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

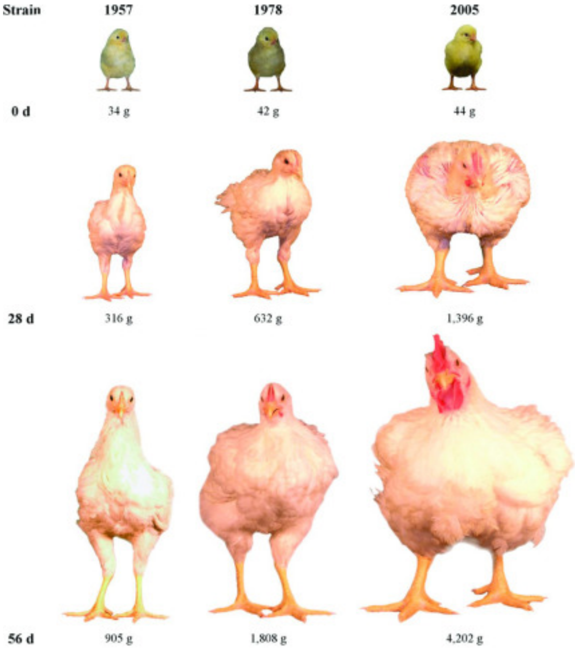
Module 3 Quantitative Genetics and 20th Century Breeding

Module 4 Advancing Forest Genetics with Recent Technology

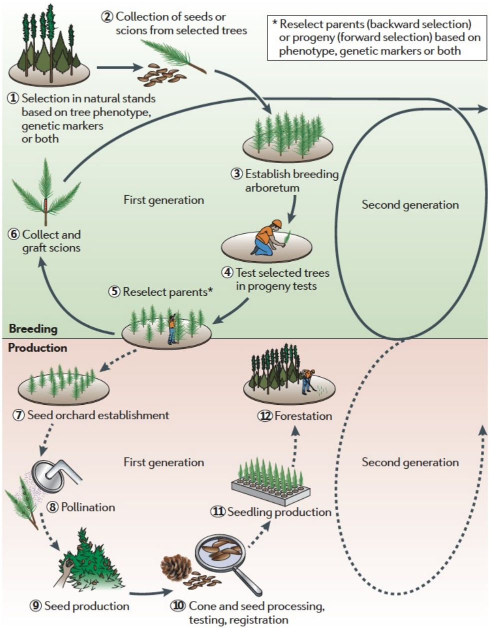
Selective breeding has been extraordinarily successful in human history

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!

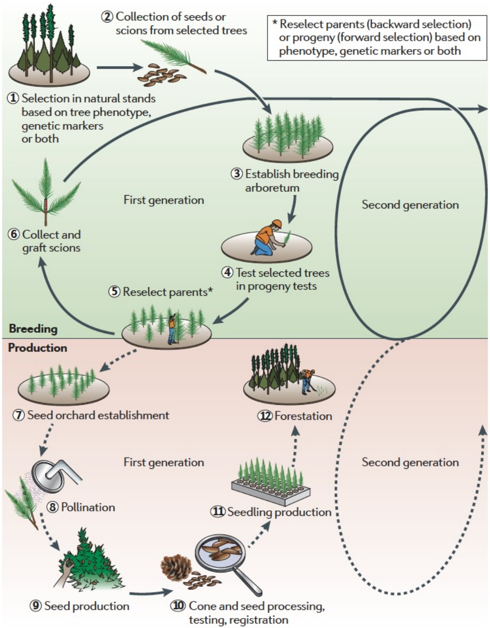


Modified from Figure 1 - Zuidhof et al. 2014



Conventional Tree Breeding

Image from Neale and Kremer 2011

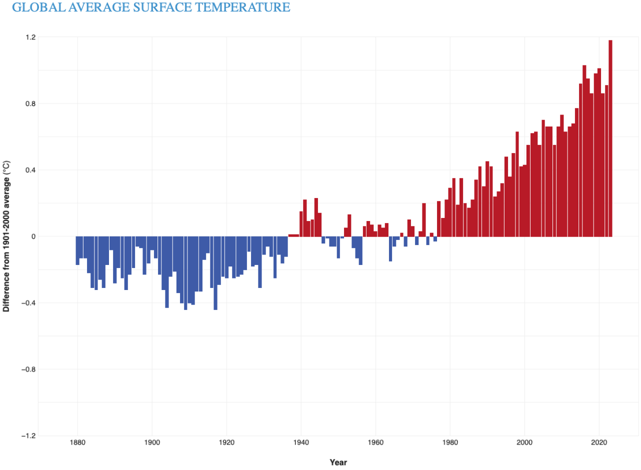


Conventional Tree Breeding

Each cycle takes from 20-30 years!

Image from Neale and Kremer 2011

Can We Afford to Take That Long?



https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature - *Hopefully the webpage still exists...*

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?

High Throughput Genotyping

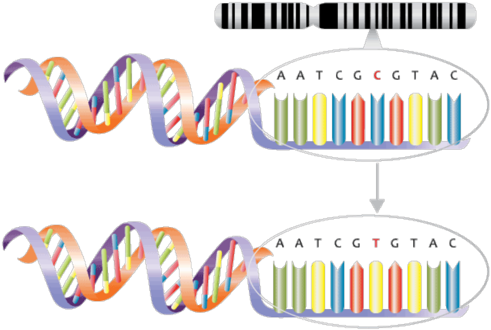
In Module 1, we discussed how we can use high-throughput sequencing methods to reconstruct the genome of an individual

But, if we are studying genetic variation we need data for numerous individuals

We disussed different types of genetic markers, but for the rest of this module, we are going to focus on SNPs as these are the main sort of data analysed these days

Recap - Single Nucleotide Polymorphism

A SNP (we often pronounce it as *snip*) is a DNA polymorphism at a particular base pair in the genome



Recap - Identifying SNPs

How can we identify SNPs?

- The first copy of the genome sequence can serve as a reference
- The fragment sequences from a new individual are compared to the reference to determine SNPs
- Technical error can also cause false SNPs, then multiple reads are required to remove errors

Reference	CCGTTAGAGT**T**ACAATTCGA
Read 2	TTAGAGT**A**ACAA
Read 3	CCGTTAGAGT**T**A
Read 4	T**T**ACAATTCGA
Read 5	GAGT**A**ACAA
Read 6	TTAGAGT**A**ACAAT

Recap - Using Genetic Markers

- Wang et al (2018 - Genome Biology) conducted a **genome-wide association study** (GWAS) to identify the genetic basi s of adaptation to photoperiod in European Aspen (*Populus tremula*)

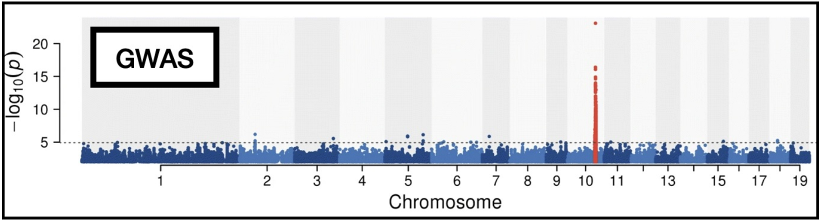
  They analysed > 4*million SNPs*

This slide also appeared in Lecture 1.6

Recap - Using Genetic Markers

- Wang et al (2018 - Genome Biology) conducted a **genome-wide association study** (GWAS) to identify the genetic basi s of adaptation to photoperiod in European Aspen (*Populus tremula*)

  They analysed > 4million SNPs



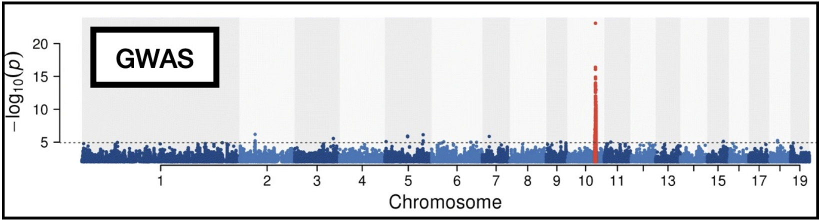
The genomic region highlighted by the red peak includes a gene that controls flowering time

This slide also appeared in Lecture 1.6

Recap - Using Genetic Markers

- Wang et al (2018 - Genome Biology) conducted a **genome-wide association study** (GWAS) to identify the genetic basi s of adaptation to photoperiod in European Aspen (*Populus tremula*)

  They analysed > 4million SNPs

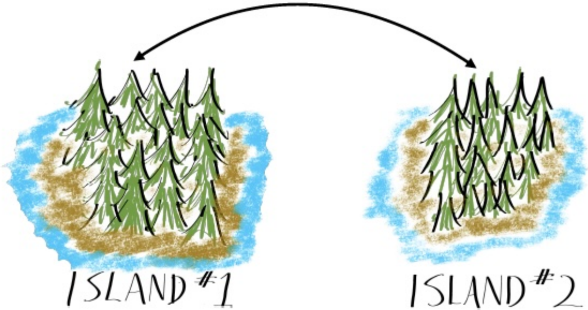


The genomic region highlighted by the red peak includes a gene that controls flowering time

BUT, there were more than 900 SNPs identified in the read peak that includes several different genes

This slide also appeared in Lecture 1.6

Let's Build a Model!



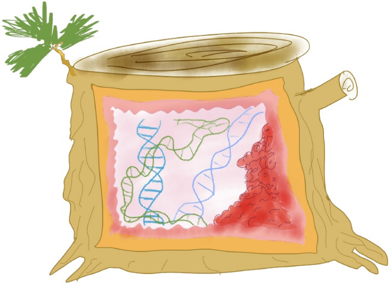
- Imagine a tree species inhabiting two islands
- Pollen flows between them

- 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

The Genetics of our Model

The Genetics of our Model



- Diploid individuals

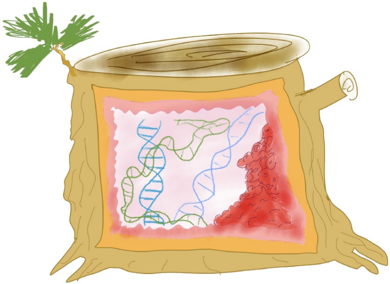
The details on this slide are just for understanding the model. You *will not* be tested on memorizing these numbers and/or parameters

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

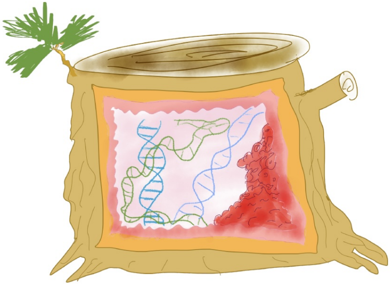
The Genetics of our Model



- 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

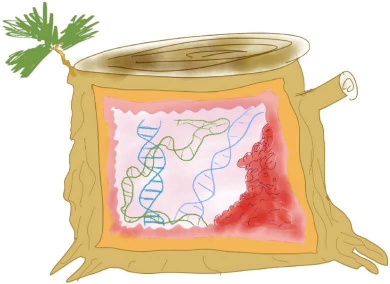
The Genetics of our Model



- 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

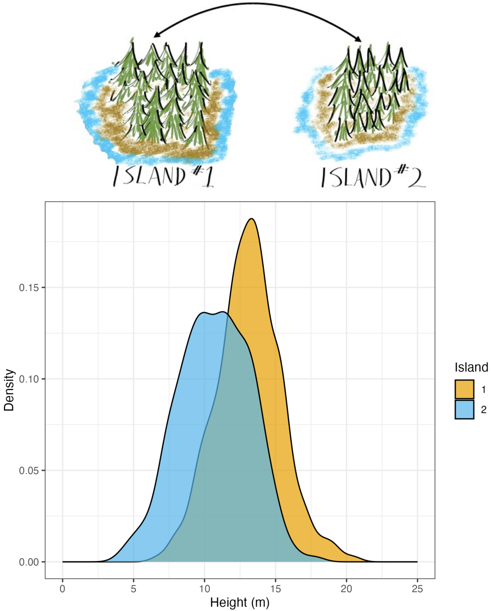
The Genetics of our Model



- 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

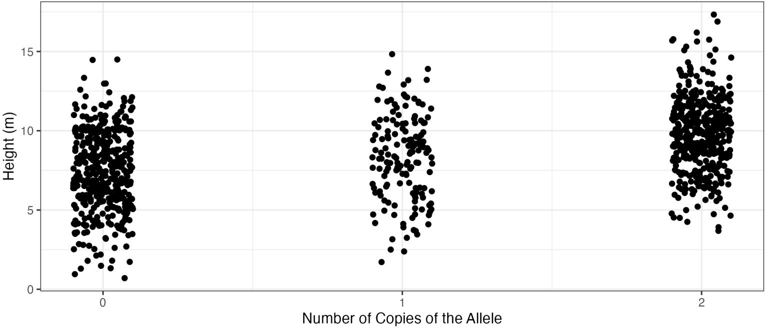
β: The effect of marker j on the trait

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

ϵ: The effect of the environment on the trait

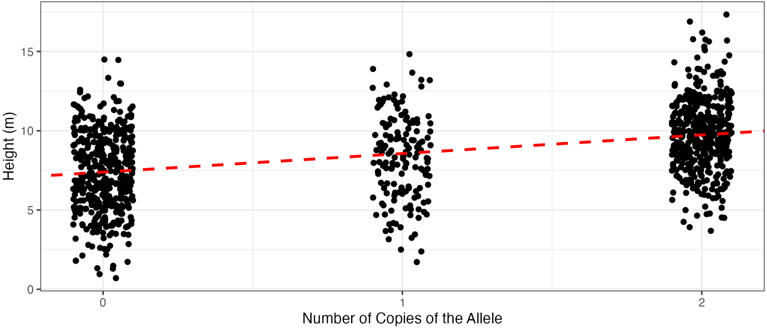
By fitting a regression to the data we can estimate the effect size of a particular marker on a trait and test for statistical significance...

Testing Association at a Single Marker



The allele is at a frequency of 0.16 across both islands

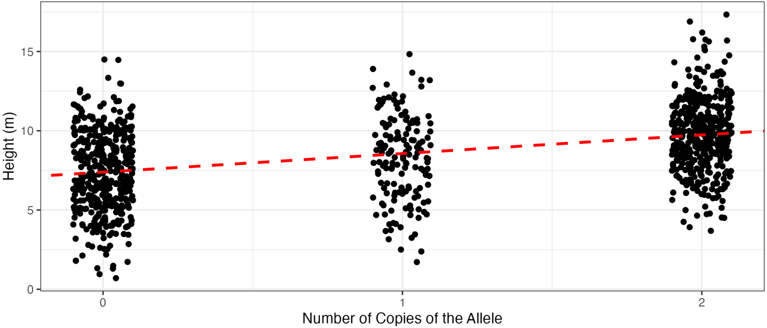
Testing Association at a Single Marker



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?

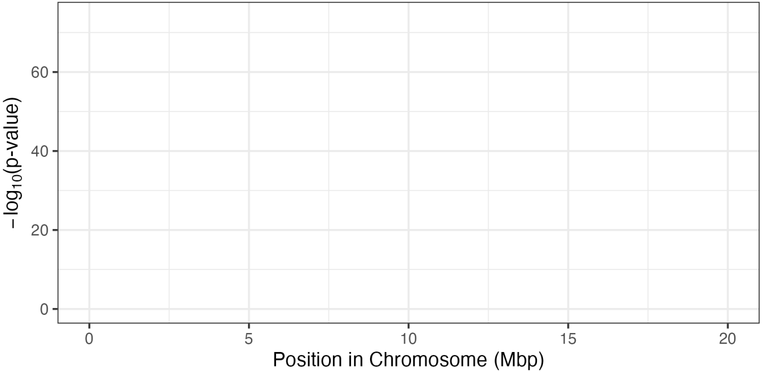
Genome-Wide Association Study

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

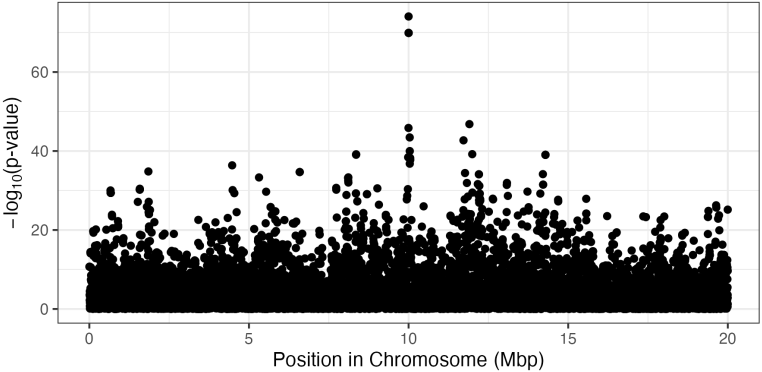
Now, let's look at each SNP across the whole genome



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

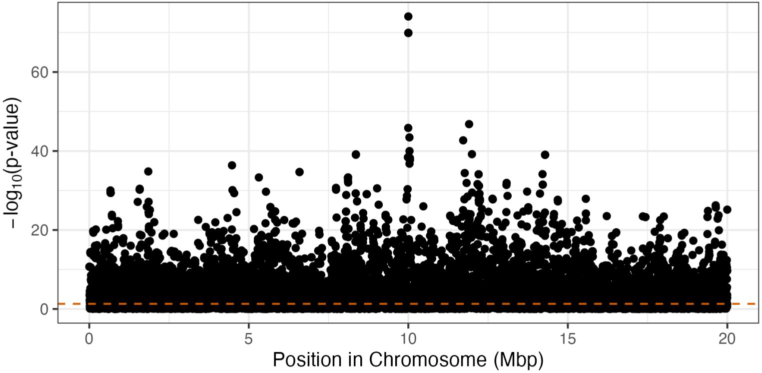
Now, let's look at each SNP across the whole genome



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

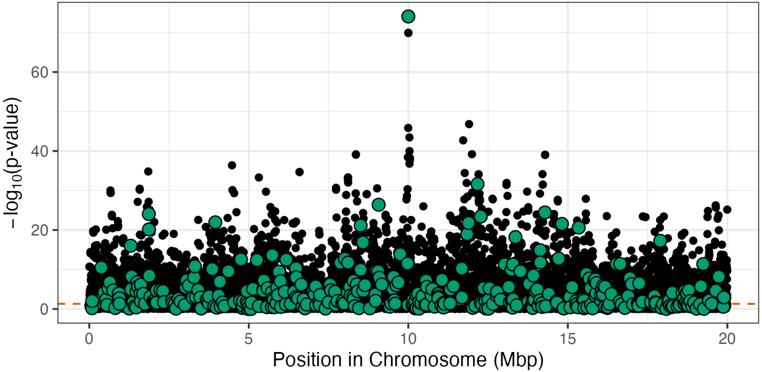
Genome-Wide Association Study

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

Can anyone think of any problems with the approach we have taken so far?

Association Genetics Caveat: Population Structure/Relatedness

British people *really* like tea

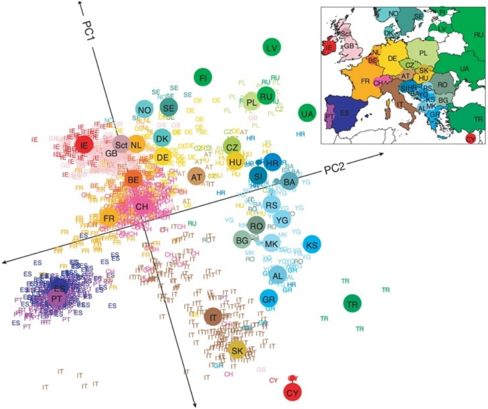


Figure modified from: Novembre et al 2008 - *Nature*

Association Genetics Caveat: Population Structure/Relatedness

British people *really* like tea

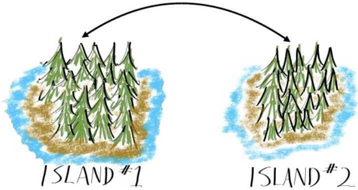
Genetic drift has led to slight differences in allele frequencies in the UK compared to other places

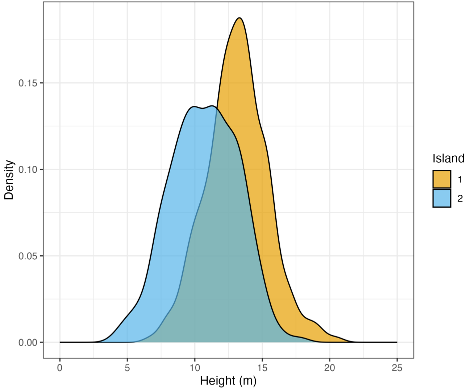
Could lead to genetic associations for tea preference

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*)

Genome-Wide Association Study

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

We corrected for population structure using a matrix of relatedness among individuals (see Lecture 4.3)

Association Genetics Caveat: Multiple Testing

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

Association Genetics Caveat: Multiple Testing

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

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

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

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

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...