FRST302: Forest Genetics

Lecture 1.6: Identifying Genetic Variation

Lecture 1.5 - Recap

- Describe gene structure
- The various roles of RNA
- Describe gene expression
- Identifying functional regions in a genome

Terminology Check

Genome - the complete set of genetic information for an organism

Transcriptome - the set of all RNA molecules, including mRNA, tRNA, and rRNA produced by the genome of an organism or a cell

***Proteome** – the set of all proteins produced by an organism or a cell

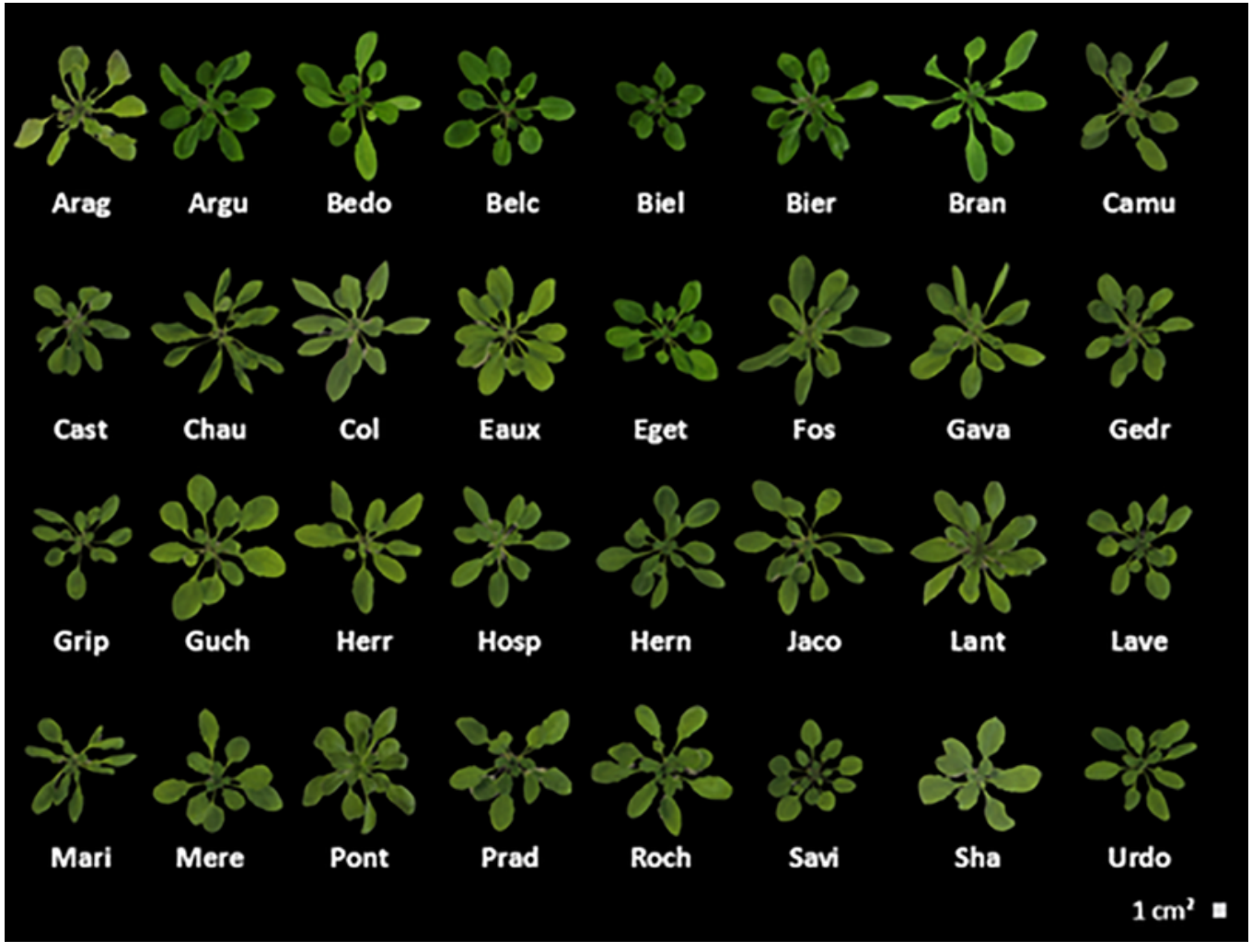
***Exome** - the subset of the genome that encodes for proteins

* We haven't mentioned these up til now

Learning Outcomes

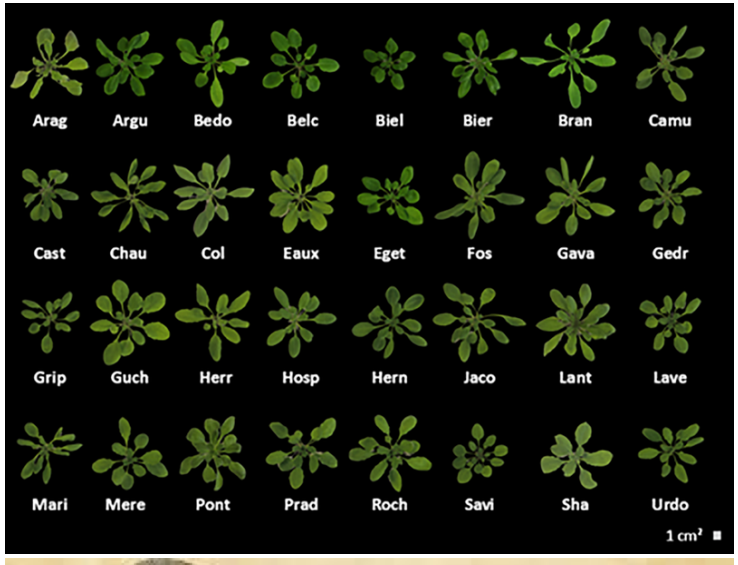
- The importance of genetic variation
- Understand phenotypic variation and its connection to genetic variation
- Different ways to study genetic variation
- Single Nucleotide Polymorphisms (SNPs) - detection, use and limitations

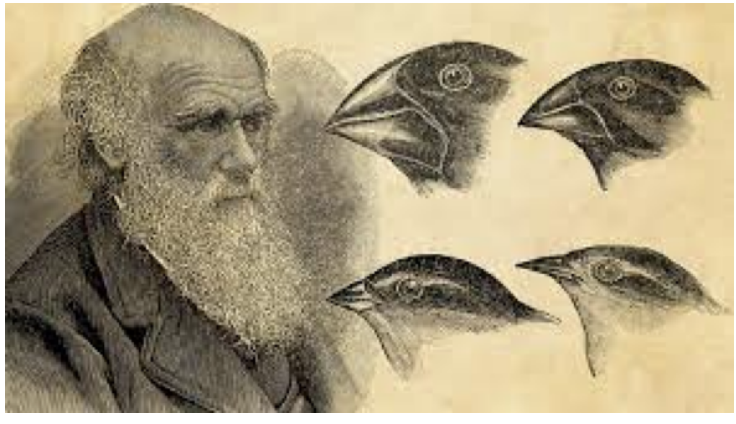
The Importance of Variation



Variation in rosettes of *Arabidopsis thaliana* individuals grown under controlled conditions in a growth chamber; Modified from Duruflé et al 2019 - *Front Plant Sci.*

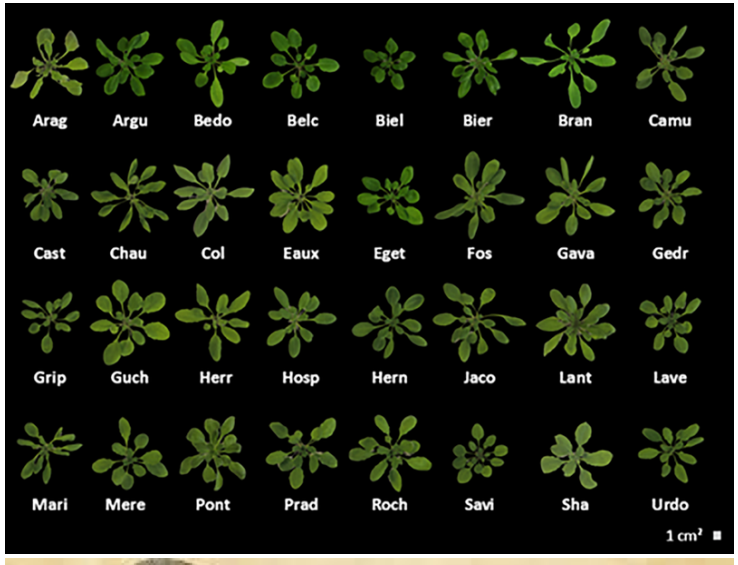
The Importance of Variation

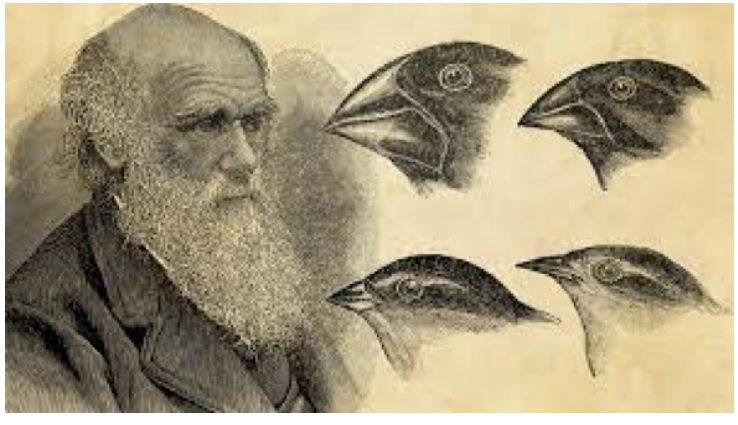




– Genetic variation is the foundation of evolution

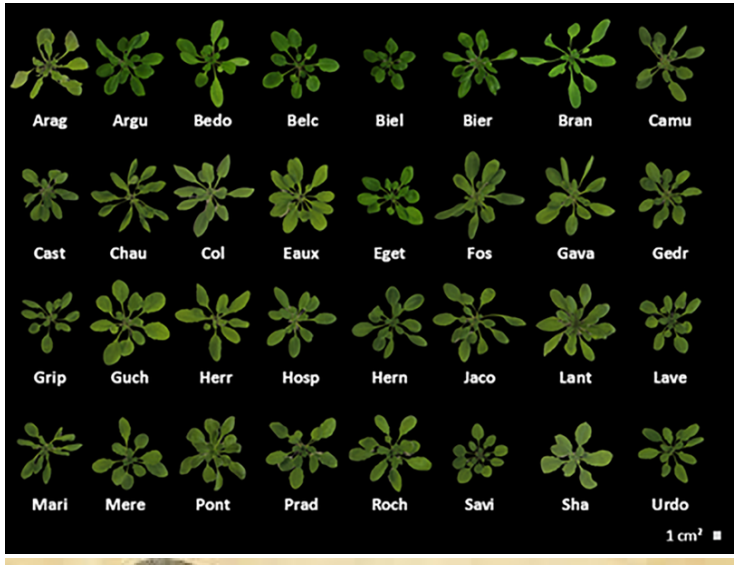
The Importance of Variation

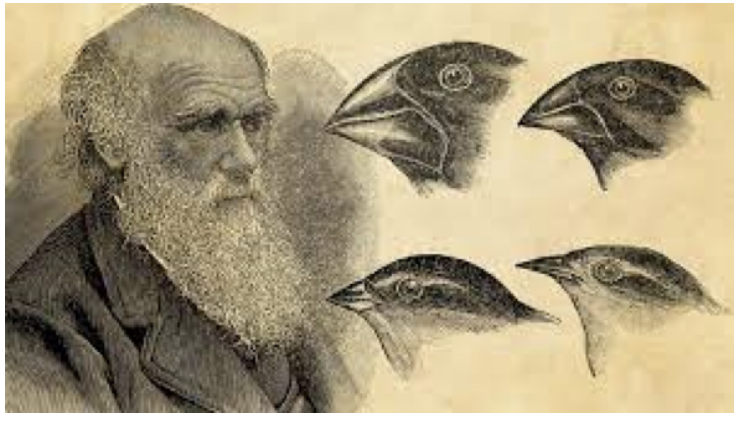




- Genetic variation is the foundation of evolution
- It predicts the response to selection and breeding
  Tongli will greatly expand on this

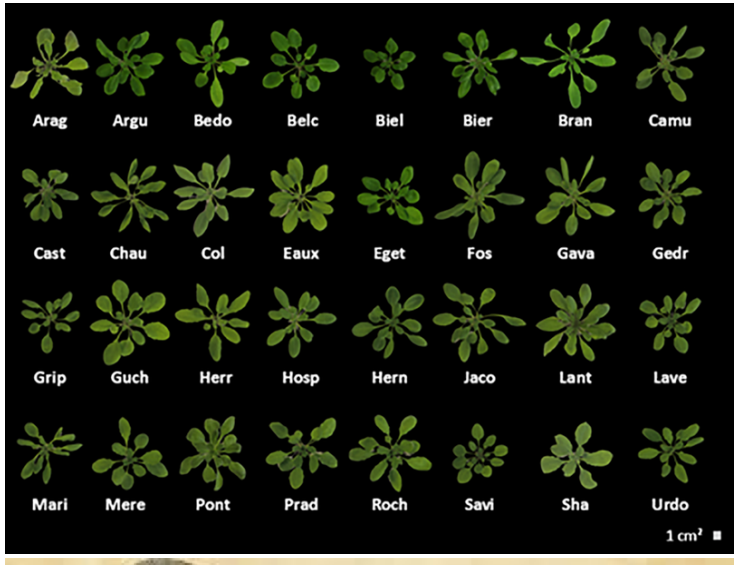
The Importance of Variation

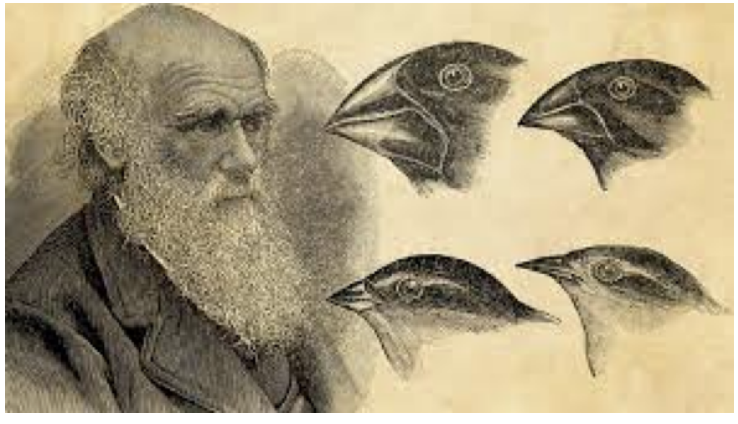




- Genetic variation is the foundation of evolution
- It predicts the response to selection and breeding
  Tongli will greatly expand on this
- Preservation of genetic diversity is a major goal of conservation

The Importance of Variation





- Genetic variation is the foundation of evolution
- It predicts the response to selection and breeding
  Tongli will greatly expand on this
- Preservation of genetic diversity is a major goal of conservation

We can study genetic variation by examining patterns of phenotypic variation
(e.g. within families - *we've already done this*!)

Common Modes of Trait Variation

Continuous phenotypic variation - traits measured on a numerical scale (e.g. height, diameter, chlorophyll fluorescence)

These trees are from the same origin and were grown in the same environment!

Common Modes of Trait Variation

Discrete phenotypic variation - traits that exhibit categorical differences

Common Modes of Trait Variation

Ordinal phenotypic variation - discrete traits with some informative order

Resistance to Dothistroma needle blight

Difficulties Working With Phenotypes

What are some difficulties you might face when using phenotypic variation to study genetic variation?

Difficulties Working With Phenotypes

What are some difficulties you might face when using phenotypic variation to study genetic variation?

$Phenotype = Genotype + Environment$

Different genotypes may express the same or similar phenotypes

Phenotypic plasticity[*See the assigned reading!*]

Time consuming

Am I measuring the right thing?

Difficulties Working With Phenotypes

What are some difficulties you might face when using phenotypic variation to study genetic variation?

$Phenotype = Genotype + Environment$

Different genotypes may express the same or similar phenotypes

Phenotypic plasticity[*See the assigned reading!*]

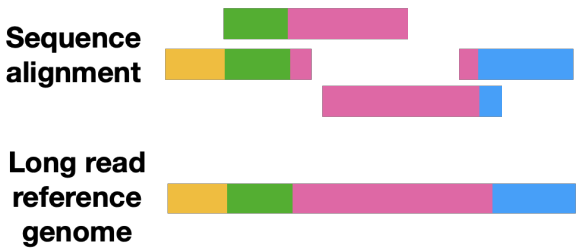
Time consuming

Am I measuring the right thing?

Directly working with genetic material is desirable!

Now What!?

We've built a reference genome, we have identifed genes and gene regulatory regions



How are we going to study genetic variation in our species of interest?

Types of Genetic Variation

Based on what we've covered in the past lectures, how could we detect genetic variation?

This is **not** *an exhaustive list!*

Types of Genetic Variation

Based on what we've covered in the past lectures, how could we detect genetic variation?

1. Differences in gene products (e.g. nonsynonymous point mutations)
2. Variation in RNA abundances (e.g. changes to the architechture of gene regulation)
3. Differences in genome structure (e.g. chromosomal rearrangements)
4. Differences in DNA sequences (e.g. nucleotide substitutions)

This is **not** *an exhaustive list!*

1. Differences in Gene Products

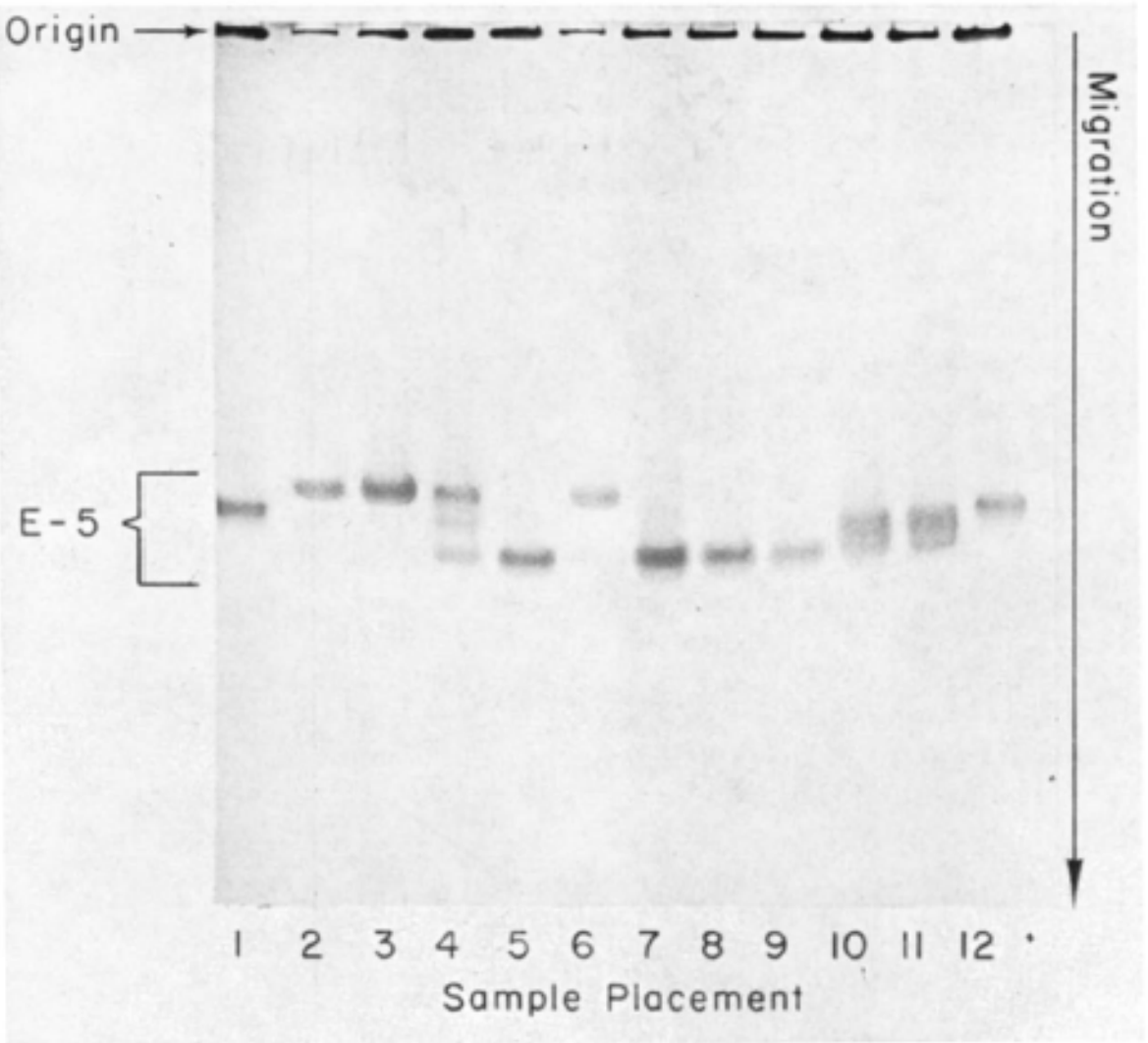
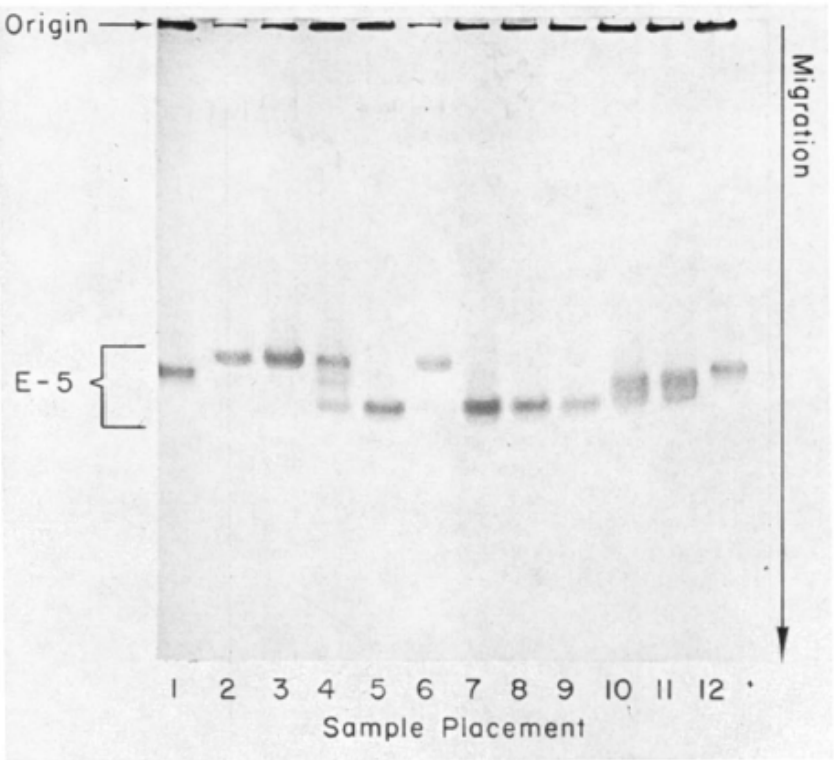


Figure from Lewontin and Hubby 1967 - *Genetics* - One of the most important papers in genetic history!

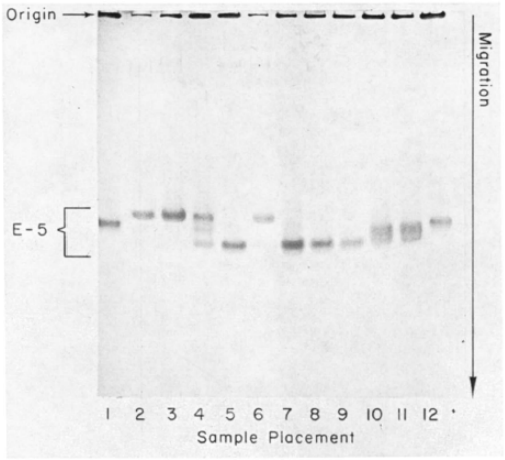
1. Differences in Gene Products



- Extract specific enzymes and run them in starch gels
- Ectrophoretically polymorphic enzymes, or **allozymes**, separate on the gels
- Different alleles and genotypes can readily by seen from the gel itself
- Widely used from the 1970s-90s

Figure from Lewontin and Hubby 1967 - *Genetics* - One of the most important papers in genetic history!

Different Approaches



Allozymes

Pros

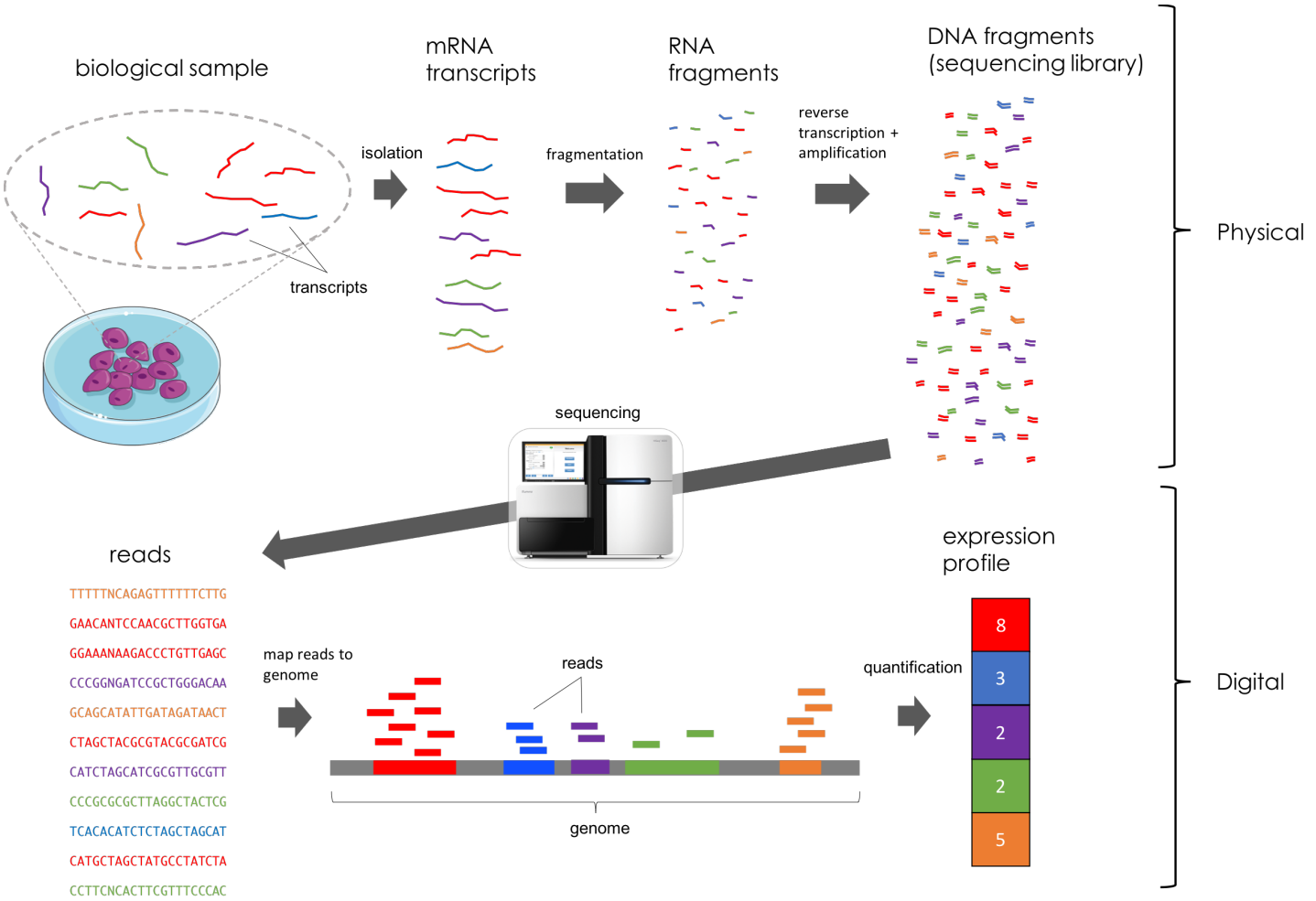
- Easy sample prep
- No computation
- Cheap, no specialized equipment
- ...

Cons

- Hazardous - toxic chemicals
- Very low throughput
- ...

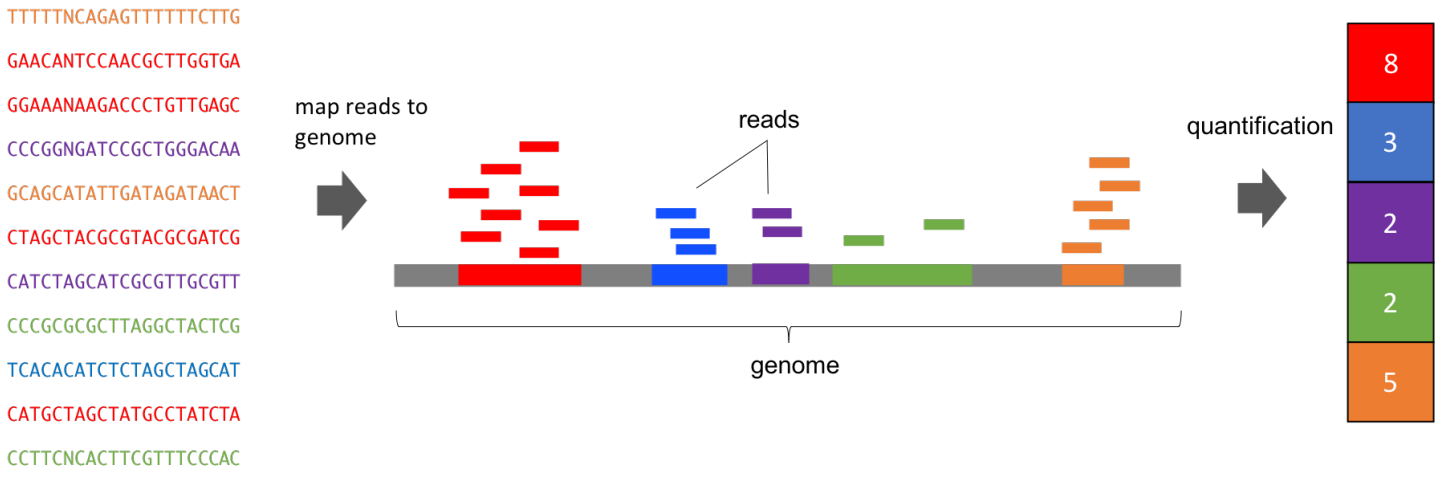
2. Variation in RNA Abundance

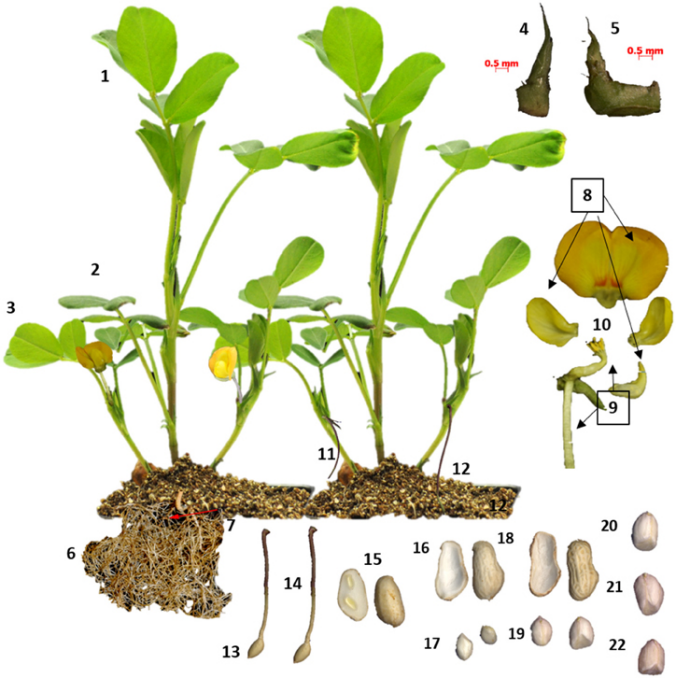
Differences in gene regulation are likely a major driver of phenotypic differences



In maize, *≈60% of variation in agriculturally important traits can be linked to variation in gene regulation*

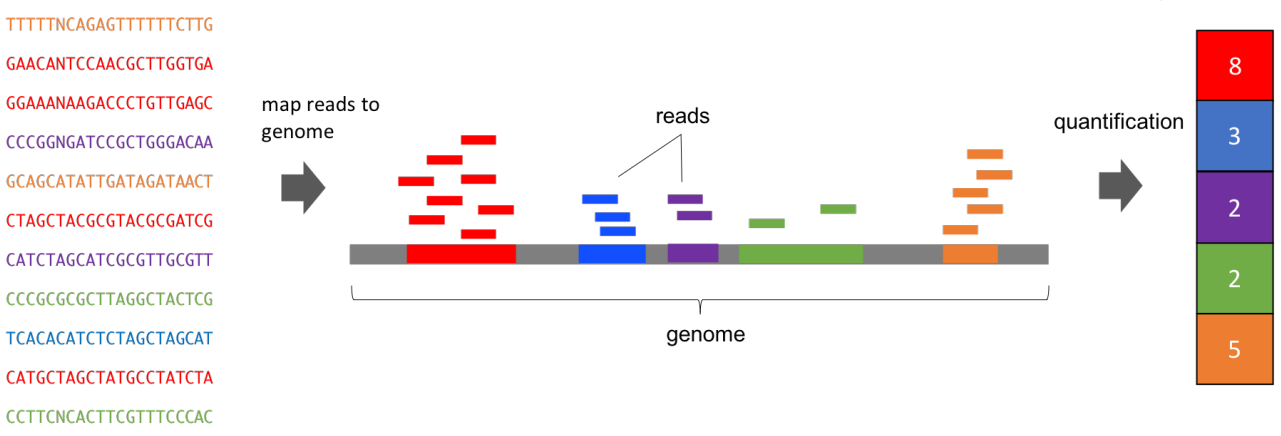
2. Variation in RNA Abundance





- Compare expression of different transcripts among multiple individuals/populations
- Hypotheses regarding particular tissues and parts of the lifecycle can be devised and tested
- Genes/transcripts that show highly variable levels of expression may contribute to phenotypic differences

Different Approaches



Differentially Expressed Genes with RNA-Seq

Pros

- Hypothesis driven
- Biologically interpretable
- Captures sequence variation too
- ...

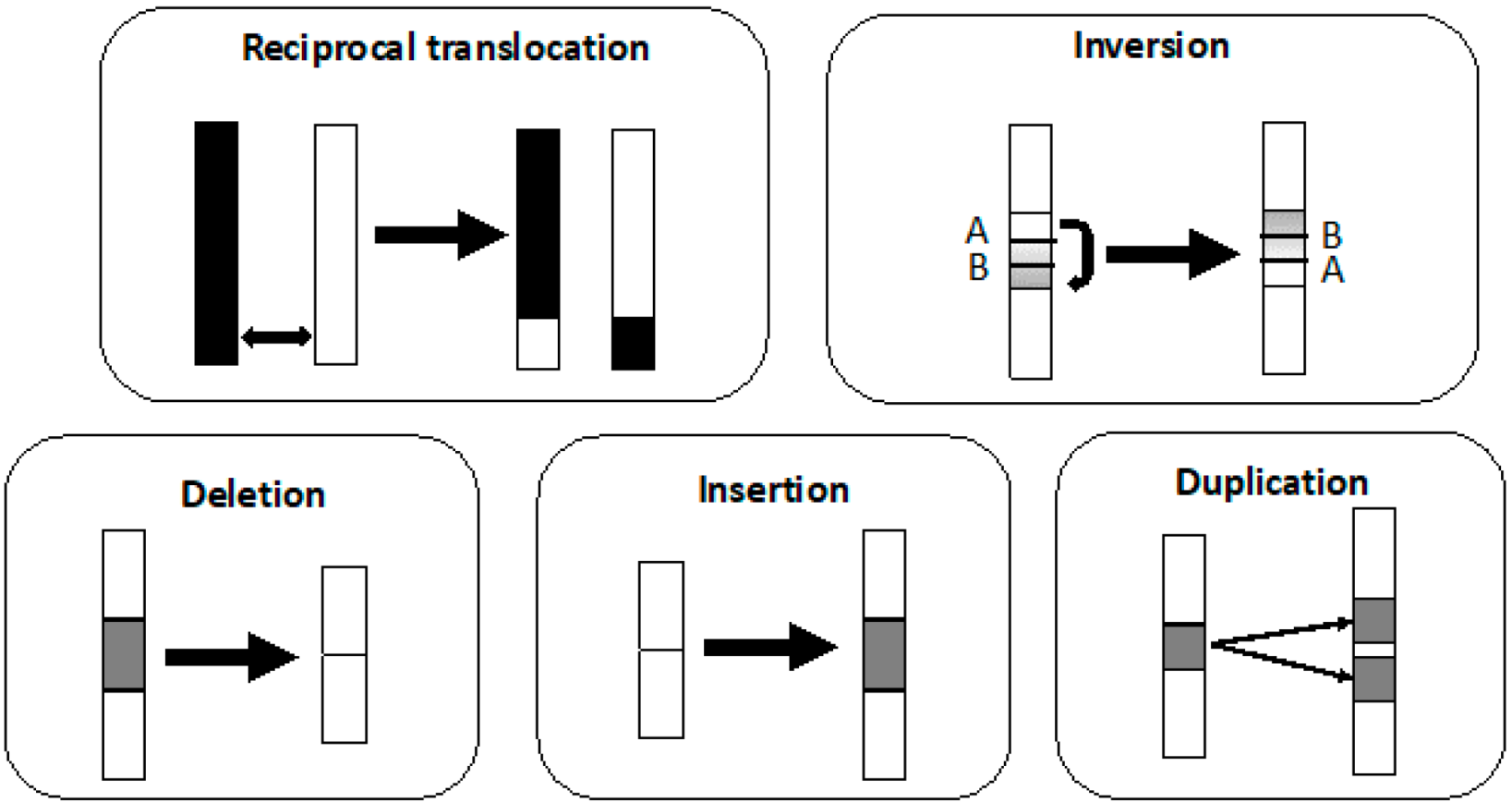
Cons

- Extremely large search space (e.g. tissues, life stages, environments)
- Statistically finicky
- ...

3. Differences in Genome Structure

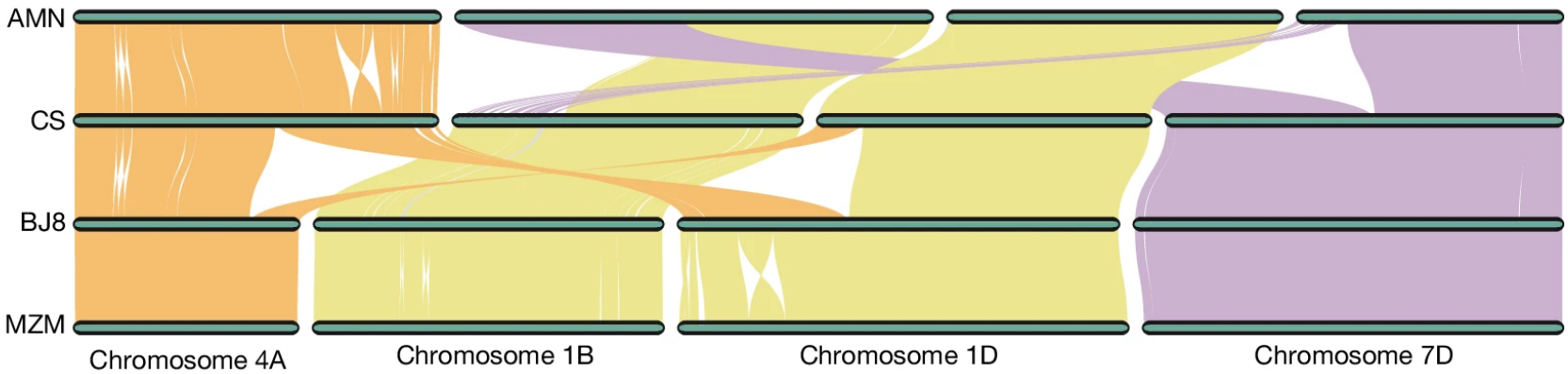
Variation in genome structure is *relatively* common

Can range from large scale (differences in chromosome structure) to small scale (insertion/deletion of a single basepair)



3. Differences in Genome Structure

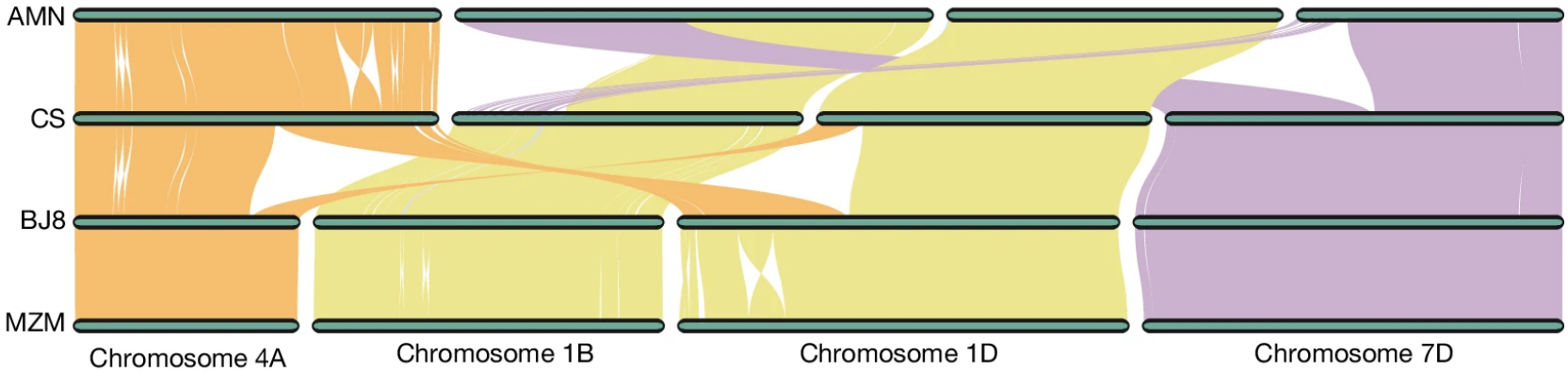
Can be identifed by assembling and comparing whole genomes



Bread wheat is hexaploid with a genome size of <3Gbp

3. Differences in Genome Structure

Can be identifed by assembling and comparing whole genomes

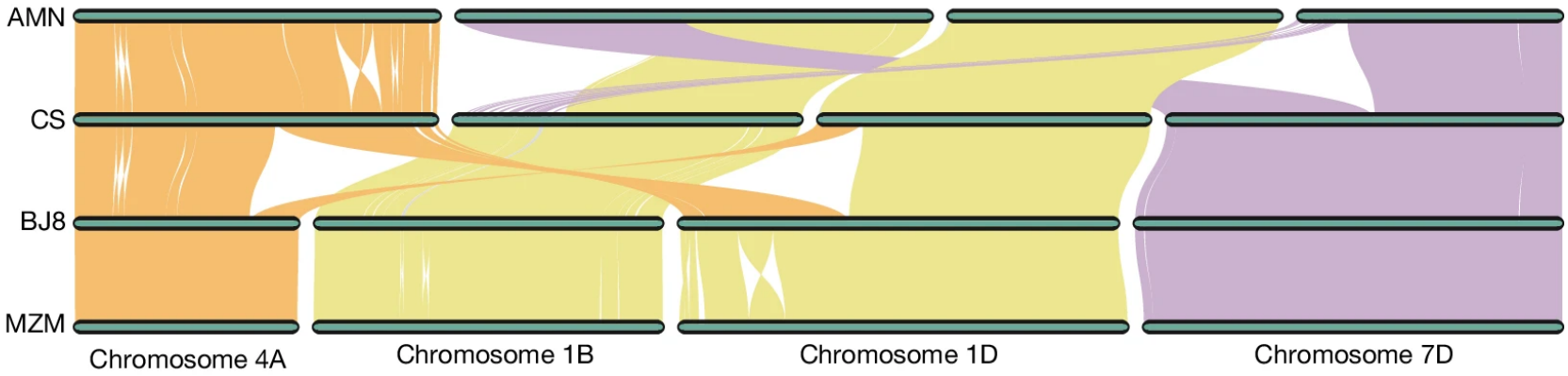


E.g. by assembling the genomes of 17 individuals of wheat, Jiao et al (2025 - Science) found >120,000 structural variants individually >5kbp - that's roughly 600Mbp of difference in genome structure

Demonstrated that the development of Winter wheat was via gene duplication!

Bread wheat is hexaploid with a genome size of <3Gbp

Different Approaches



Assembling and comparing whole genomes

Pros

- *This is the future of genetics*
- Potentially identifies all genetic variation
- Can get closer to causality
- ...

Cons

- Currently prohibitively expensive
- Does not solve statistical inherent in genetic analysis
- Intense computation
- ...

Questions?

Questions?

Let's take a short break

Sample Exam Questions

Sample Exam Questions

Under pure blending inheritance an individual's trait is simply the arithmetic average of its parents trait values. By what fraction would you expect variation in that trait to decrease in the population each generation?

A ...

B ...

C ...

D ...

Sample Exam Questions

Imagine Mendel had studied additional traits of his peas. Which of the following hypothetical traits would have made it difficult for him to formulate his laws:

A ...

B ...

C ...

D ...

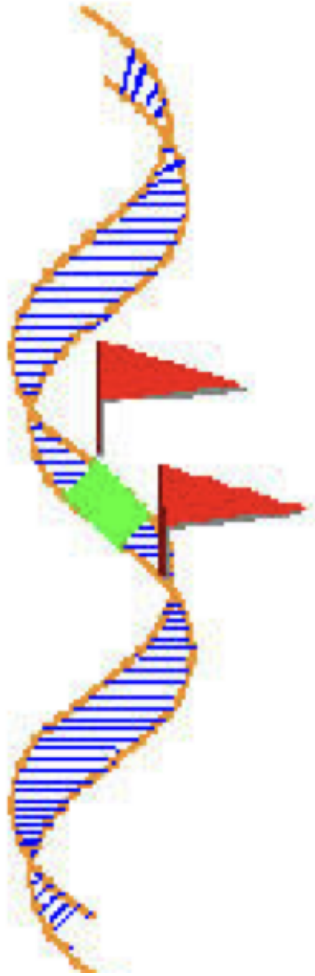
4. Differences in DNA Sequences

Genetic Markers

A **genetic marker** is a gene/gene product or a DNA sequence that can be used to:

- Identify individuals or species
- Determine relationships among them
- Locate known genes
- Study genetic variation

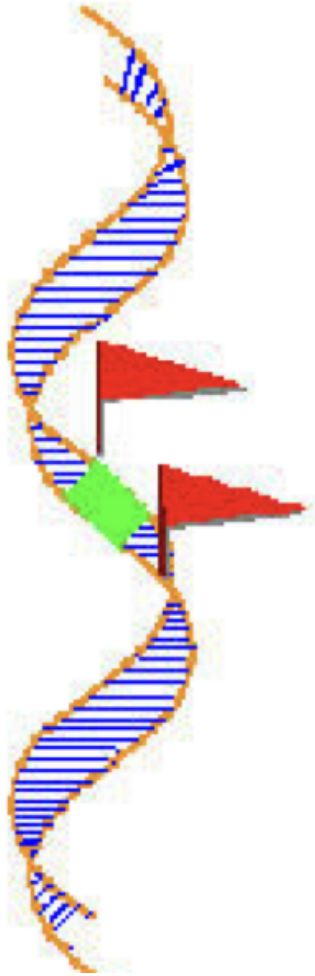
The development of genetic markers represents a major acheivement for biology in the 20th century



Genetic Markers

Desirable features of genetic markers:

- Easily distinguish hetero- and homozygotes
- HIghly polymorphic
- Reproducible/reliable
- Distributed throughout the genome
- Fast and inexpensive
- Non-invasive/non-destructive collection of data



Different Types of DNA Markers

- RFLP - Restriction Fragment Length Polymorphism
- RAPD - Random Amplified Polymorphic DNA
- AFLP - Amplified Fragment Length Polymorphism

  These have been largely superceded by DNA sequencing methods

- SSR, Short Simple Repeats (also called Microsatellites)

  These are not really used in high throughput studies

- SNPs - Single Nucleotide Polymorphisms

  The sky is the limit!

Different Types of DNA Markers

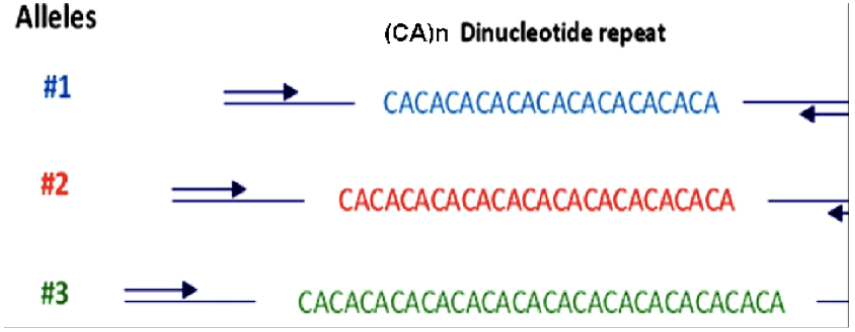
- RFLP - Restriction Fragment Length Polymorphism
- RAPD - Random Amplified Polymorphic DNA
- AFLP - Amplified Fragment Length Polymorphism

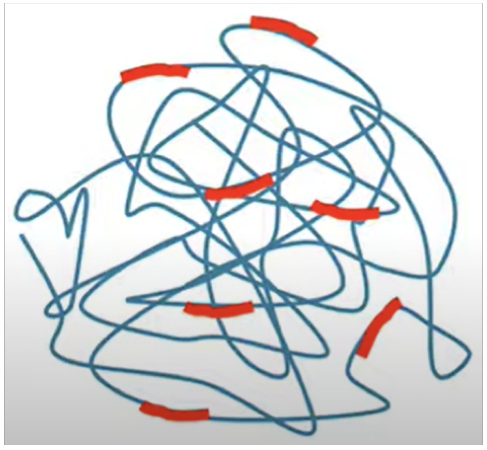
All these methods involve examining the patterns of variation observed on gels (i.e. not high throughput)
The data that arises from these methods can be hard to relate to phenotypes

Now that DNA sequencing is so cheap, these are kind of obsolete in forest genetics

Different Types of DNA Markers - Short Simple Repeats

- SSRs – also called microsatellites or short tandem repeats (STRs)
- Usually about 1-10 bps
- Distributed at thousands of locations over the genome
- High mutation rates (so very polymorphic)

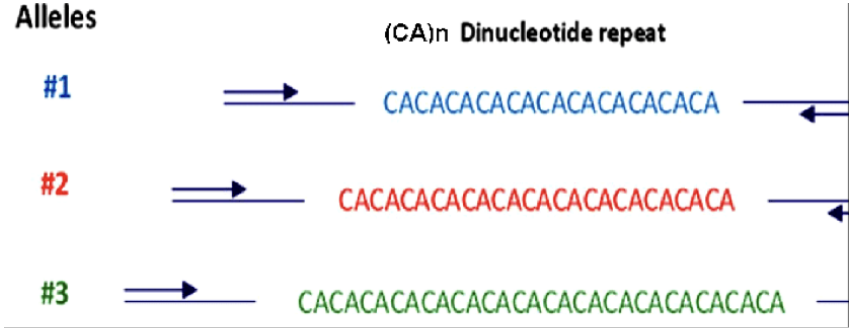


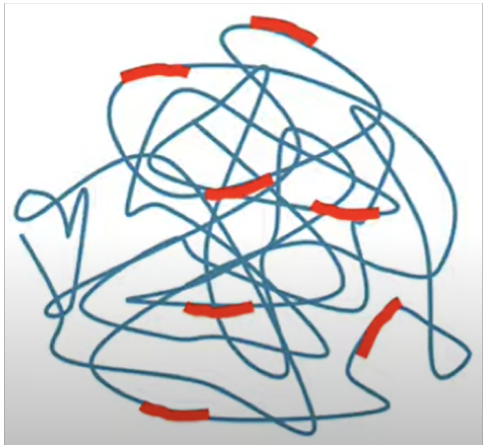


Different Types of DNA Markers - Short Simple Repeats

Cons

- Requires specific primers for target species
- Primer design is labour intensive
- Null alleles - if primer fails to bind a sample can appear homozygous

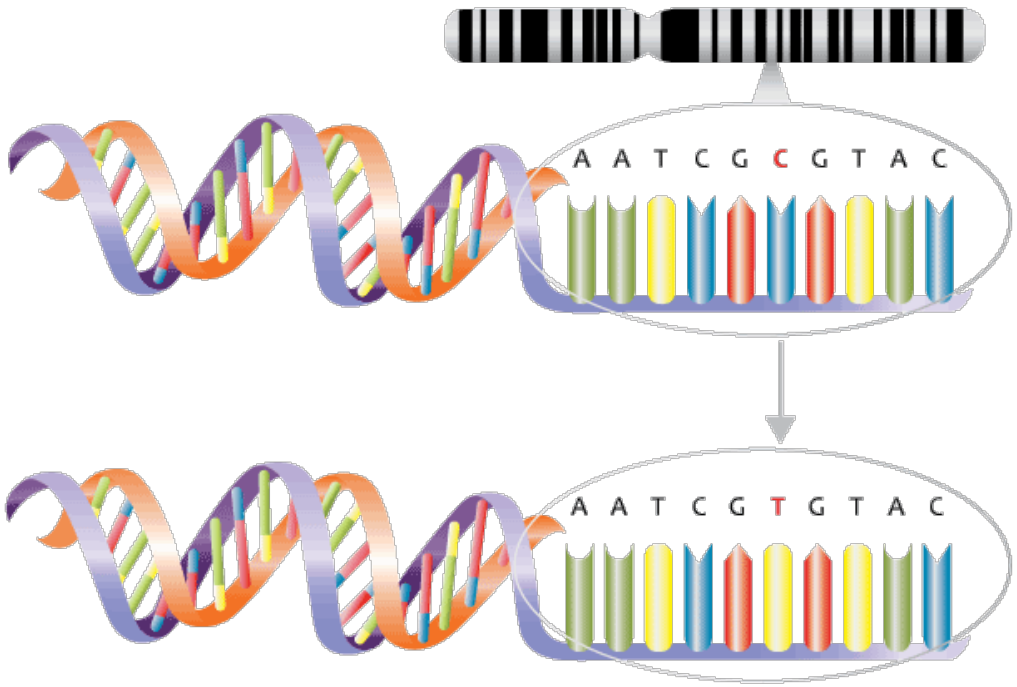




Now that DNA sequencing is so cheap, these are less widely used than previously, but still useful in non-model systems

Single Nucleotide Polymorphism

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

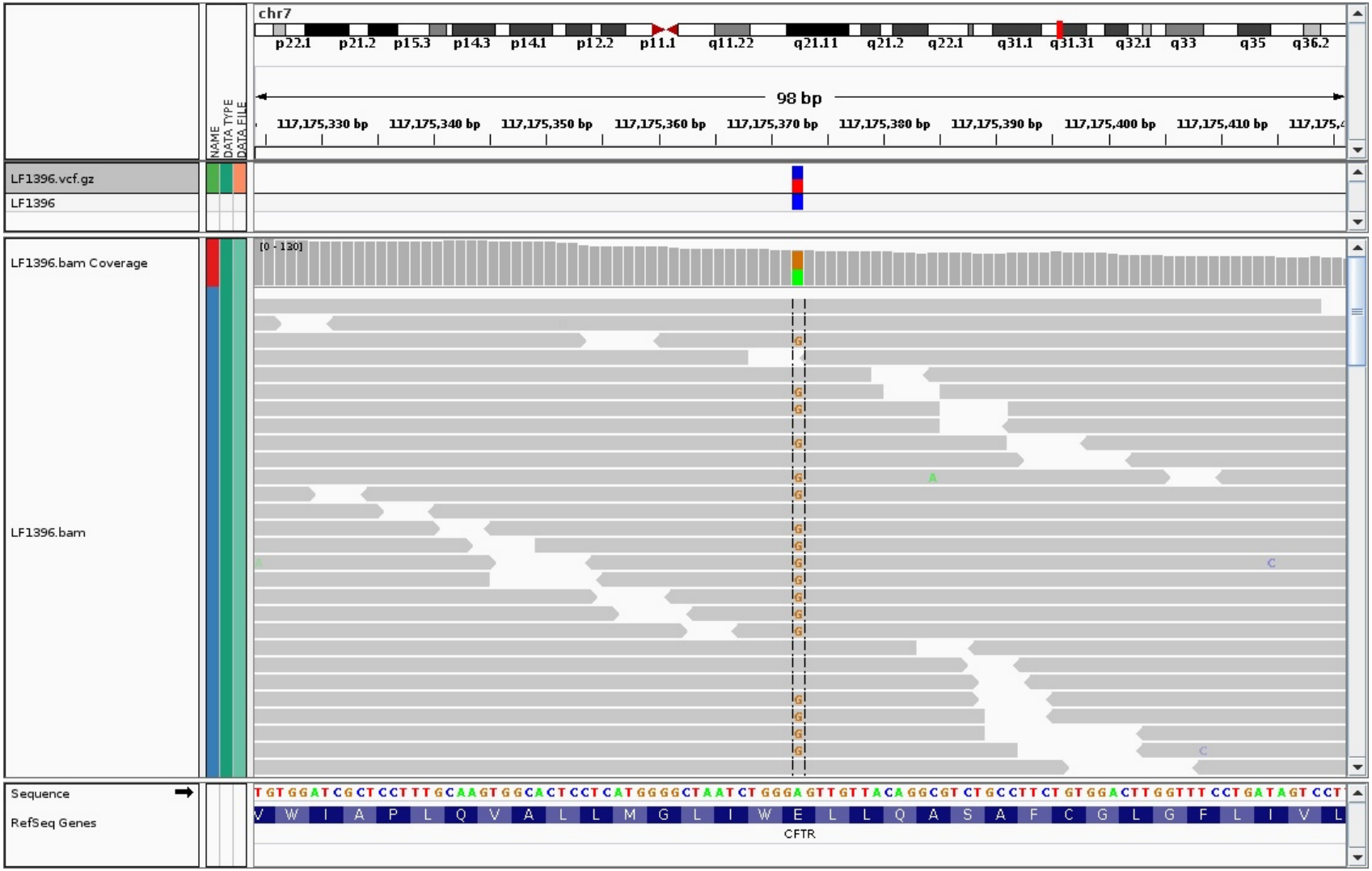


Identifying SNPs

How can we identify SNPs?

Identifying SNPs

How can we identify SNPs?



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

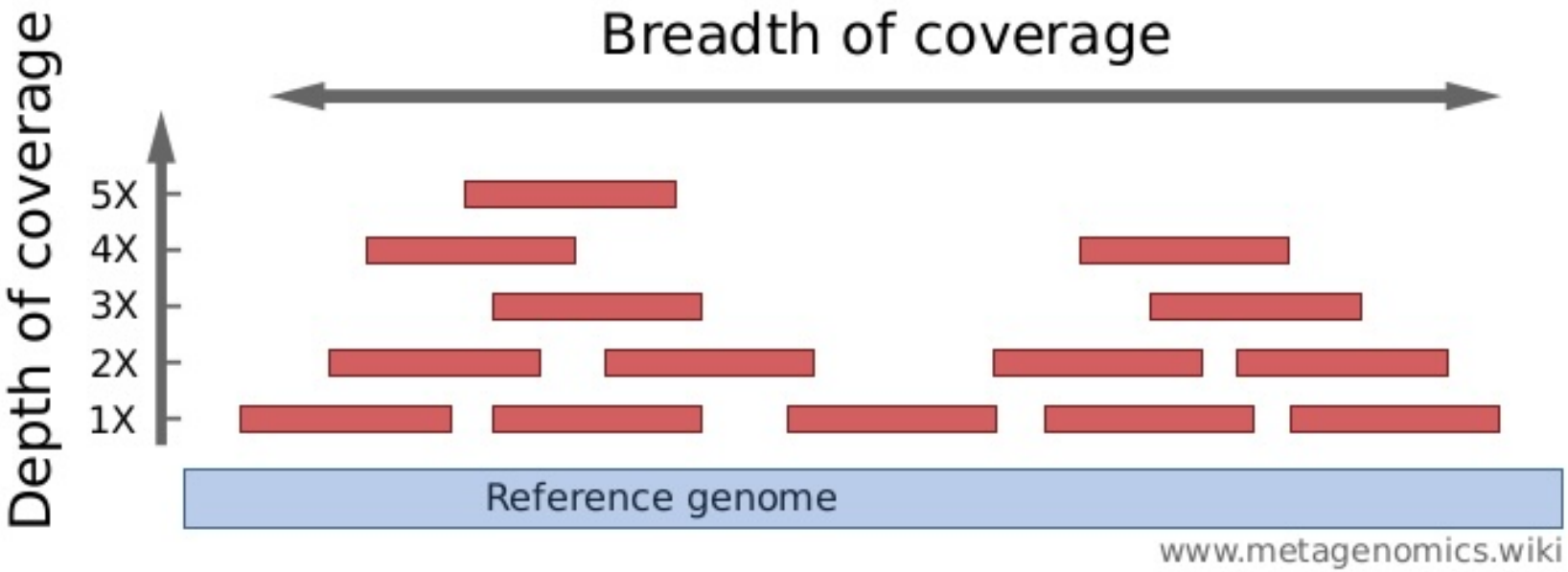
Repeat for as many individuals as you need/can afford!

Genome Re-sequencing

We usually refer to the sequencing individuals after building a reference genome as **resequencing**

There are many, many different ways to prepare the DNA for resequencing (we'll touch on these in Module 4)

All high-throughput methods require the user to think about **sequencing coverage**



SNPs as Markers

There can never be more than 4 alleles (i.e. A, T, C or G)

Most SNPs only exhibit rwo alleles

SNPs occur throughout the genome

Easy to build statistical models to study them and relate them to phenotypes

SNPs occur throughout the genome

Different Approaches

SNPs

Pros

- Easy to model them statistically
- Millions of them in typical eukaryotic genomes (often found at around 0.1% of sites)
- ...

Cons

- Lots of computation /bioinformatics to identify them
- Hard to link to function in many cases
- Many statistical issues
- ...

Analysis of SNPs is now a cornerstone of modern genomics (*not just in trees!*)

Types of Genetic Variation

Remember all the different things that we may use genomics for in forestry? Think about how you could tackle them with the different techniques we have discussed today.

1. Differences in gene products (identified with allozymes)
2. Variation in RNA abundances (identified with RNA-seq)
3. Differences in genome structure (identified with whole genome alignments)
4. Differences in DNA sequences (identified by fragment analysis or re-sequencing)

This is **not** *an exhaustive list!*

Questions?

Questions?

Let's take a short break

Using Genetic Markers

- We may be interested in identifying the genetic basis of an important trait
- By looking for SNPs that have a particularly strong association with trait variation we may find important genes (e.g. targets for transformation etc.)

Using Genetic Markers

- We may be interested in identifying the genetic basis of an important trait
- By looking for SNPs that have a particularly strong association with trait variation we may find important genes (e.g. targets for transformation etc.)

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

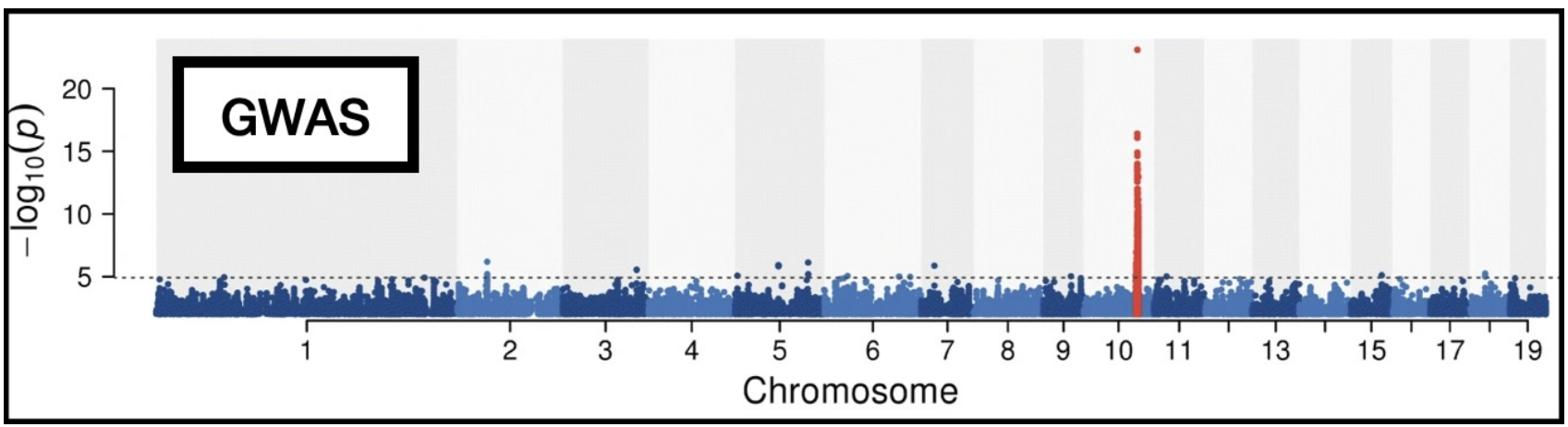
  They analysed $>$ *4million SNPs*

Using Genetic Markers

- We may be interested in identifying the genetic basis of an important trait
- By looking for SNPs that have a particularly strong association with trait variation we may find important genes (e.g. targets for transformation etc.)

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

  They analysed $>$ *4million SNPs*



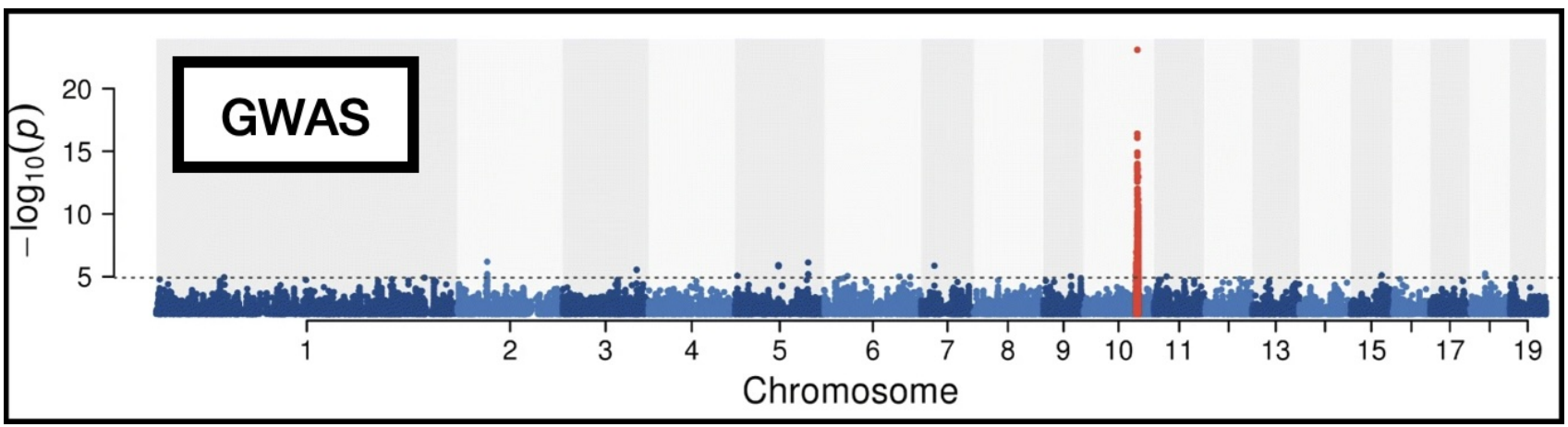
The genomic region highlighted by the red peak includes a gene that controls flowering time

Using Genetic Markers

- We may be interested in identifying the genetic basis of an important trait
- By looking for SNPs that have a particularly strong association with trait variation we may find important genes (e.g. targets for transformation etc.)

- Wang et al (2018 - Genome Biology) conducted a **genome-wide association study** (GWAS) to identify the genetic basis 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

Statistical Issues in Genetic Analysis

There is much more that can be said that would fit in a single course, let alone a lecture slide, so we'll return to this in Module 4

For the time being, consider the following issues with genetic analysis:

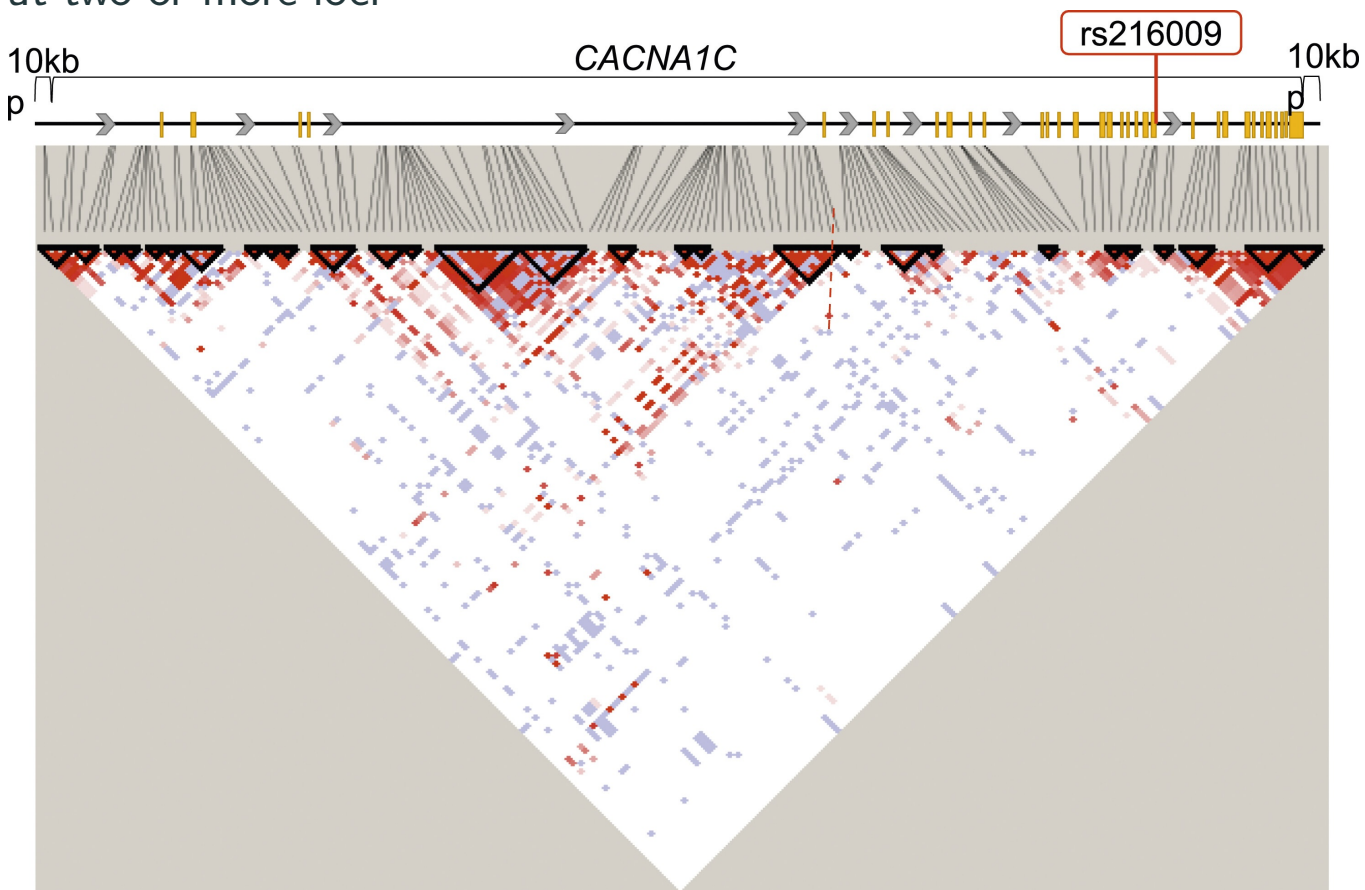
- Correlation $\neq$ Causation
- Small effect loci require extremely large samples to adequately analyse
- Linkage and Linkage Disequilibrium

Linkage Disequilibrium

Due to genetic linkage, sites in close proximity on chromosomes will often be inherited together

This generates a statistical relationship between the patterns of variation seen at different sites in the genome

We use the term **linkage disequilibrium** to refer to non-random associations of alleles at two or more loci



LD Introduces statistical non-independence into our analyses that we need to think

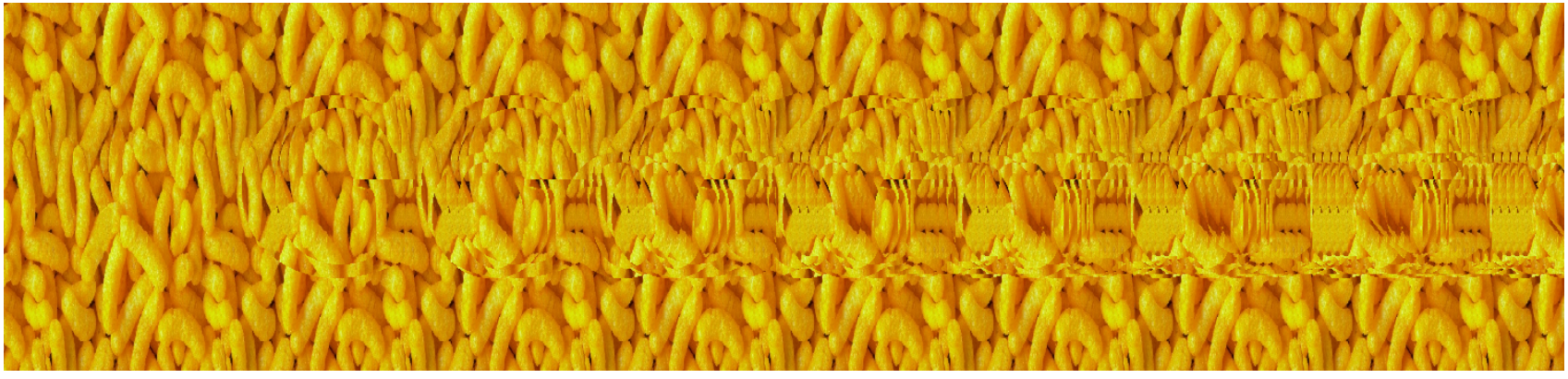
Learning Outcomes

- The importance of genetic variation
- Understand phenotypic variation and its connection to genetic variation
- Different ways to study genetic variation
- Single Nucleotide Polymorphisms (SNPs) - detection, use and limitations

Module Test 1

Don't prioritize memorization

Most questions will not test rote learning



See you for the Module 1 review in a couple of weeks!