Commonly used statistical models in QTL

Presenter: Shiyan Miao

Introduction of myself

- I am Shiyan Miao a third-year student in Bachelor of Science, major in Statistics and minor in Mathematics.
- I used to study Finance in SUIBE, and then I transfer to ANU to change my major.

Selection in Biology

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 to be parents, useful traits.
- **Natural selection:** the environment “scores” survival and reproduction.
- **Artificial selection:** breeders “score” and choose the parents.

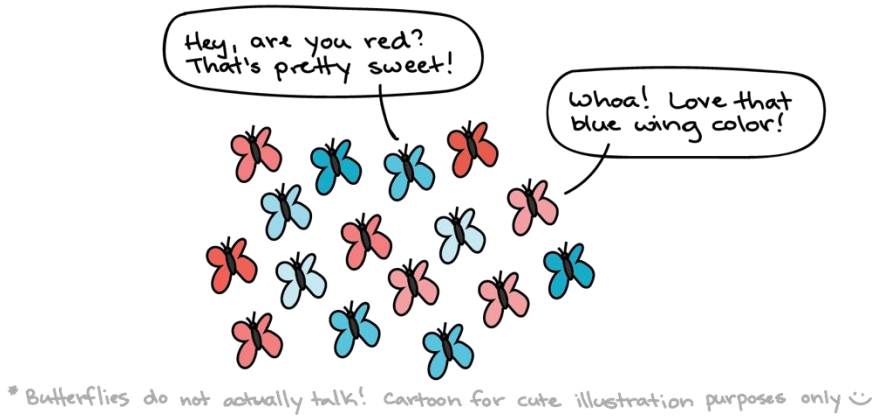
(Falconer & Mackay, 1996.)

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



(Khan Academy Darwin-evolution-natural-selection)

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

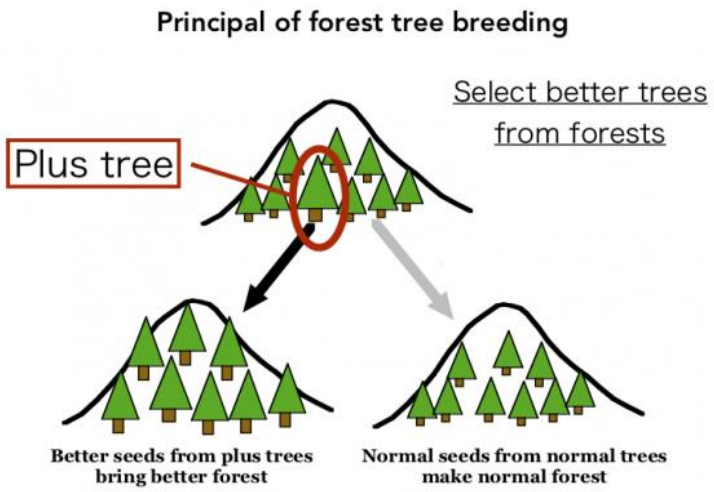
Limitations - Need many locations & seasons to separate genetics from weather and soil.

(Collard et al., 2005.)

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.



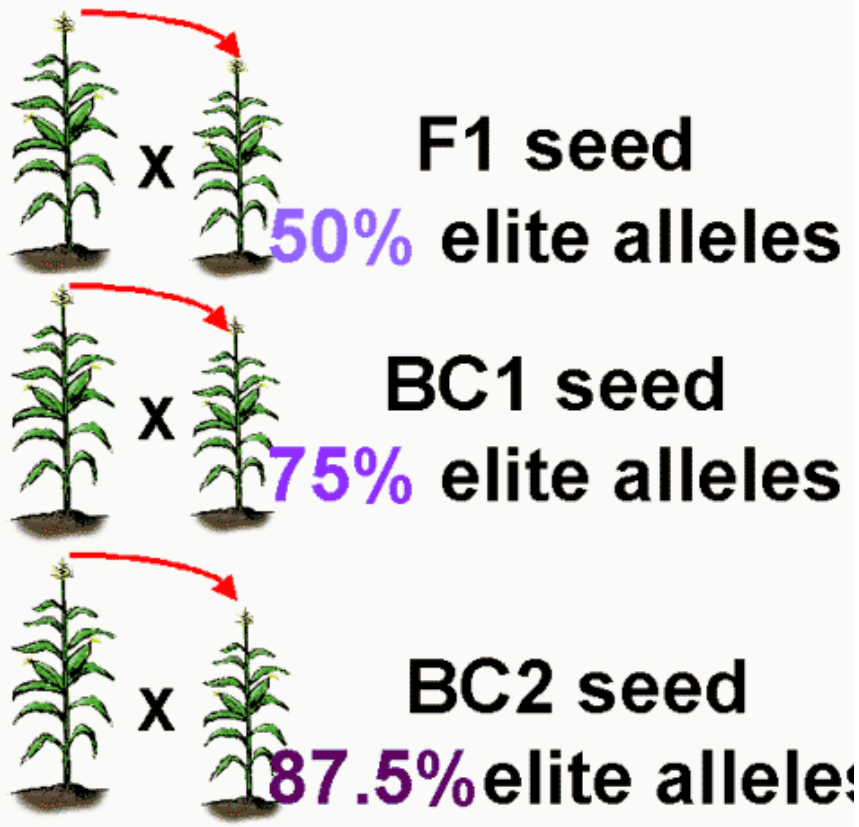
(https://www.ffpri.go.jp/hokuiku/en/research/plustree_progeny.html, Hokkaido Regional Breeding Office, Forest Tree Breeding Center)

(Falconer & Mackay, 1996.)

Backcross breeding

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

Limitations - 6–8 backcross generations



(Wu, Ma, & Casella, 2007.)

(https://passel2.unl.edu/view/lesson/c55bd4193305/3 , Plant and Soil Sciences eLibrary 2025)

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

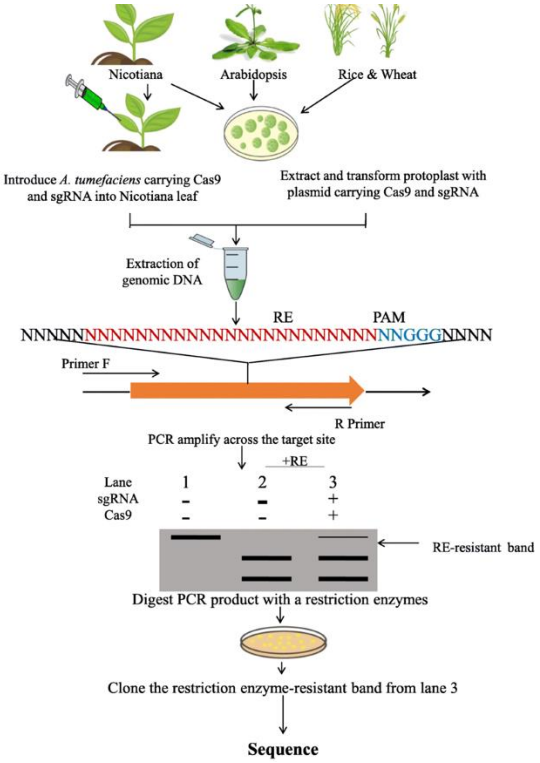
So we need faster, more precise tools (Marker-assisted selection) !

(Collard et al., 2005.)

Marker-assisted selection

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

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



Hasan, N., Choudhary, S., Naaz, N. *et al.* Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. *J Genet Eng Biotechnol* **19**, 128 (2021). https://doi.org/10.1186/s43141-021-00231-1

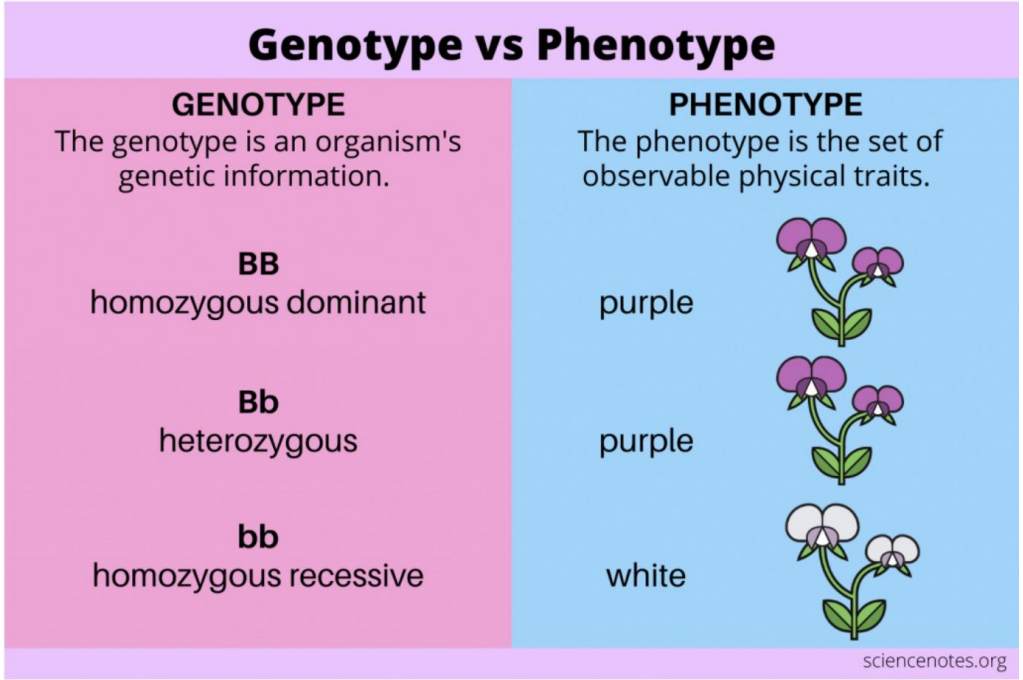
(Collard et al., 2005.)

Why MAS helps?

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

(Collard et al., 2005.)

Connect phenotype and genotype



(https://sciencenotes.org/genotype-vs-phenotype-definitions-and-examples/, Science Notes and Projects)

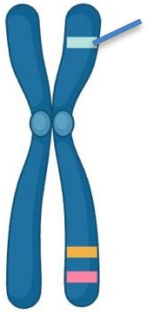
(Mackay et al., 2009.)

Quantitative trait loci (QTL)

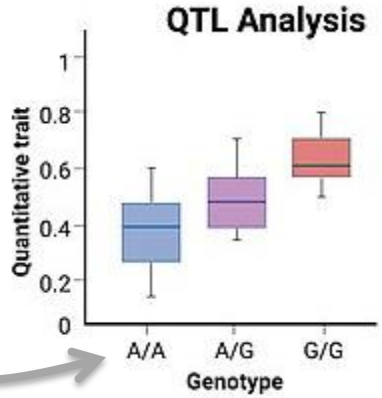
What is a QTL?

A stretch of DNA that control the quantitative trait.

To find QTL:



A/A, A/G, G/G



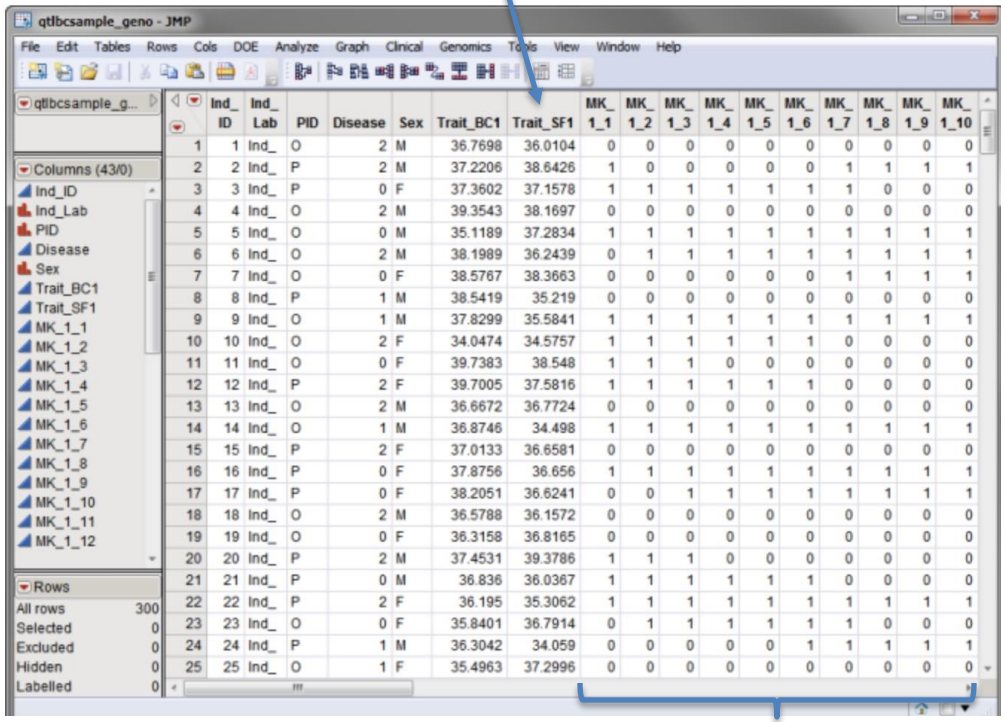
Sequence “markers”

Statistical analysis of marker – trait variation

(Mackay et al., 2009.)

QTL data

Phenotype



	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

Marker genotype

Genotype are encoded as 0 or 1

0: AA

1: GG

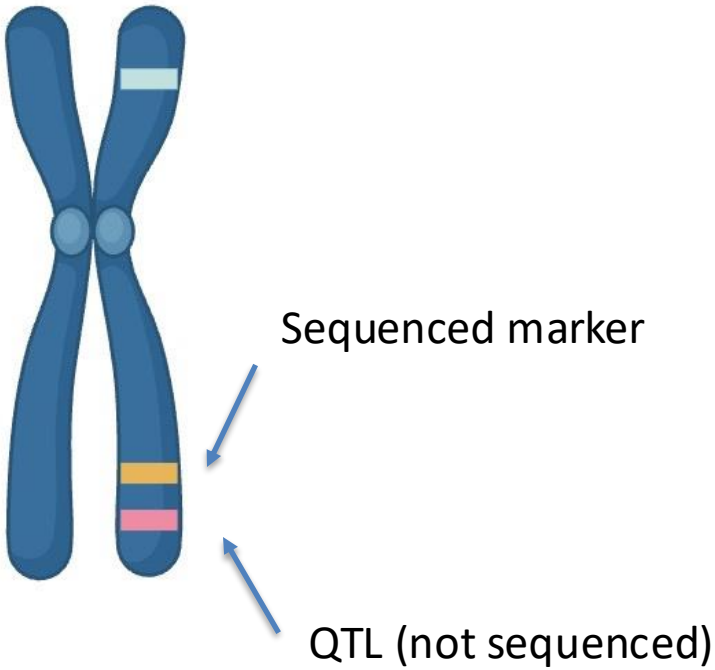
(Mackay et al., 2009.)

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

What if a marker is not on a QTL?

Due to financial reason, we can't sequence the entire genome

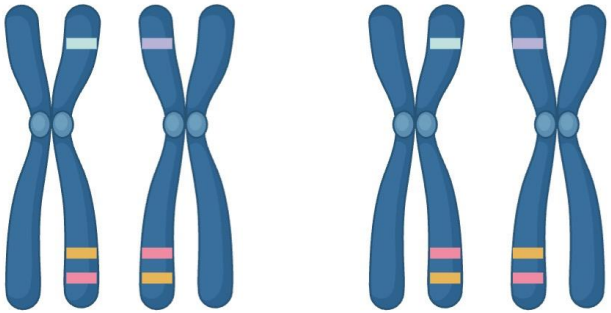
We can still make inference of QTL based on sequenced markers due to **linkage between them**



(Mackay et al., 2009.)

Linkage & Recombination frequency

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

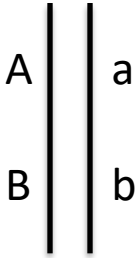


Recombination:

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

Unlinked: $r = 0.5$

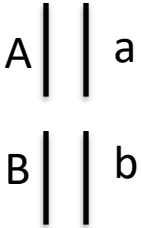
Tight linkage $r \approx 0$



AB ab

$r = 0$

Complete linkage



AB ab Ab aB (recombined)

$r = 0.5$

A and B can freely combine

Statistical model for detecting QTL

(Foster, 2006.)

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:

$$P = G + residual$$

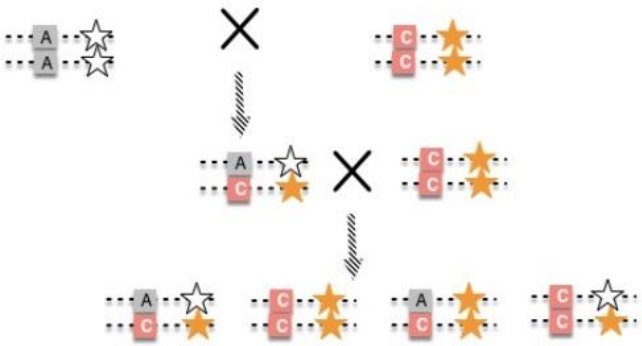
Phenotype

Genetic effect

Everything else

(Falconer & Mackay, 1996.)

Experimental population

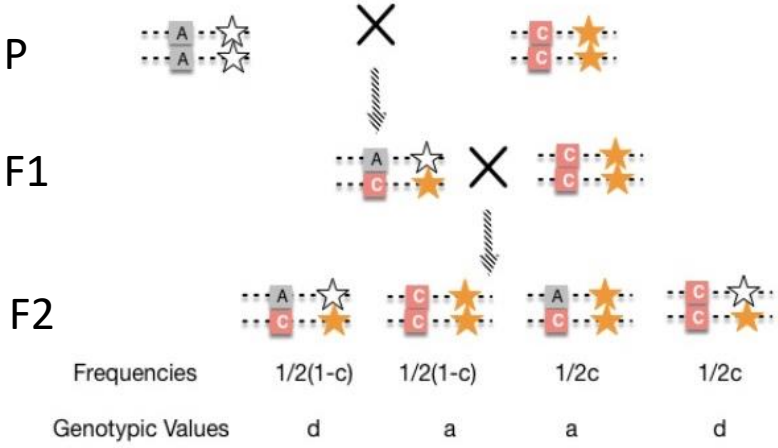


- Intercross (F2 X F2) is another frequently used population

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, Vedantu) (Collard et al., 2005. Foster, 2006.)

Single marker model



c here is the recombination freq.

Mean of the AC genotype:

$$\frac{1/2(1-c)\times d + 1/2c \times a}{1/2}$$

$$= \mathrm{d}(1-\mathrm{c}) + \mathrm{ca}$$

Mean of the CC genotype:

$$\frac{1/2(1-c)\times a + 1/2c \times d}{1/2}$$

$$= a(1-\mathrm{c}) + \mathrm{cd}$$

Difference between AC and CC

$$\mathrm{d}(1-\mathrm{c}) + \mathrm{ca} - (a(1-\mathrm{c}) + \mathrm{cd})$$

$$= (\mathrm{d}-\mathrm{a})(1-2\mathrm{c})$$

(https://jyanglab.com/slides/2024-agro931/week15/w15_c1.html#20 Jyang Lab)

Single Marker Model in Backcross

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

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

$$\text{where } \beta = \mu_{\mathrm{AC}} - \mu_{\mathrm{CC}} = (\mathrm{d} - \mathrm{a})(1 - 2\mathrm{c}) \quad \text{(t-test)}$$

(Foster, 2006.)

Recombination Frequency

From the case above, we can see the difference between AC and CC is

$$(\mathrm{d} - \mathrm{a})(1 - 2\mathrm{c})$$

1. $(\mathrm{d} - \mathrm{a}) = 0$

No genetic effects

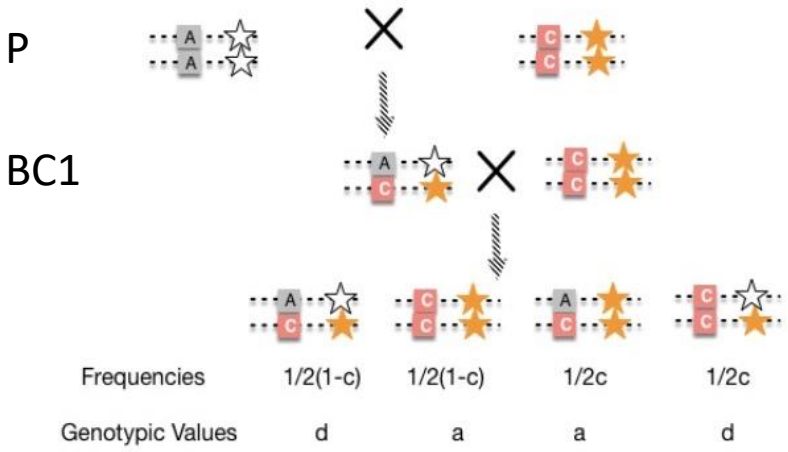
2. $(1 - 2\mathrm{c}) = 0 \rightarrow \mathrm{c} = 1/2$

Marker and the QTL is unlinked.

Mixture Distribution

- We do not know the exact QTL genotype of each sample, but we do know the probability of its occurrence.

- In BC1, with marker genotypes **AC** and **CC**, the underlying QTL genotypes are only two: $Qq\ (heterozygote), qq\ (homozygote)$



c here is the recombination freq.

(Foster, 2006.)

Conditional Probabilities & Mixture Density Function

- Let the recombination fraction between the marker (AC/CC) and QTL be c.

- If the marker genotype is **AC**:

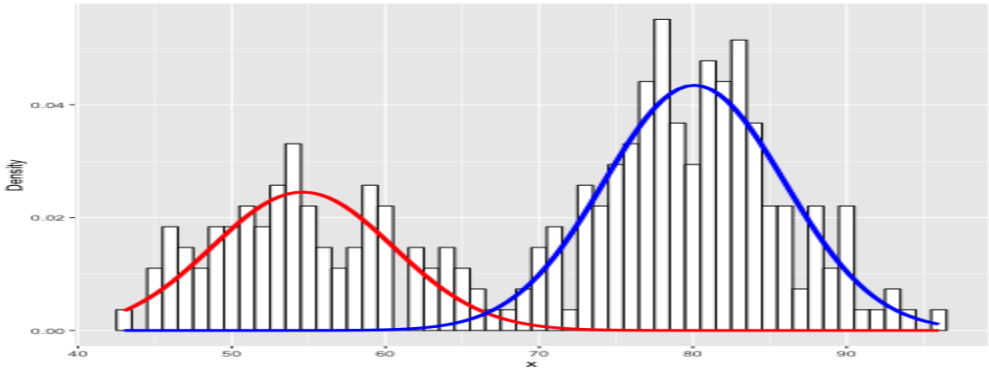
$$P(Qq|AC) = 1 - c, P(qq|AC) = c$$

- If the marker genotype is **CC**:

$$P(Qq|CC) = c, P(qq|CC) = 1 - c$$

(Foster, 2006.)

Mixture Distribution



The overall phenotypic distribution (ignoring markers) is a two-component normal mixture:

$$f(z|AC) = (1-c)\emptyset\left(z;\mu_{Qq},\sigma^2\right) + c\,\emptyset\left(z;\mu_{qq},\sigma^2\right)$$
$$f(z|CC) = c\,\emptyset\left(z;\mu_{Qq},\sigma^2\right) + (1-c)\,\emptyset\left(z;\mu_{qq},\sigma^2\right)$$

(Foster, 2006.)

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

(Foster, 2006.)

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 $\rightarrow$ strong shrinkage ($\beta_j \approx 0$)
Small λ_j $\rightarrow$ large effects

(Foster, 2006.)

- **3、Penalized regression-Elastic Net**

$$\min_{\beta} \underbrace{\|y - Z\beta\|^2}_{\text{Fitting residuals}} + \underbrace{\lambda_1 \sum_{j=1}^{p} |\beta_j|}_{\text{L1 penalty}} + \underbrace{\lambda_2 \sum_{j=1}^{p} \beta_j^2}_{\text{L2 penalty}}$$

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.

(Foster, ch. 1, citing Hoerl & Kennard, 1970a, 1970b; Lindley & Smith, 1972; Frank & Friedman, 1993; Breiman, 1995; Tibshirani, 1996; Fu, 1998; Zou & Hastie, 2005)

Recommendation for the models

- **Single-Marker Model:** Few QTLs with moderate–large effects; quick screening on backcross.

- **Multiple-Marker Model:** Many small/medium effects; need conditional effects while controlling other loci.

- **Elastic Net Strategy:** Combines sparsity + grouping of correlated markers → stable selection; Controls overfitting with shrinkage.

What we will do in the future

- Develop an interactive R shiny app, which will allow user to simulate quantitative traits.

- Sample size calculation to save the cost of sequencing markers when handle with large sample size.

Reference

1. Broman, K. W., & Sen, Ś. (2009). *A guide to QTL mapping with R/qtl*. Springer.
2. Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica, 142*, 169–196.
3. Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). Prentice Hall.
4. Foster, S. (2006). *The LASSO linear mixed model for mapping quantitative trait loci* [Doctoral dissertation, The University of Adelaide].
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
6. Mackay, T. F. C., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: Challenges and prospects. *Nature Reviews Genetics, 10*, 565–577.
7. Miller, A. (2002). *Subset selection in regression* (2nd ed.). Chapman & Hall/CRC.
8. Wu, R., Ma, C.-X., & Casella, G. (2007). *Statistical genetics of quantitative traits: Linkage, maps and QTL*. Springer.
9. Yi, N., & Shriner, D. (2008). Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses. *Heredity, 100*, 240–252.