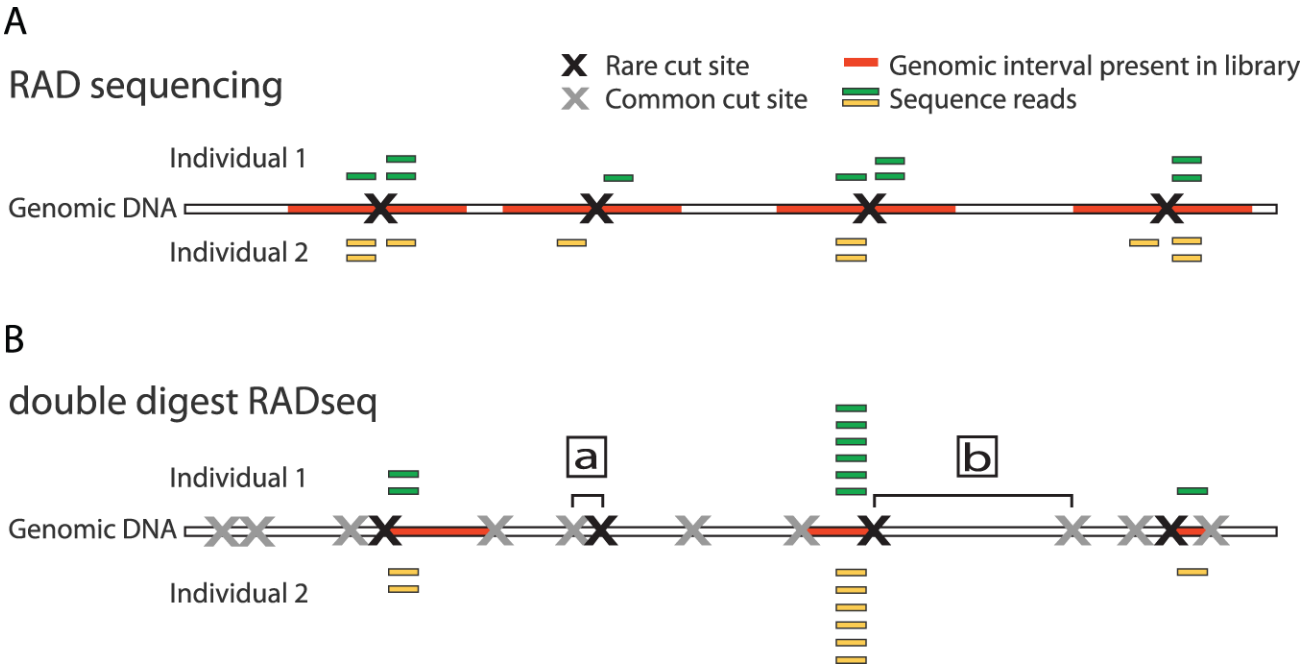
Brief Recap

Lab 2 – May 13th

Systematic DNA fragmentation

Most NGS protocols start with preparation of libraries by **shearing** the DNA



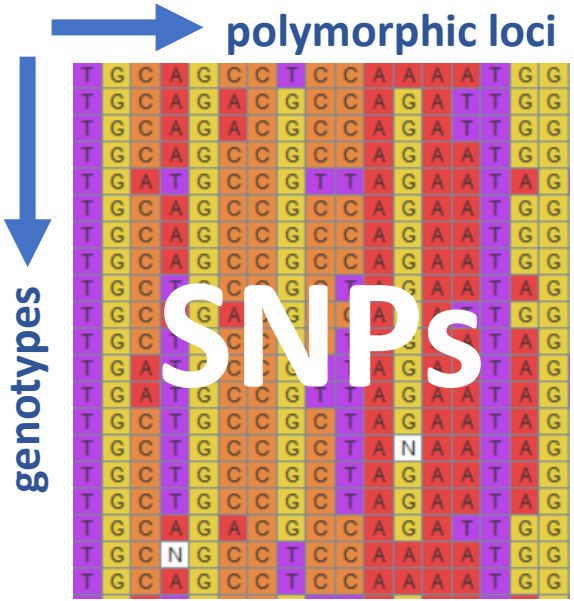
Peterson et al., 2014

Single-nucleotide polymorphism

DEFINITION: a germline substitution of a single nucleotide at a specific position in the genome and is present in a sufficiently large fraction of the population (1% or more).

Reference ATTCGC**T**CAGATTACAAACTACTTA

Ind 3 ATTCGC**A**CAGATTACAAACTACTTA



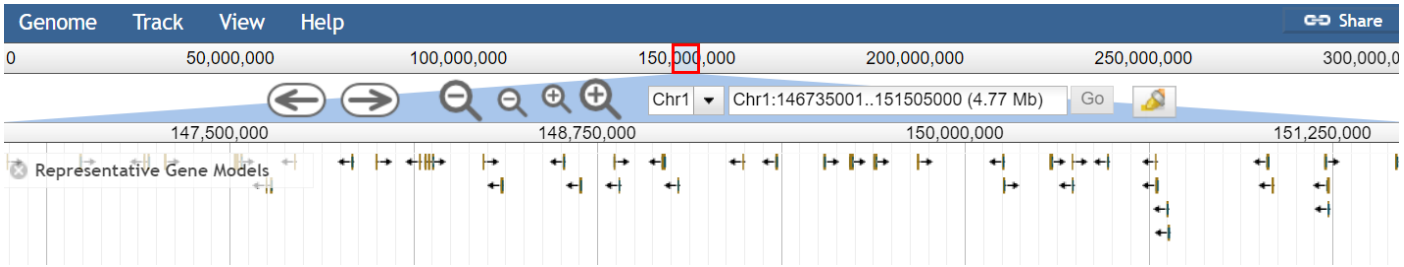
Map genotype-trait associations – A primer of Genome Wide Association Studies (GWAS)

Our working hypothesis: there are one or more «genetic factors» somewhere on the genome affecting a trait of interest

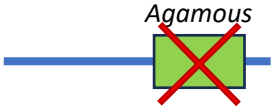
Gene X

We already know it's not an easy job:

- Most interesting traits are controlled by **multiple genetic factors**
- Eukaryotic Genomes are complex; loci may interact
- It is not really like finding a needle in a haystack; it is **finding a needle in pile of needles**



Reverse genetics



Gene(s)

What trait arises from the perturbation of a DNA sequence?

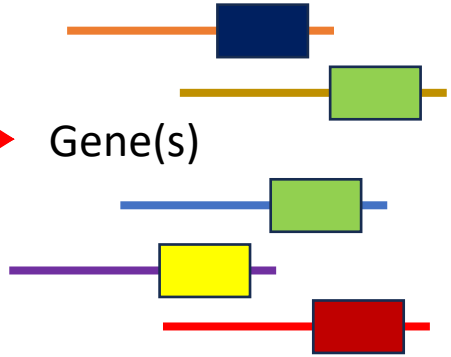
Trait(s)

Forward genetics

Trait(s)

Is variation of a trait associated with genotypic variation?

Gene(s)



A recipe for forward genetics: genome-wide association studies (GWAS)

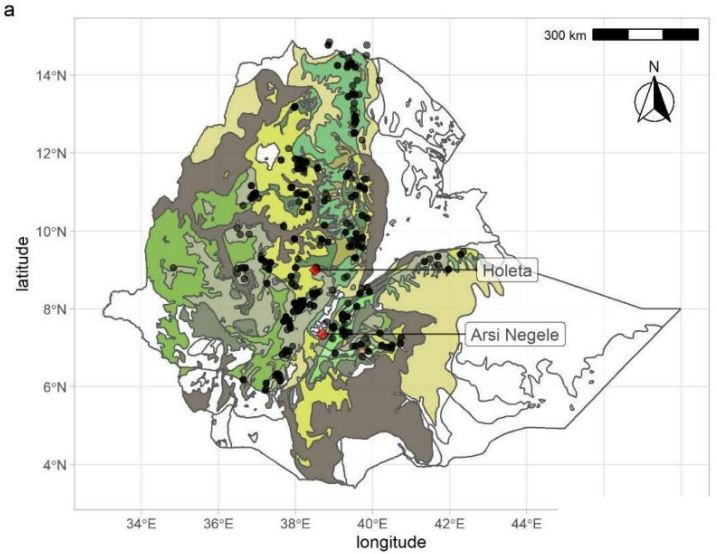
Our ingredients:

1. **Genetic materials**, a set of genetic resources in which variation is present for certain traits
2. **Phenotypic values** measured on the set of genetic materials and representing variation of interest
3. **Molecular markers** typed on the set of genetic materials; most commonly SNPs, which are bi-allelic and distributed genome wide
4. **Appropriate statistics** to connect genotypes and phenotypes; many methods, same underlying reasoning

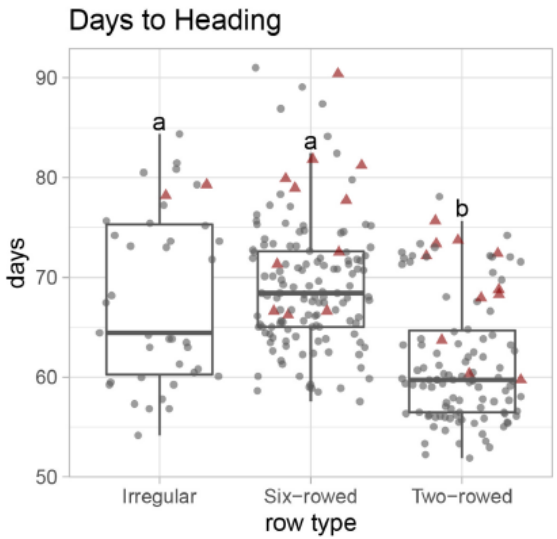
The recipe at work (see Caproni, Lakew et al 2023 in the shared folder)

Research question: climate change is affecting seasonal rainfall distribution in Ethiopia; there is the need to steer breeding towards early flowering genotypes to improve local adaptation; plant genetic resources may have useful alleles to contribute to this

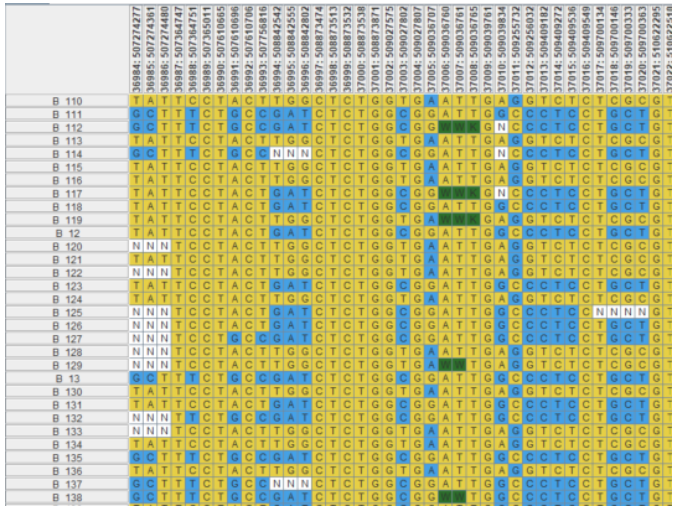
1. **Genetic materials**: A representative collection about 400 Ethiopian barley landraces and breeding lines

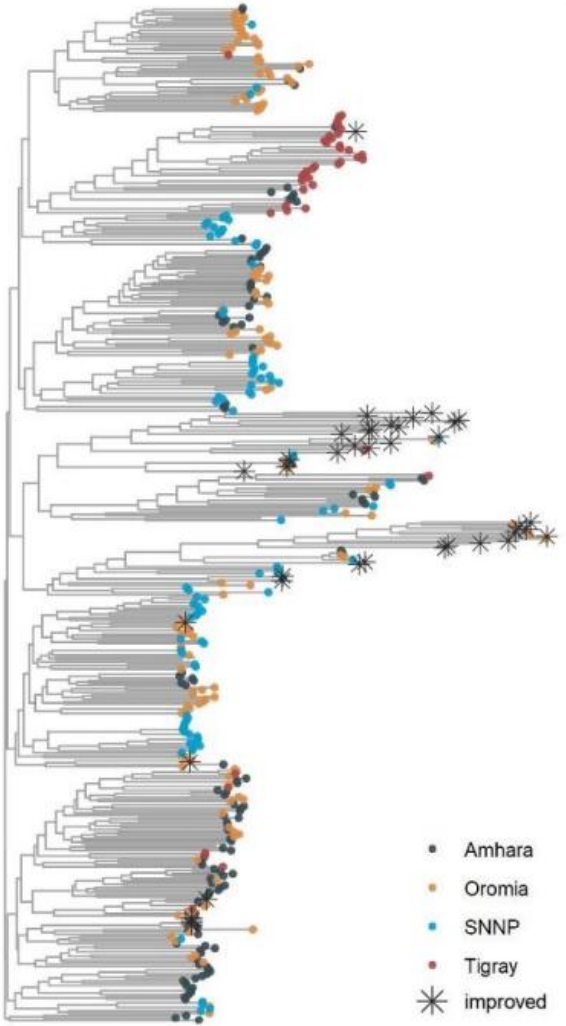


2. **Phentoypic values**: Days to flowering measured on all genotypes for which genotypic data is also available



3. **Molecular markers**: 3K SNPs describing the diversity of genetic materials across the whole genome





start_tassel.pl

```
(base) genomics@genomics-vm:~$ start_tassel.pl
```

A recipe for forward genetics: genome-wide association studies (GWAS)

Our ingredients:

1. **Genetic materials**, a set of genetic resources in which variation is present for certain traits
2. **Phenotypic values** measured on the set of genetic materials and representing variation of interest
3. **Molecular markers** typed on the set of genetic materials; most commonly SNPs, which are bi-allelic and distributed genome wide
4. **Appropriate statistics** to connect genotypes and phenotypes; many methods, same underlying reasoning

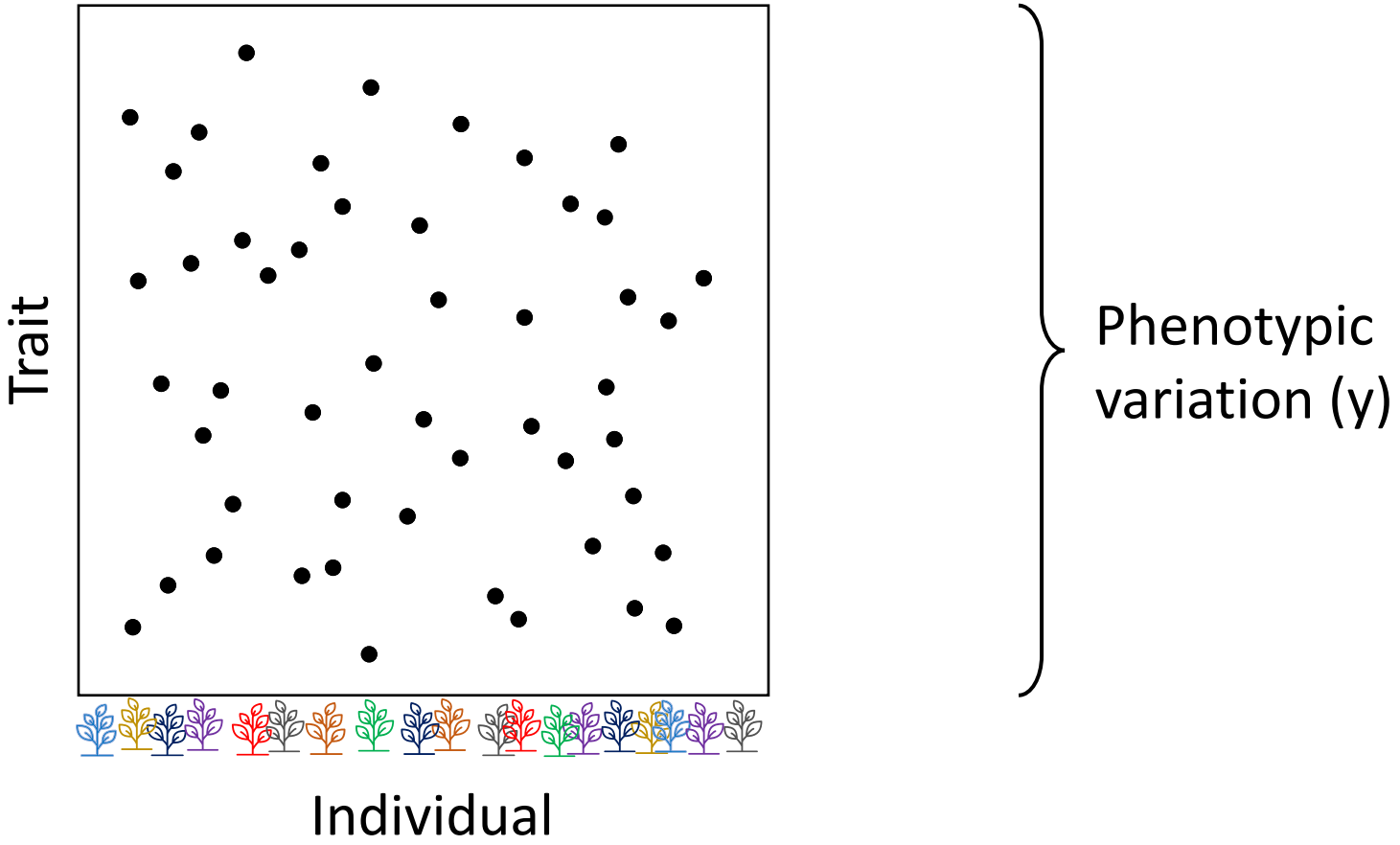
- Many different methods, same underlying reasoning: is there any given allele (marker) associated with the value of the trait of interest?
- In other words, we want to know whether our response variable (y, the phentoype) is associated with our explanatory variable (x, the marker)
- We can address this in a simple statistical framework based on a linear model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$$

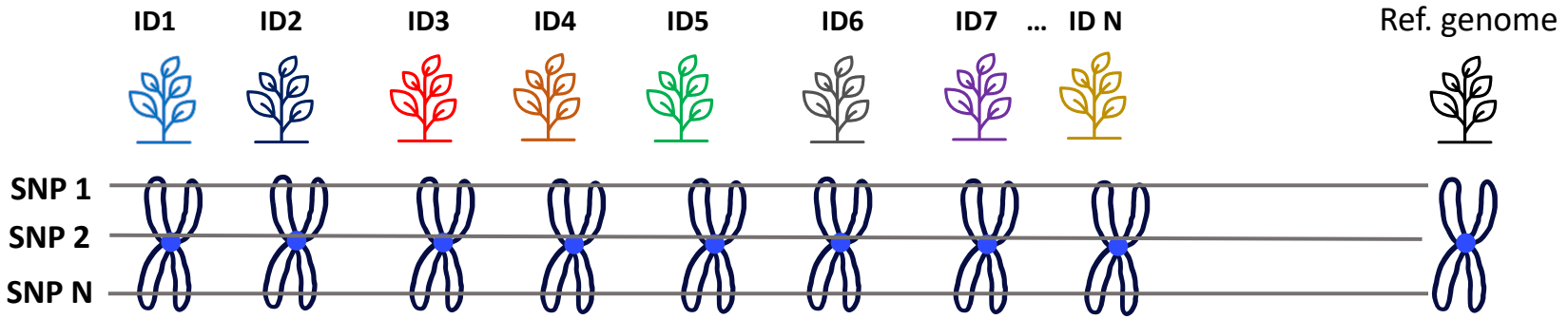
4. Appropriate statistics

4. Appropriate statistics



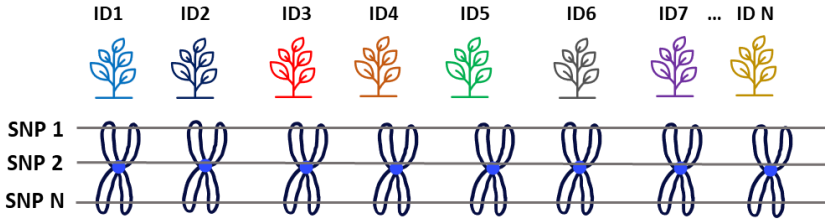
- Each individual is different from the others; when we genotype them with SNPs, we obtain biallelic markers at each locus, with different outputs depending on their allelic diversity
- We don't really need to worry about nucleotides; let's rather think in terms of alleles, and let's call the allele **0** when it is the same as the reference genome and **1** when it is different

Homozyogous reference: 00 Heterozygous: 01 Homozygous aternative: 11

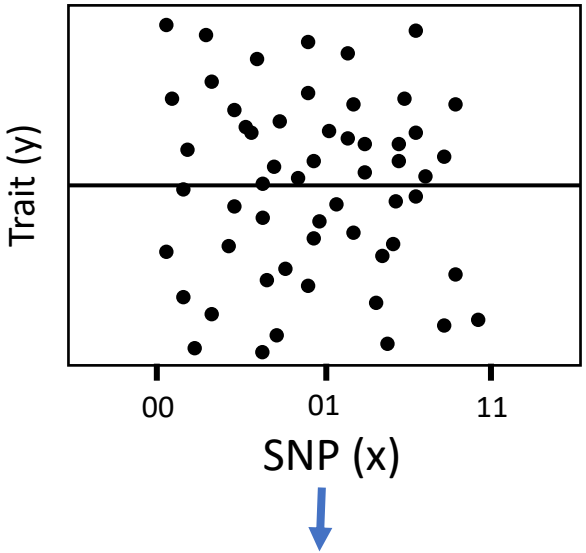


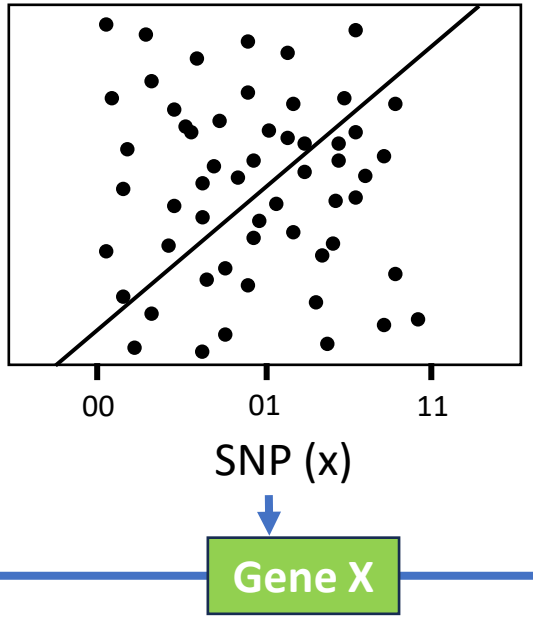
SNP 1	00	00	01	01	11	00	11	00
SNP 2	01	11	00	01	00	11	11	11
SNP *N*	11	00	00	11	11	11	00	01

Running a GWAS fitting a linear model to connect phenotypes and alleles at each locus



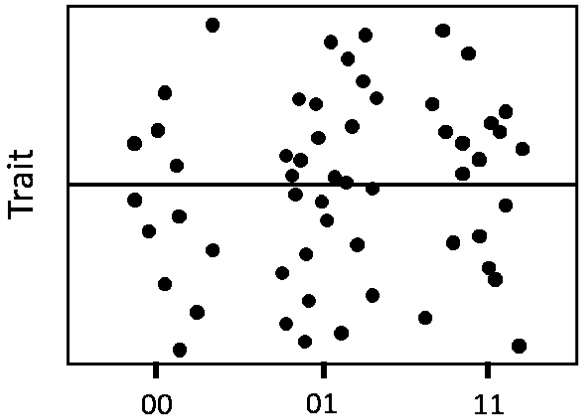
SNP 1	00	00	01	01	11	00	11	00
SNP 2	01	11	00	01	00	11	11	11
SNP *N*	11	00	00	11	11	11	00	01





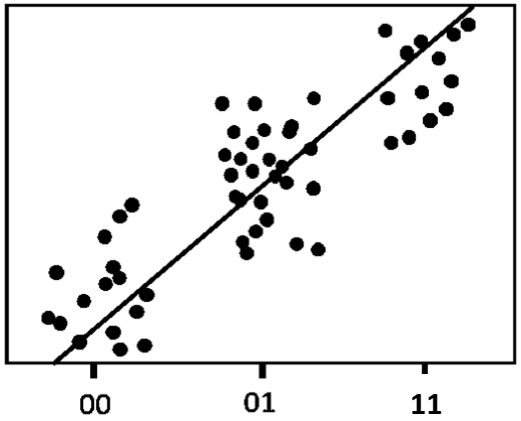
No association; this is the outcome expected on most tests (as most of the markers/loci have nothing to do with the trait)

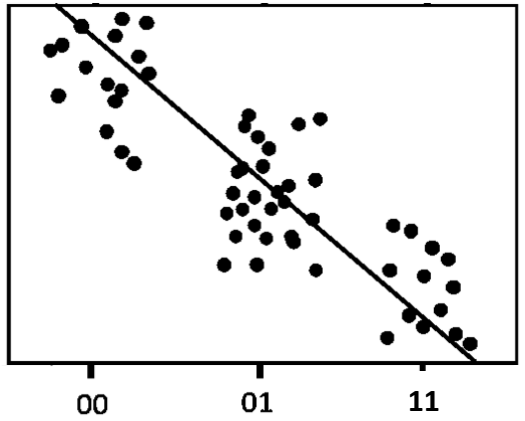
$$y = \beta_0 + \cancel{\beta_1 x} + \varepsilon$$



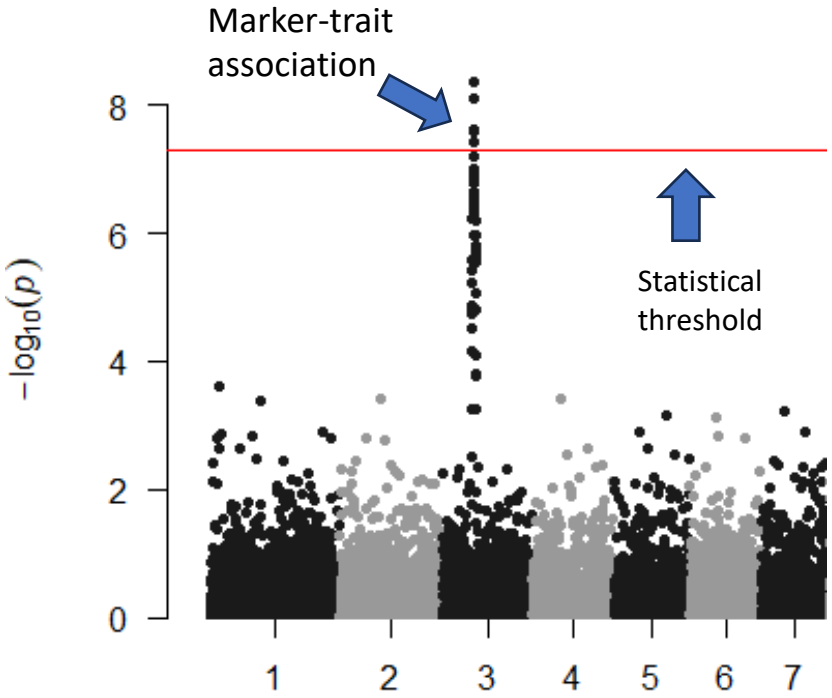
Association; it seems that the response variable is associated with the explanatory variable, and we expect it to happen rarely. To what extent the association is significant, the statistics tells us

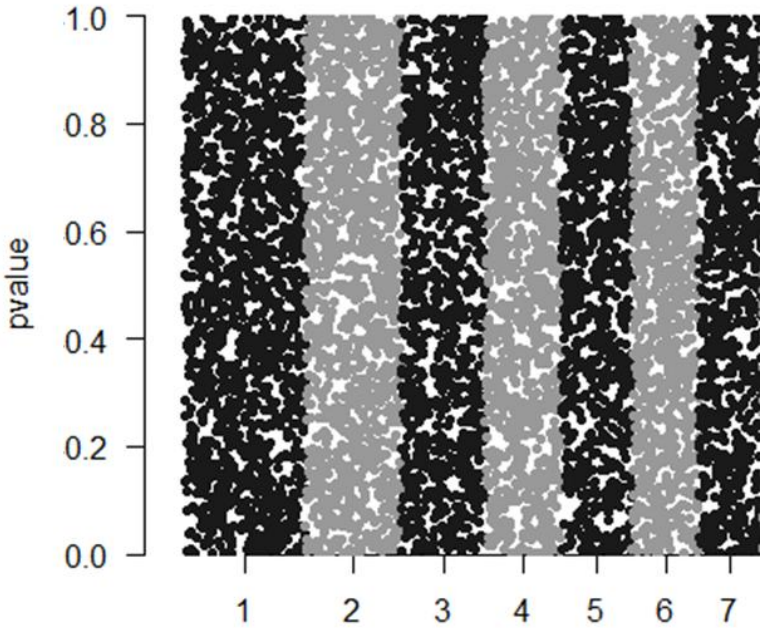
$$y = \beta_0 + \beta_1 x + \varepsilon$$





- The model is tested on all markers; if you have 1M markers, that's 1M tests!
- Each test is specific to a marker, which is specific to a genomic location
- The common representation of the outcome is a **Manhattan plot** which puts together position on the genome (x) and significance of the associated test (y)

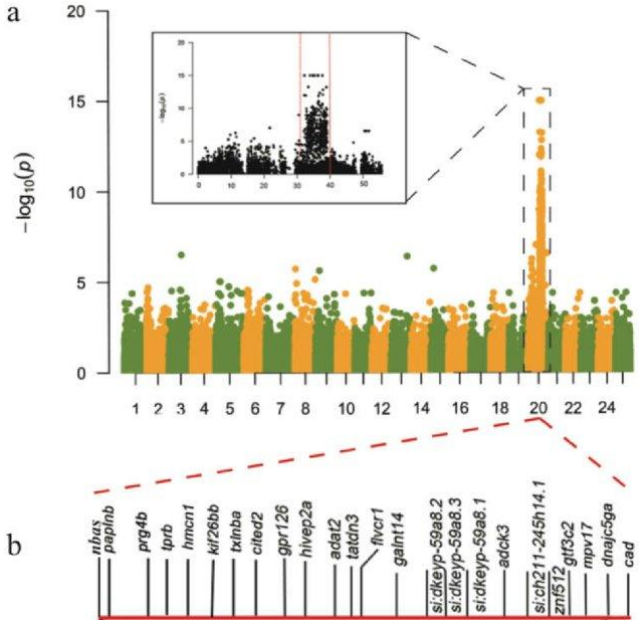




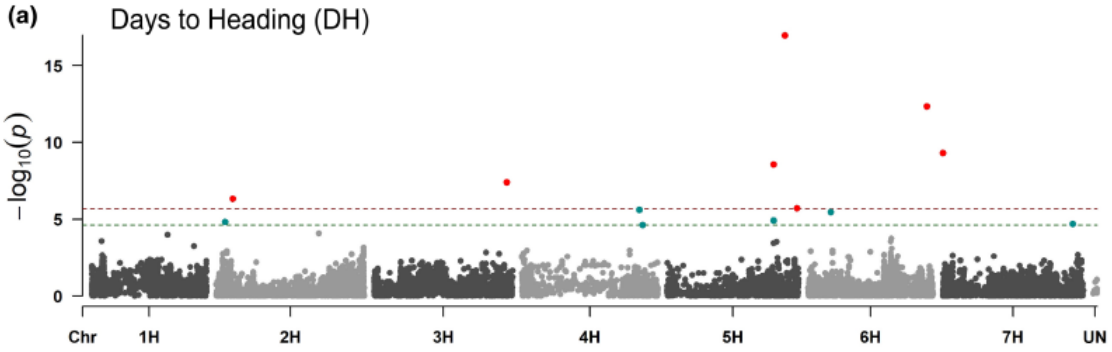
Remember that SNP markers, however many they may be, seldom represent the full extent of variation in the genome

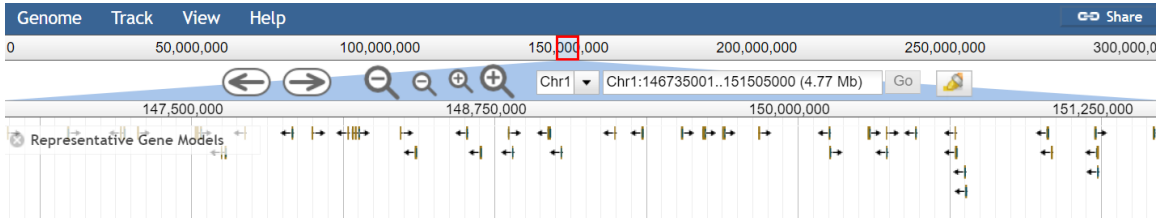
- Markers are our **proxy** to represent variation in the DNA level; they are the mean to an end and not the end itself

The reason why we capture the «effect» of a specific genetic factor on the value of the trait through GWAS is that **linkage disequlibrium (LD) exists between the marker and the causative variant**



Back to Ethiopian barley genetic resources now





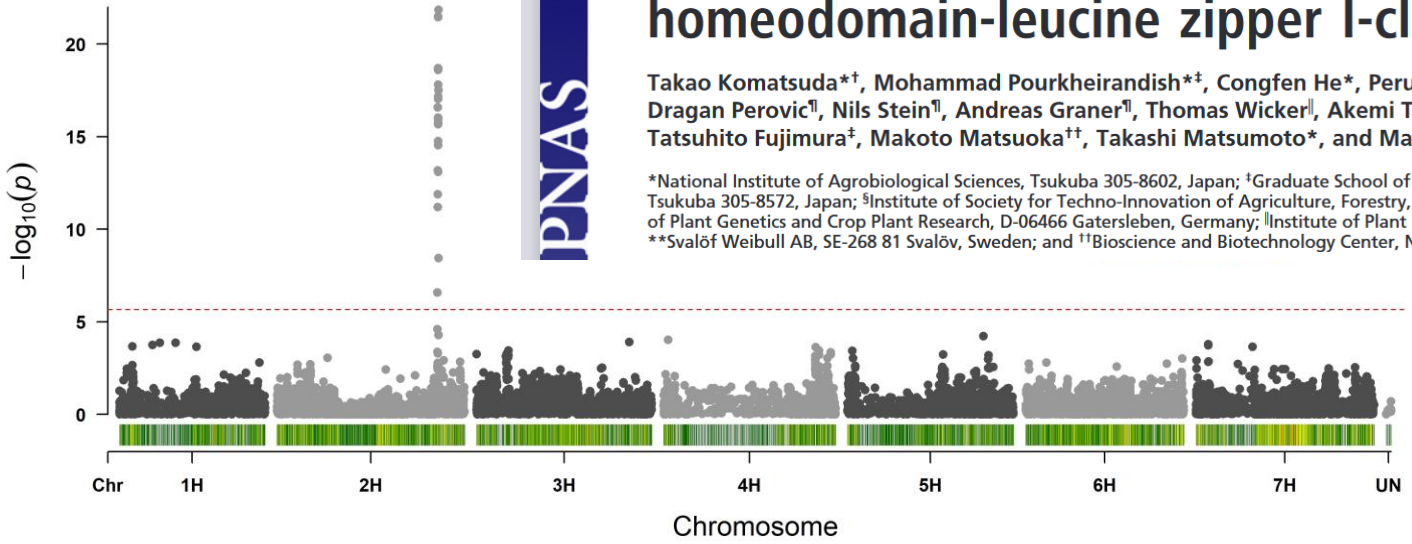
What's next?

- Characterize gene models in the region
- Develop segregating populations to fine map genetic elements
- Design cheap markers tagging loci of interest
- Derive sequences to be tested with **reverse genetics**

GWAS: proof of concept

Mapping lateral spikelet fertility

Vrs1



PNAS

Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene

Takao Komatsuda*†, Mohammad Pourkheirandish*‡, Congfen He*, Perumal Azhaguvel*, Hiroyuki Kanamori§, Dragan Perovic¶, Nils Stein¶, Andreas Graner¶, Thomas Wicker‖, Akemi Tagiri*, Udda Lundqvist**, Tatsuhito Fujimura‡, Makoto Matsuoka††, Takashi Matsumoto*, and Masahiro Yano*

*National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan; ‡Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba 305-8572, Japan; §Institute of Society for Techno-Innovation of Agriculture, Forestry, and Fisheries, Tsukuba 305-0854, Japan; ¶Leibniz Institute of Plant Genetics and Crop Plant Research, D-06466 Gatersleben, Germany; ‖Institute of Plant Biology, University of Zürich, CH-8008 Zürich, Switzerland; **Svalöf Weibull AB, SE-268 81 Svalöv, Sweden; and ††Bioscience and Biotechnology Center, Nagoya University, Nagoya 464-8601, Japan

END