FRST302: Forest Genetics

Lecture 4.1 - Accelerating Tree Improvement I

Welcome to Module 4!

Housekeeping

- Module test will be a mix of multiple choice and short answer questions
- The test questions will test your understanding of concepts and ask you to apply it
- **Terminology** will be covered on lecture slides
- **Concepts** will be covered in lectures
- If you are lost with the concepts or terminology make use of office hours, the Canvas Discussion board and ask questions in class!

Recap

Module 1 Genes, Genomes and Sequencing

Module 2 Population Genetics and Local Adaptation

Module 3 Quantitative Genetics and 20th Century Breeding

Recap

Module 1 Genes, Genomes and Sequencing

Module 2 Population Genetics and Local Adaptation

Module 3 Quantitative Genetics and 20th Century Breeding
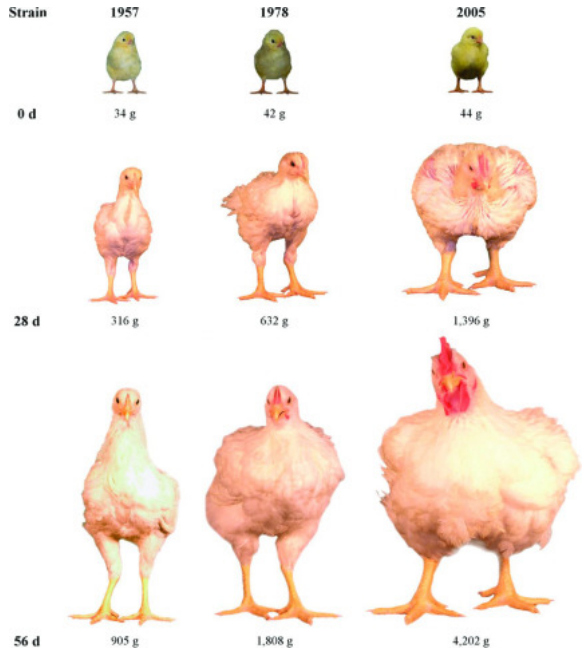
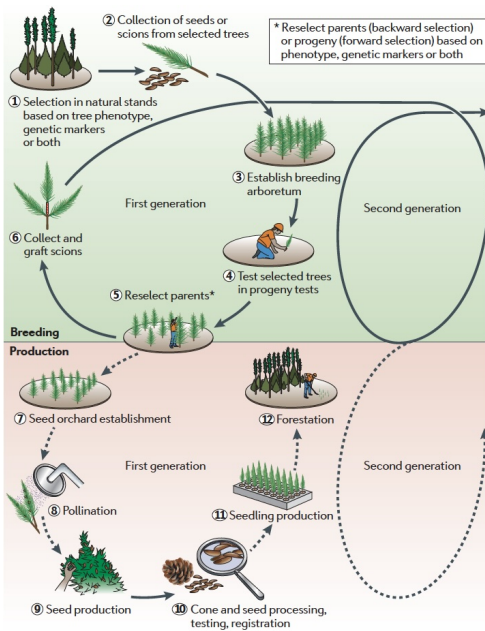**Module 4** Advancing Forest Genetics with Recent Technology

a) *Brassica rapa* morphotypes; b) *Brassica oleracea* morphotypes - from Cheng *et al* 2016

Quantitative genetic models work and have been very effective in the last 100 years!

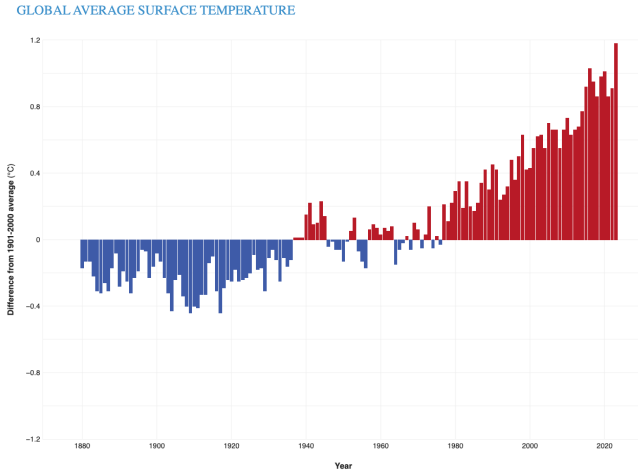| Strain	1957	1978	2005
0 d	34 g	42 g	44 g
28 d	316 g	632 g	1,396 g
56 d	905 g	1,808 g	4,202 g

Modified from Figure 1 - Zuidhof et al. 2014

Conventional Tree Breeding
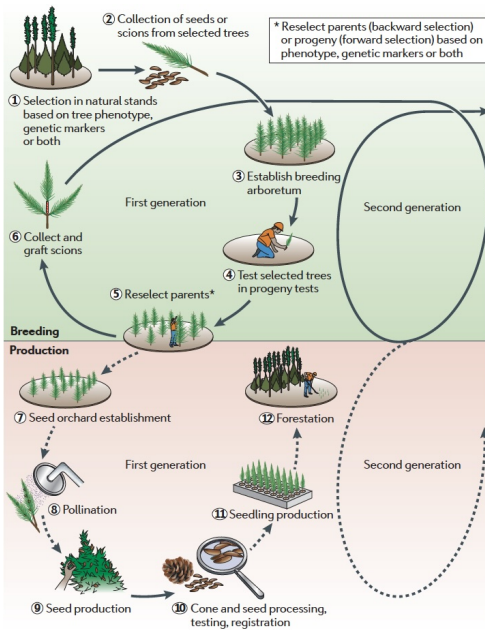
Conventional Tree Breeding
Each cycle takes from 20-30 years!

Image from Neale and Kremer 2011

Can We Afford to Take That Long?

GLOBAL AVERAGE SURFACE TEMPERATURE

How could you advance tree breeding with what you learned in the last three modules?

Many of the interventions we may want to take would involve knowing the genes that underly important traits

For species that are intractable to cultivate in a lab, how can we identify important genetic variation?

Let's Build a Model!
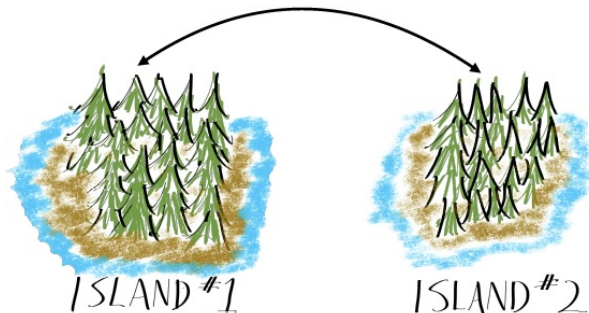
- Imagine a tree species inhabiting two islands
- Pollen flows between them

- Taller trees are favoured on Island #1
- 99.9% of the time trees pollinate individuals from their own island[a]

[a]This gives an expected F_{ST} of 0.05

- Diploid individuals

The details on this slide are just for understanding the model. You *will not* be tested on memorizing these numbers and/or parameters

The Genetics of our Model

- Diploid individuals
- The genome is composed of a single chromosome $1cM$ long

The details on this slide are just for understanding the model. You *will not* be tested on memorizing these numbers and/or parameters

The Genetics of our Model

- Diploid individuals
- The genome is composed of a single chromosome $1cM$ long
- Mutations that influence height are co-dominant, with effects drawn from a Gaussian distribution with $\sigma = 0.1$

The details on this slide are just for understanding the model. You *will not* be tested on memorizing these numbers and/or parameters

The Genetics of our Model

- Diploid individuals
- The genome is composed of a single chromosome $1cM$ long
- Mutations that influence height are co-dominant, with effects drawn from a Gaussian distribution with $\sigma = 0.1$
- Mutation rate, $\mu = 1 \times 10^{-9}/bp/generation$

The details on this slide are just for understanding the model. You *will not* be tested on memorizing these numbers and/or parameters

The Genetics of our Model

- Diploid individuals
- The genome is composed of a single chromosome $1cM$ long
- Mutations that influence height are co-dominant, with effects drawn from a Gaussian distribution with $\sigma = 0.1$
- Mutation rate, $\mu = 1 \times 10^{-9}/bp/generation$
- Heritability of height $h^2 = 0.4$

The details on this slide are just for understanding the model. You *will not* be tested on memorizing these numbers and/or parameters

Phenotypic Variation on the Islands
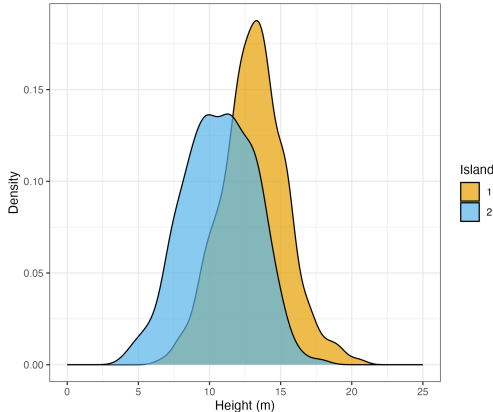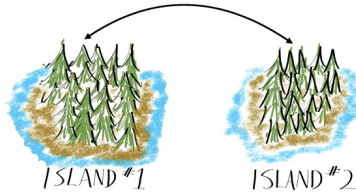
- Mean height is on 13.11 on Island 1 and 10.71 on Island 2
- This slight difference was statistically significant ($p << 0.001$)

I'll add a picture here of the chromosome with a marker...

Testing for Genetic Associations With Tree Height

To test for an association of a trait (in this case height) with SNP, we can do a statistical regression:

$$\hat{Y}_i \sim \alpha + \beta_j X_i + \epsilon$$

Testing for Genetic Associations With Tree Height

To test for an association of a trait (in this case height) with SNP, we can do a statistical regression:

$$\hat{Y}_i \sim \alpha + \beta_j X_i + \epsilon$$

$\hat{Y}$: The phenotype of individual i

α: The population mean

β: The effect of marker j on the trait

X_i: The number of copies of marker j that individual i possesses

ϵ: The effect of the environment on the trait

By fitting a regression to the data we can estimate the effect size of a particular marker on a trait and test for statistical significance...
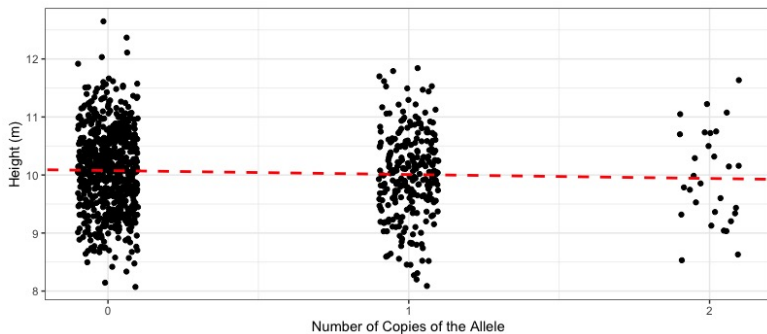
Testing Association at a Single Marker
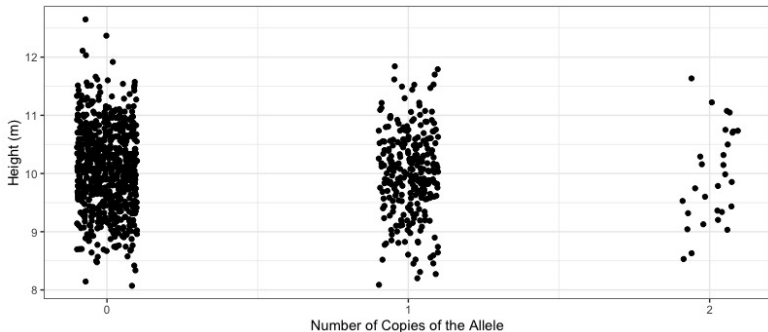
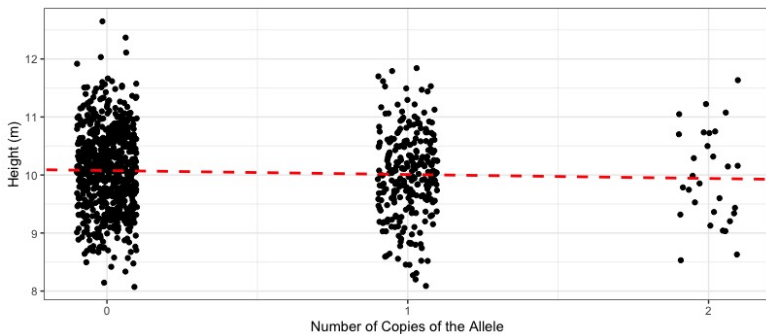*The allele is at a frequency of 0.16 across both islands*

Testing Association at a Single Marker

The estimated β for this SNP is $\beta = -0.07$

The $p - value$ for this regression was $p = 0.108$

Testing Association at a Single Marker

The estimated β for this SNP is $\beta = -0.07$

The $p-value$ for this regression was $p = 0.108$

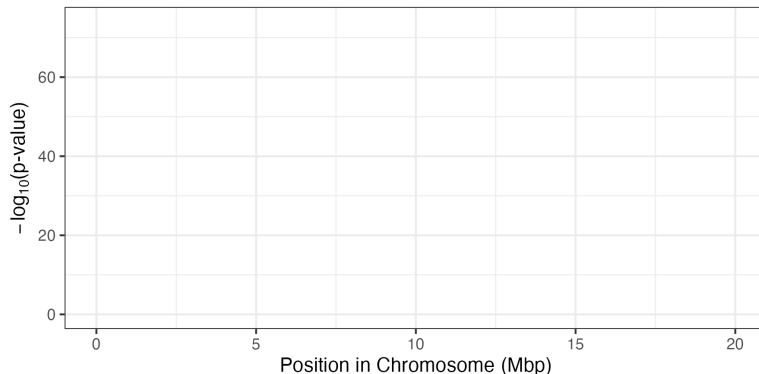What can we say about this marker?

Genome-Wide Association Study

Now, let's look at each SNP across the whole genome

There's nothing special about $-log_{10}(p - value)$, it's just the $p - value$ of an association expressed in an easy to visualise way

Genome-Wide Association Study

Now, let's look at each SNP across the whole genome

There's nothing special about $-log_{10}(p - value)$, it's just the $p - value$ of an association expressed in an easy to visualise way

Genome-Wide Association Study
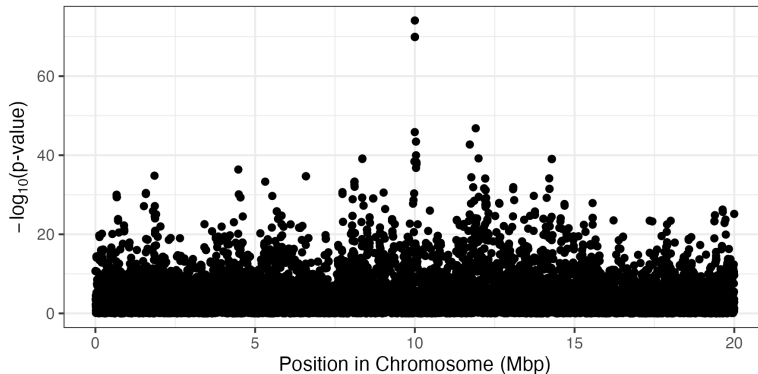
Now, let's look at each SNP across the whole genome

There's nothing special about $-log_{10}(p-value)$, it's just the $p-value$ of an association expressed in an easy to visualise way

Now, let's look at each SNP across the whole genome

There's nothing special about $-log_{10}(p - value)$, it's just the $p - value$ of an association expressed in an easy to visualise way

Genome-Wide Association Study
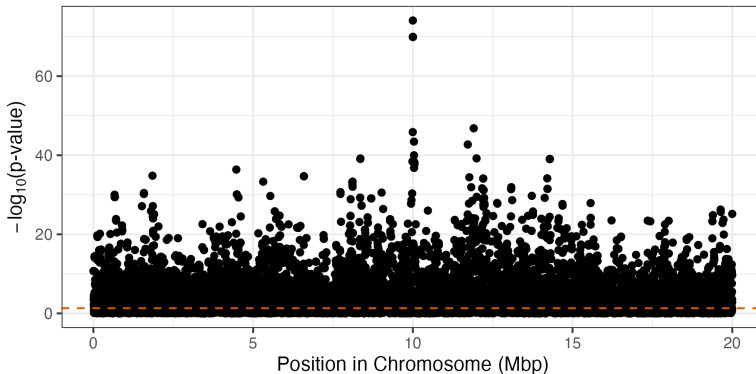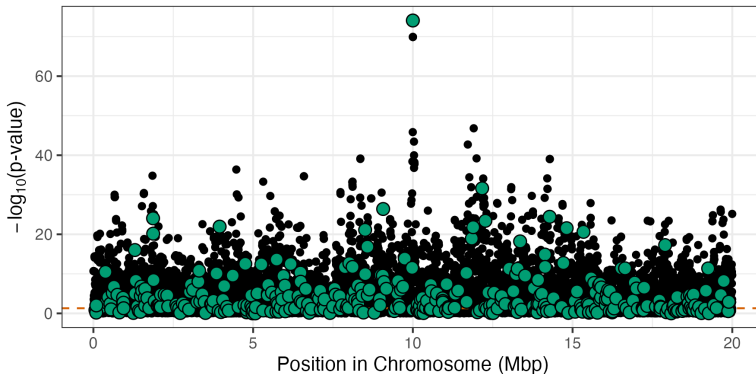
Now, let's look at each SNP across the whole genome

There's nothing special about $-log_{10}(p - value)$, it's just the $p - value$ of an association expressed in an easy to visualise way

Can anyone think of any problems with the approach we have taken so far?
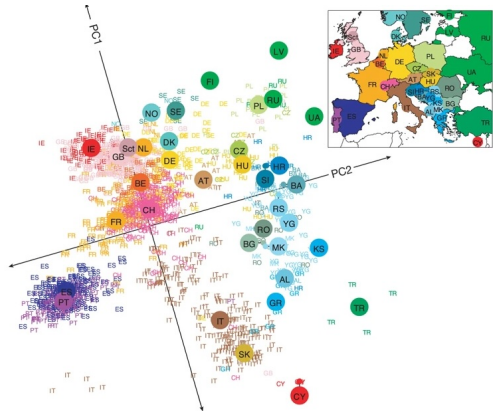
British people *really* like tea

Figure modified from: Novembre et al 2008 - *Nature*

## Association Genetics Caveat: Population Structure/Relatedness

British people *really* like tea

Genetic drift has led to slight differences in allele frequencies in the UK compared to other places

Could lead to genetic associations for tea preference

A relatedness structure among sampled individuals can lead to spurious genetic associations

---

Tea drinking in the UK has a clear cultural basis

Association Genetics Caveat: Population Structure/Relatedness
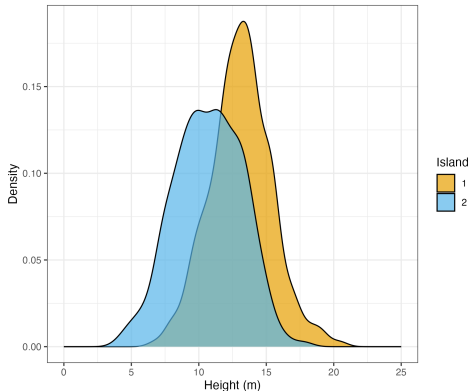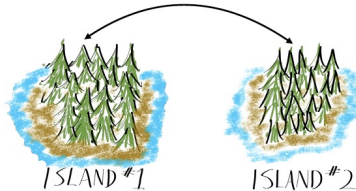
- In our model, trees are restricted to two distinct islands with a small amount of gene flow
- We can factor the relationship among individuals into our statistical model
  (*there are a variety of ways to do this, but we won't get into the specifics today*)

Association Genetics Caveat: Multiple Testing

Statisical analysis often relies on significance thresholds

In GWAS, we are conducting many tests (i.e. on each SNP) simultaneously

Association Genetics Caveat: Multiple Testing

Statisical analysis often relies on significance thresholds

In GWAS, we are conducting many tests (i.e. on each SNP) simultaneously

Let's do a simple experiment...

- Draw 100 numbers at random between 0 and 1, call this A

Association Genetics Caveat: Multiple Testing

Statisical analysis often relies on significance thresholds

In GWAS, we are conducting many tests (i.e. on each SNP) simultaneously

Let's do a simple experiment...

- Draw 100 numbers at random between 0 and 1, call this A

- Draw another 100 numbers at random between 0 and 1, call this B

Association Genetics Caveat: Multiple Testing

Statisical analysis often relies on significance thresholds

In GWAS, we are conducting many tests (i.e. on each SNP) simultaneously

Let's do a simple experiment...

- Draw 100 numbers at random between 0 and 1, call this A

- Draw another 100 numbers at random between 0 and 1, call this B

- Calculate the means of A and B

Association Genetics Caveat: Multiple Testing

Statisical analysis often relies on significance thresholds

In GWAS, we are conducting many tests (i.e. on each SNP) simultaneously

Let's do a simple experiment...

- Draw 100 numbers at random between 0 and 1, call this A

- Draw another 100 numbers at random between 0 and 1, call this B

- Calculate the means of A and B

- What is the expected difference between A and B?

Association Genetics Caveat: Multiple Testing

Statisical analysis often relies on significance thresholds

In GWAS, we are conducting many tests (i.e. on each SNP) simultaneously

Let's do a simple experiment...

- Draw 100 numbers at random between 0 and 1, call this A

- Draw another 100 numbers at random between 0 and 1, call this B

- Calculate the means of A and B

- What is the expected difference between A and B?

Genome-Wide Association Study - Redo

Let's now correct for the issues we have identified so far...