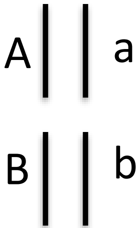
Recombination:

$$r = \Pr(recombinant\ gamete)$$

Unlinked: $r = 0.5$

Tight linkage $r \approx 0$





AB Ab aB ab

$r = 0.5$

Single Marker Model in Backcross

- In a backcross population, each autosomal locus has only two genotypes AA AB. $G_i \in \{0,1\}$, $0 = AA, 1 = AB$.
- Work with the two group means μ_{AA}, μ_{AB}, $\Delta = \mu_{\mathrm{AB}} - \mu_{\mathrm{AA}}$

$$y_i = \mu + \beta G_i + \varepsilon_i \quad ,$$

$$\text{where } \beta = \mu_{\mathrm{AB}} - \mu_{\mathrm{AA}}$$

P AA × BB

F1 AB × AA

F2 AA or AB

Classical one-way ANOVA (genotype as the factor)

- ~~ANOVA~~
- $H_0\colon\ \mu_{AA} = \mu_{AB}$
- Classical one-way ANOVA
- The between-group sum of squares $SS_B = \sum_j n_j(\bar{y}_j - \bar{y})^2$
- The within-group sum of squares $SS_W = \sum_j \sum_{i\in j}(y_{ij} - \bar{y}_j)^2$

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

with $\mathrm{k} = 2.$

Mixture Distribution

- Let the left and right flanking markers be M_1, M_2, and the candidate QTL position be Q.
- $r_{12} = Recomb(M_1, M_2),\ r_{1Q} = Recomb(M_1, Q),$
  $$r_{2Q} = Recomb(M_2, Q).$$
- (1) Left = AA, Right = AA

$$\Pr(Q = AA) = \frac{(1 - r_{1Q})(1 - r_{2Q})}{1 - r_{12}},$$

$$\Pr(Q = AB) = \frac{r_{1Q} r_{2Q}}{1 - r_{12}}$$

$Q = AA$:”no recomb left”+ :”no recomb right”= $(1 - r_{1Q})(1 - r_{2Q})$

- (2) Left = AA, Right = AB

$$\Pr(Q = AA) = \frac{(1 - r_{1Q})r_{2Q}}{r_{12}},$$

$$\Pr(Q = AB) = \frac{r_{1Q}(1 - r_{2Q})}{r_{12}}$$

- (3) Left = AB, Right = AA

$$\Pr(Q = AA) = \frac{r_{1Q}(1 - r_{2Q})}{r_{12}},$$

$$\Pr(Q = AB) = \frac{(1 - r_{1Q})r_{2Q}}{r_{12}}$$

If the flanks are **different** (AA,AB or AB,AA), there was an **odd** number of crossovers; the event probability is r_{12}

- (4) Left = AB, Right = AB

$$\Pr(Q = AA) = \frac{r_{1Q} r_{2Q}}{1 - r_{12}},$$

$$\Pr(Q = AB) = \frac{(1 - r_{1Q})(1 - r_{2Q})}{1 - r_{12}}$$

If the two flanking marker genotypes are the same (AA,AA or AB,AB), then there must be an even number of crossovers between M_1, M_2.
The probability of that conditioning event is $1 - r_{12}$.

Multiple markers model

- Goal: consider multiple markers simultaneously for a continuous trait, rather than analysing one marker at a time.

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i,\ \ \varepsilon_i \sim N(0, \sigma^2)$$

Phenotype of the $i-th$ individual

The effect size of marker β_j

The numeric coding of marker j for individual i

Limitation in Multiple Marker Model

- **High dimensionality:** The number of markers p is often comparable to, or even much larger than, the sample size n $(p \gg n)$;
- **Collinearity:** linkage disequilibrium (LD) makes columns highly correlated.
- As a result, plain OLS is unstable or even not identifiable.

Strategies for Multiple Marker Model

- **1、Stepwise selection**

simple but unstable under high dimension; ignores model uncertainty.

- **2、Bayesian shrinkage**

$$\beta_j \mid \sigma^2, \lambda_j \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda_j}\right), \qquad \lambda_j \sim \mathrm{Gamma}(\alpha, \theta)$$

Marker-specific Gaussian shrinkage.

Large λ_j → strong shrinkage ($\beta_j \approx 0$)

Small λ_j → large effects

• 3、Penalized regression-Elastic Net

$$\min_{\beta} \|y - Z\beta\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

Ridge$(\boldsymbol{\lambda_2 > 0, \lambda_1 = 0})$: shrinks coefficients; no variable selection; robust under strong LD.

Lasso $(\boldsymbol{\lambda_1 > 0, \lambda_2 = 0})$: can shrink some β_j exactly to 0, but tends to pick one of several highly correlated markers.

Elastic net: combines both; good for *groups* of correlated markers.