

INSTITUTE
OF PLANT
SCIENCES



Sant'Anna
School of Advanced Studies – Pisa

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

September 25th

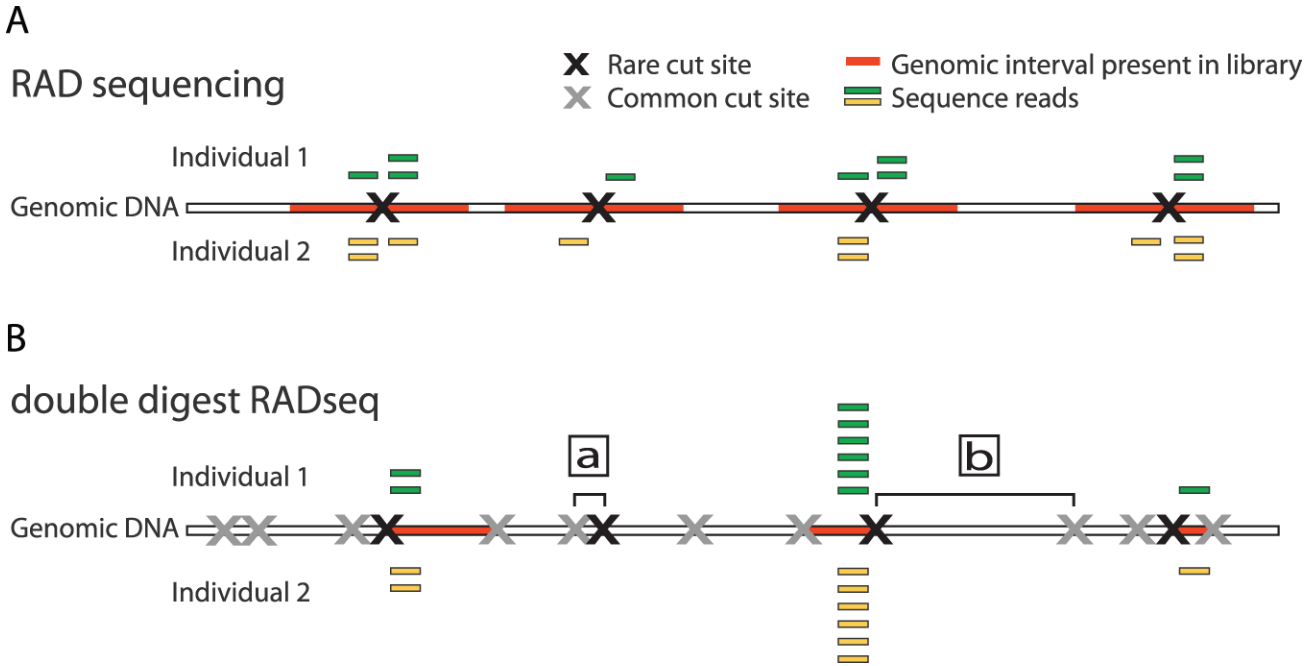


Brief Recap

September 25th

Systematic DNA fragmentation

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

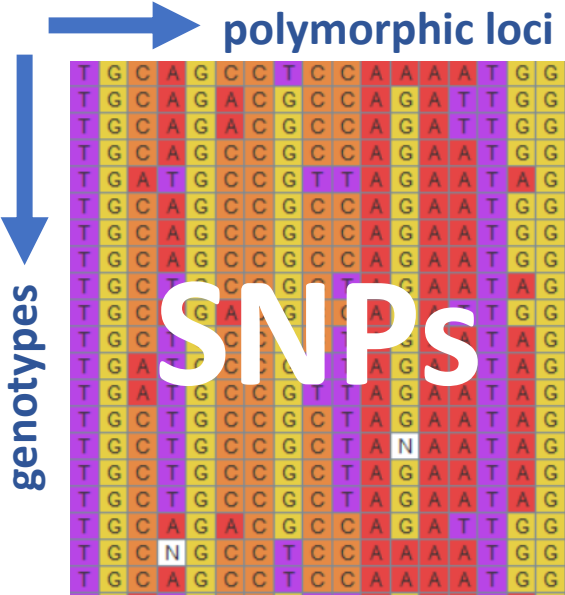


Single-nucleotide polymorphism

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 ATTCGCTCAGATTACAACTACTTA

Ind 3 ATTCGC**A**CAGATTACAACTACTTA



INSTITUTE
OF PLANT
SCIENCES



Sant'Anna
School of Advanced Studies – Pisa

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

September 25th

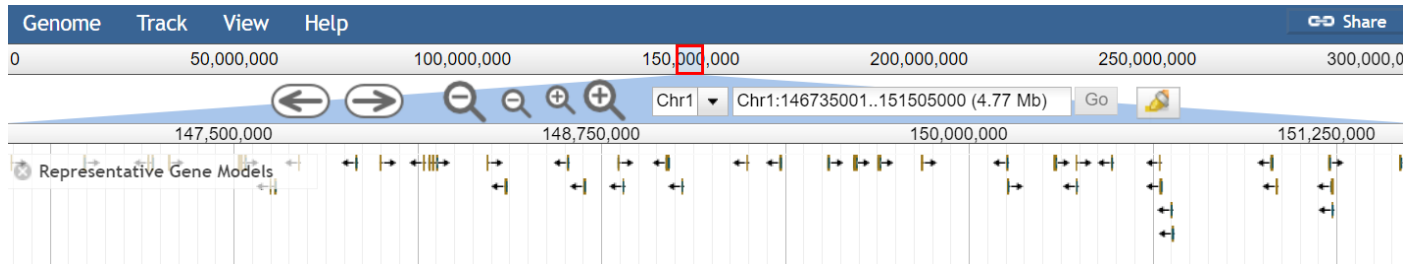


Our working hypothesis: there are one or more «genetic factors» somewhere on the genome affecting a trait of interest

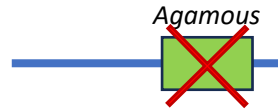
Gene X

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)



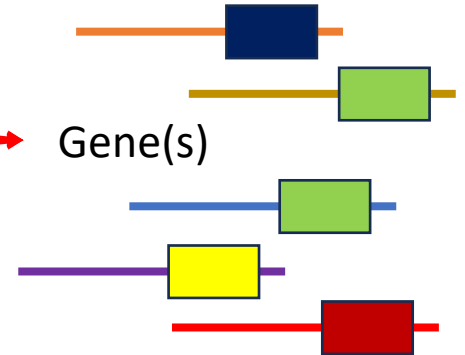
Forward genetics

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:

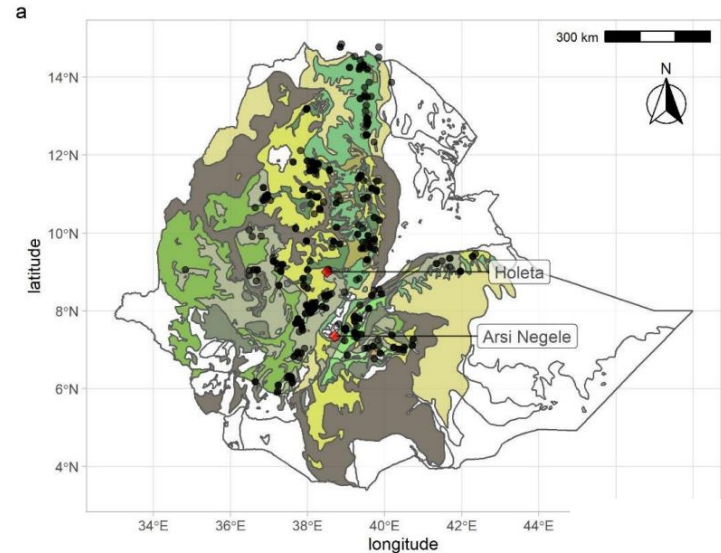
1. **Genetic materials**, a set of genetic resources in which variation is present for certain traits
2. **Phenotypic values** measured on the set of genetic materials and representing variation of interest
3. **Molecular markers** typed on the set of genetic materials; most commonly SNPs, which are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



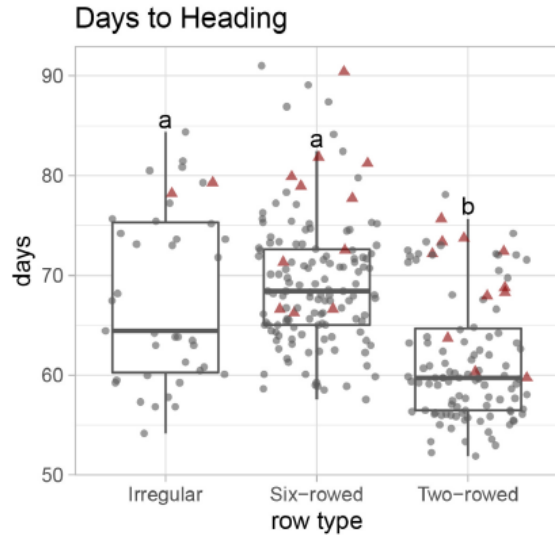
The recipe at work (see Caproni, Lakew et al 2023 in the shared folder)

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

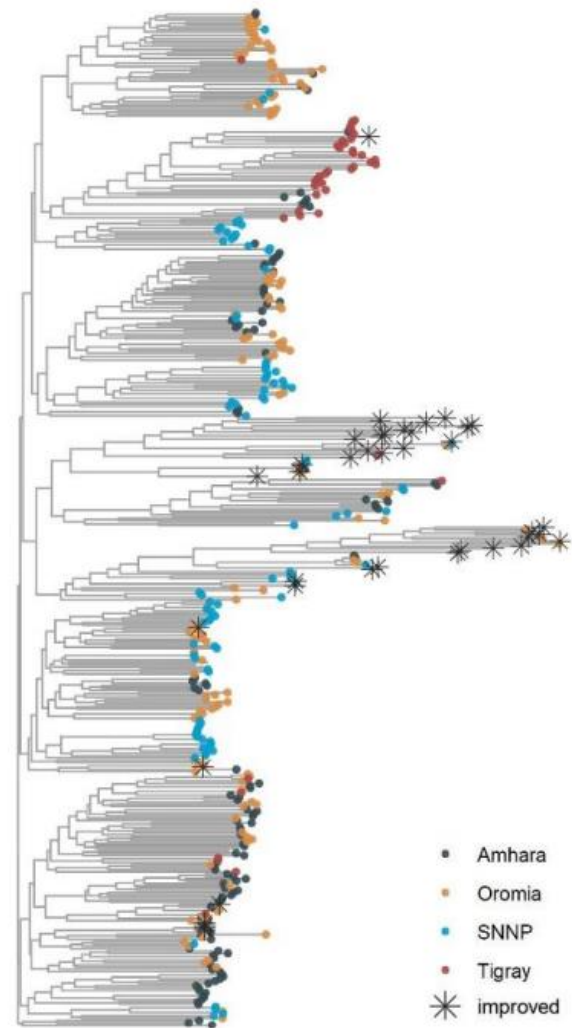


2. **Phenotypic values:** Days to flowering measured on all genotypes for which genotypic data is also available



3. Molecular markers: 30K SNPs describing the diversity of genetic materials across the whole genome

	36884: 507274277	36885: 507274361	36886: 507274480	36887: 507364747	36888: 507365011	36889: 507616096	36890: 507616706	36891: 507616816	36892: 507616906	36893: 507658816	36894: 508842542	36895: 508842555	36896: 508842802	36897: 508872474	36898: 508872532	36899: 508872532	37000: 508872538	37001: 508872871	37002: 509027575	37003: 509027862	37004: 509027867	37005: 509030707	37006: 509030707	37007: 509030701	37008: 509030705	37009: 509030834	37010: 509030834	37011: 509255732	37012: 509256032	37013: 509400912	37014: 509400912	37015: 509400952	37016: 509400952	37017: 509700134	37018: 509700146	37019: 509700333	37020: 509700363	37021: 510822285	37022: 510822285
B 110	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 111	G C T T T T C G C G C G G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 112	G C T T T T C G C G C G G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 113	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 114	G C T T T T C G C G C G G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 115	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 116	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 117	T A T T C C G A C T T G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 118	T A T T C C G A C T T G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 119	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 12	T A T T C C G A C T T G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 120	N N N T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 121	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 122	N N N T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 123	T A T T C C G A C T T G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 124	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 125	N N N T C C G A C T T G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 126	N N N T C C G A C T T G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 127	N N N T C C G C C G G A T T T G G C C C C T C T C T C G C C G																																						
B 128	N N N T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 129	N N N T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 13	G C T T T T C G C G C G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 130	T A T T C C G A C T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 131	T A T T C C G A C T T G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 132	N N N T C C G C C G G A T T T G G C C C C T C T C T C G C C G																																						
B 133	N N N T C C G A C T T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 134	T A T T C C G A C T T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 135	G C T T T T C G C G C G A T G C G A T T T G A G G G T C T C T C G C C G																																						
B 136	T A T T C C G A C T T T G G C C T C G T G G T G A A T T T G A G G G T C T C T C G C C G																																						
B 137	G C T T T T C G C C N N N T C G A T T T G A G G G T C T C T C G C C G																																						
B 138	G C T T T T C G C C G A T G C G A T T T G A G G G T C T C T C G C C G																																						



- Amhara
- Oromia
- SNNP
- Tigray
- ✳ improved

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



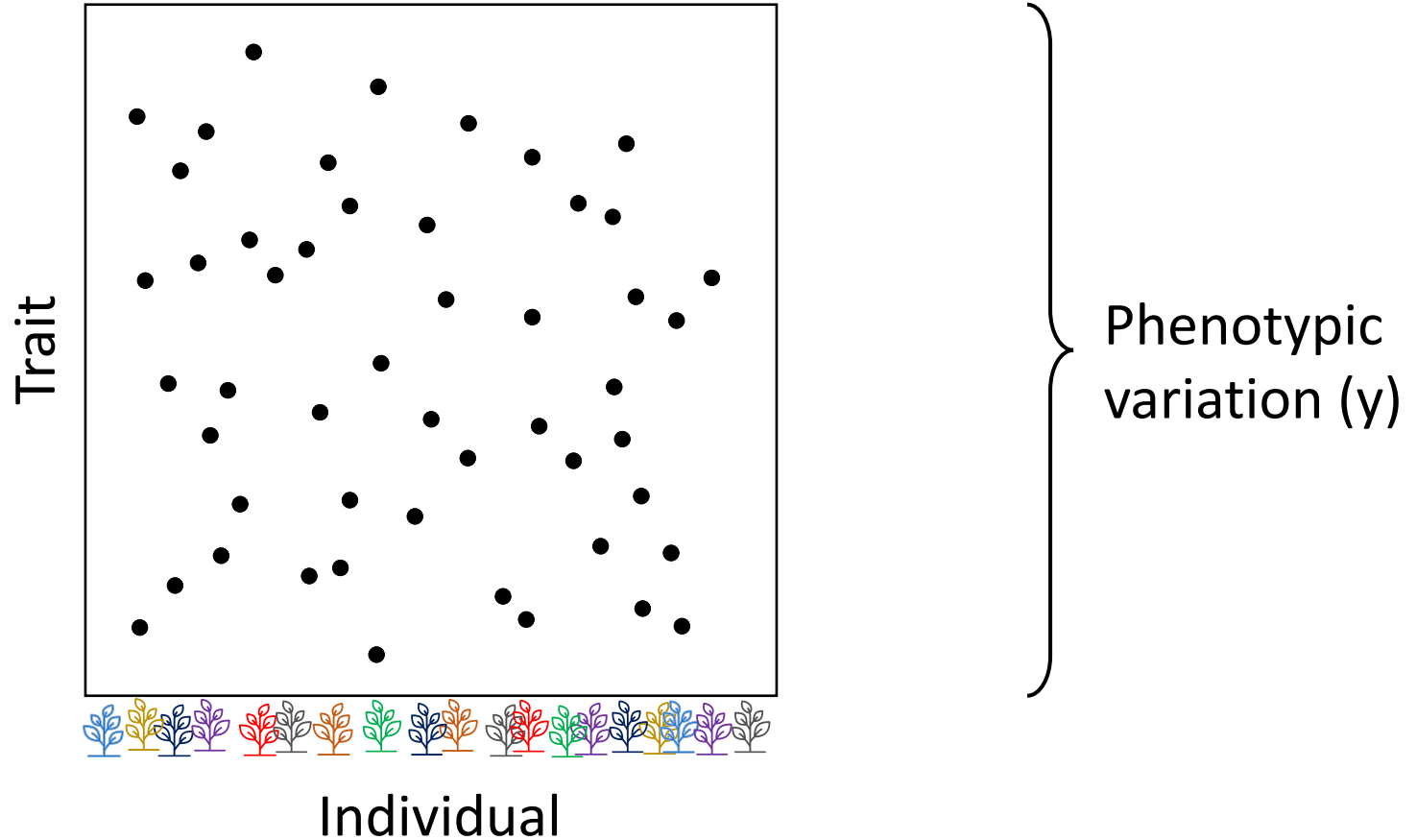
- Many different methods, same underlying reasoning: is there any given allele (marker) associated with the value of the trait of interest?
- In other words, we want to know whether our response variable (y , the phenotype) is associated with our explanatory variable (x , the marker)
- We can address 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

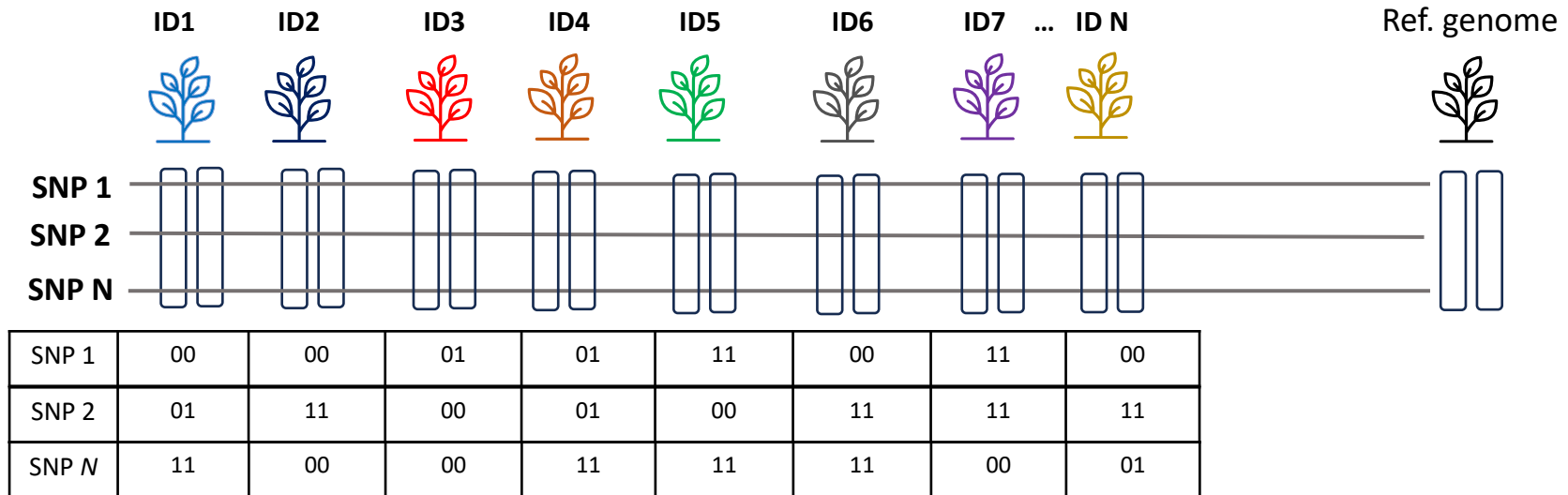


- Each individual is different from the others; when we genotype them with SNPs, we obtain biallelic markers at each locus, with different outputs depending on their allelic diversity
- We don't really need to worry about nucleotides; let's rather think in terms of alleles, and let's call the allele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

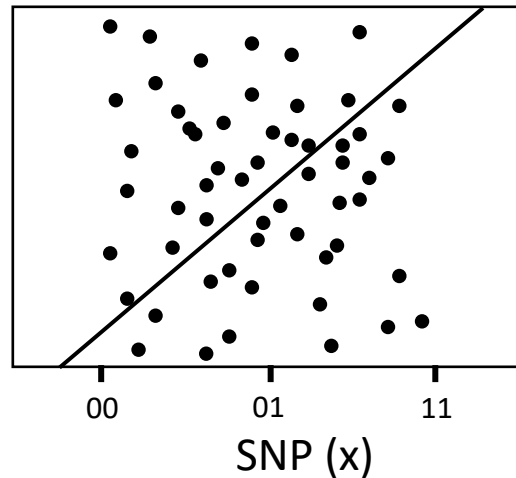
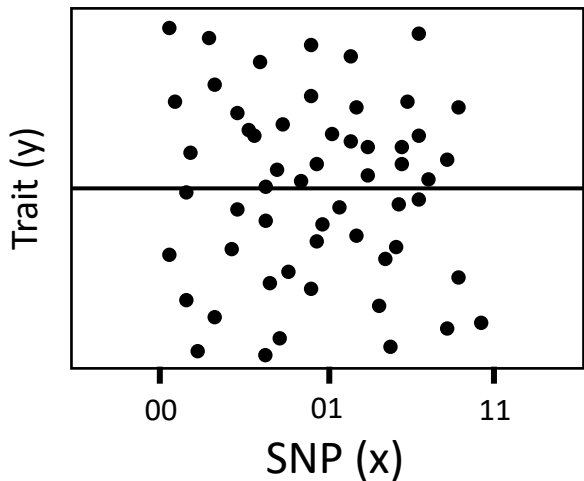
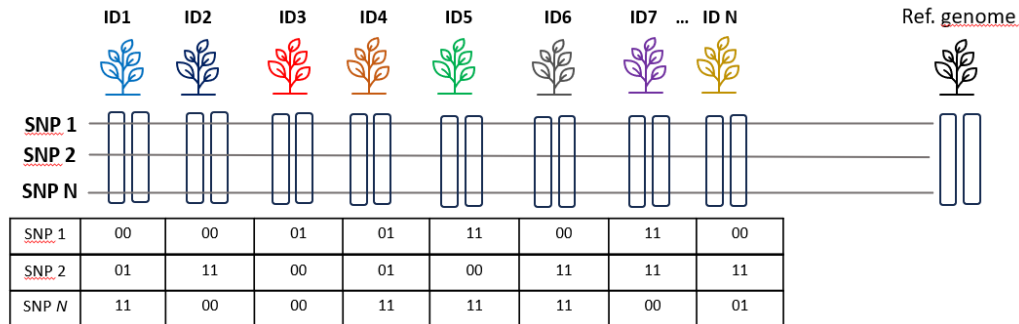
Homozygous reference: 00

Heterozygous: 01

Homozygous alternative: 11

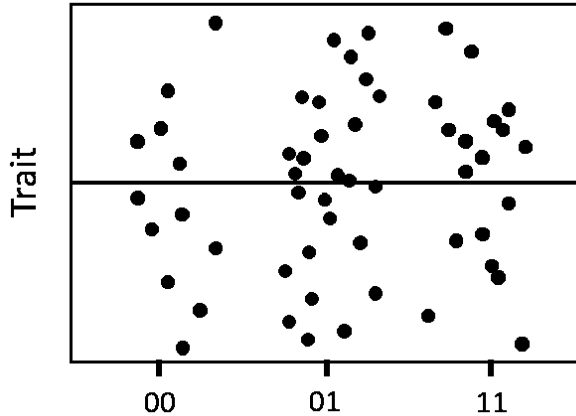


Running a GWAS fitting a linear model to connect phenotypes and alleles at each locus



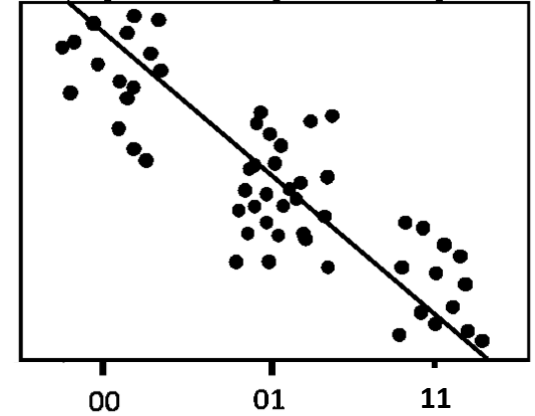
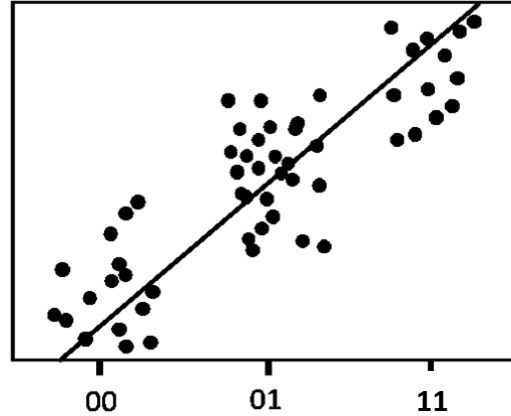
No association; this is the outcome expected on most tests (as most of the markers/loci have nothing to do with the trait)

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

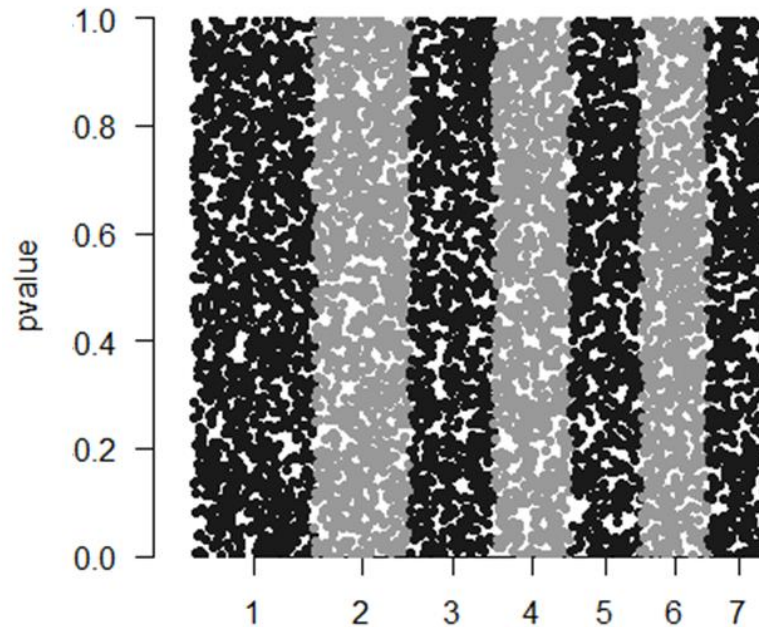
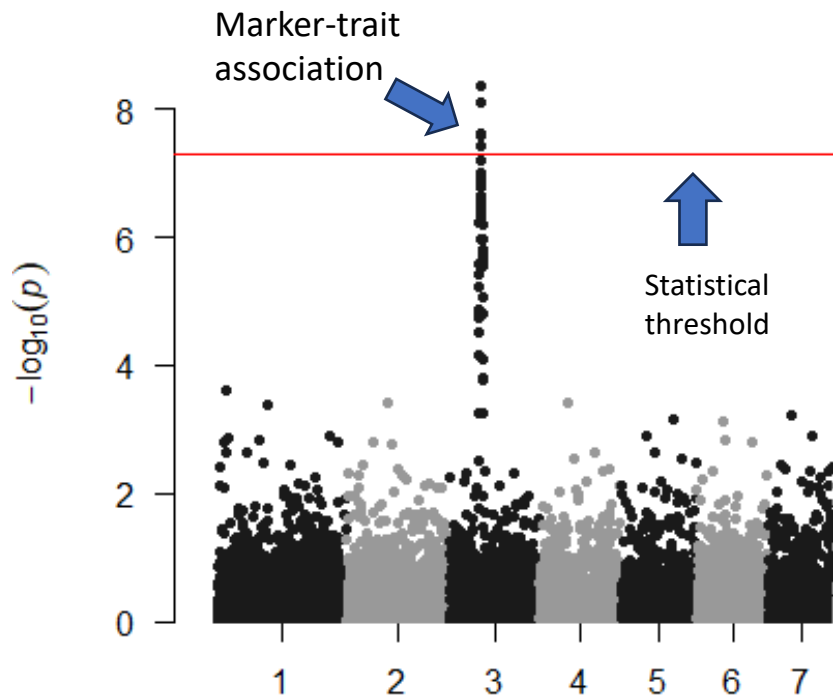


Association; it seems that the response variable is associated with the explanatory variable, and we expect it to happen rarely. To what extent the association is significant, the statistics tells us

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



- 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)

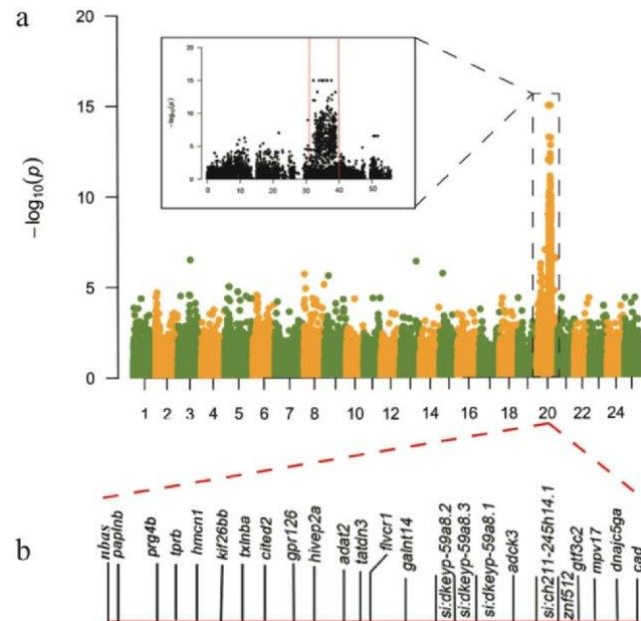


Remember that SNP markers, however many they may be, seldom represent the full extent of variation in the genome

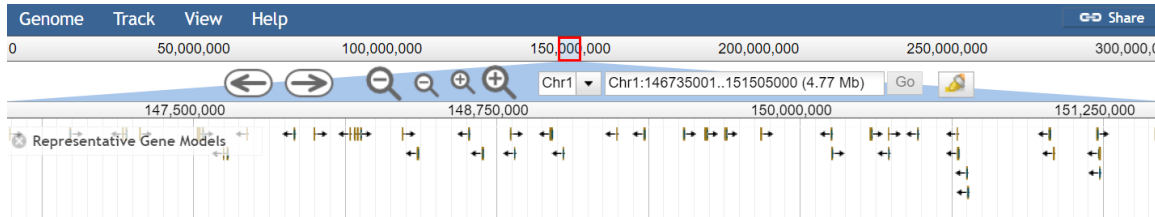
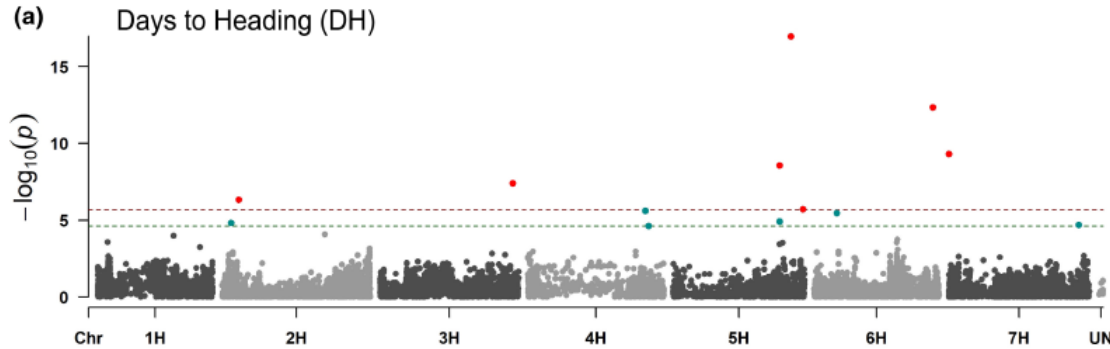
- Markers are our **proxy** to represent variation in the DNA level; they are the mean to an end and not the end itself



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



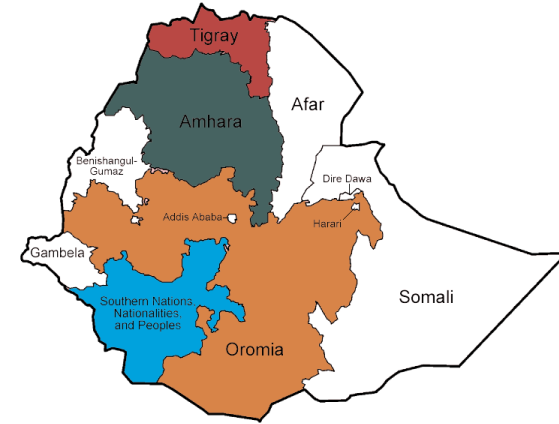
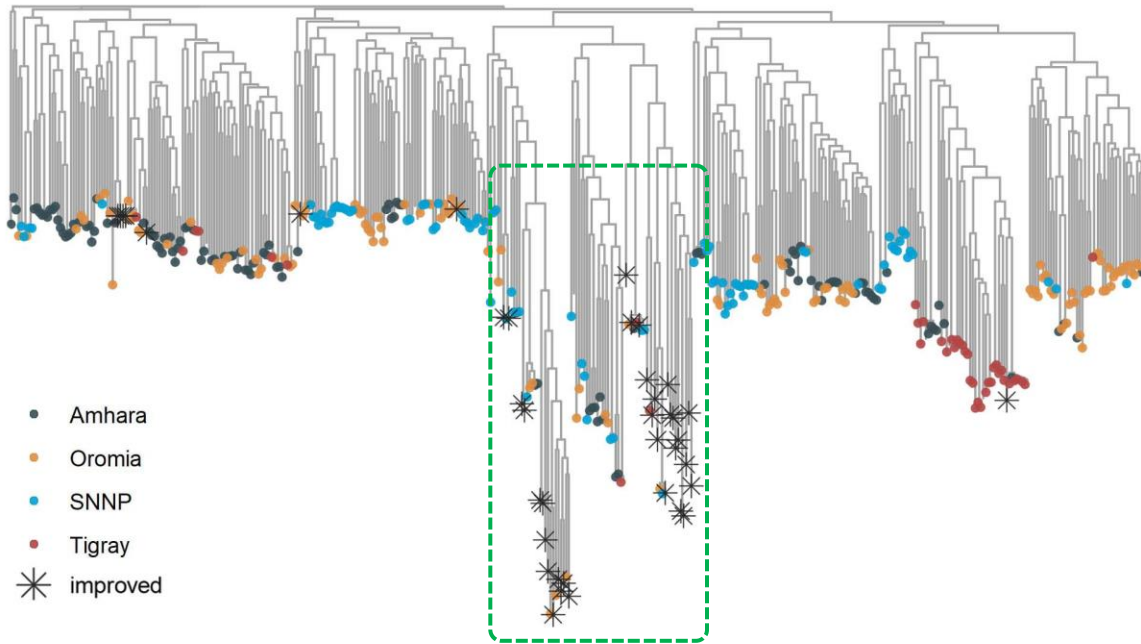
Back to Ethiopian barley genetic resources now



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**

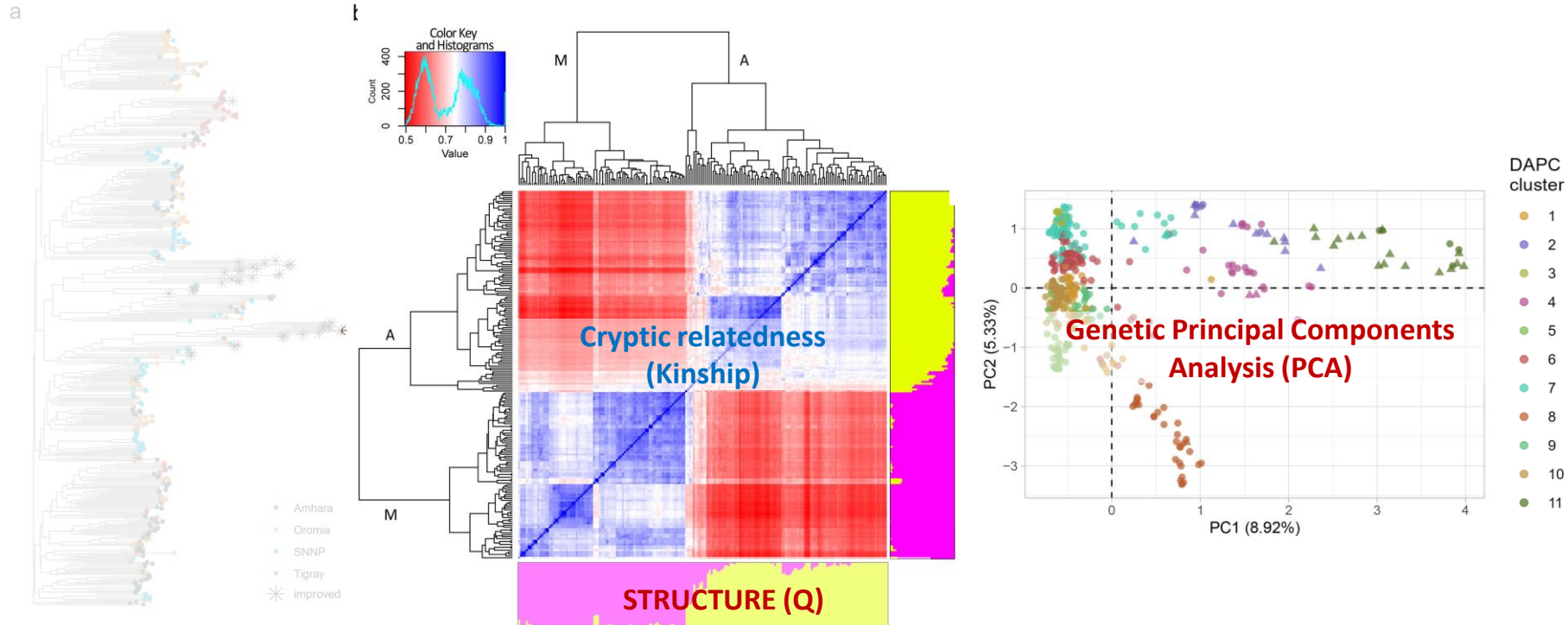
Genetic diversity and structure



- 2064 K LD-pruned SNP to study structure
- Most **improved materials** set apart
- Regional provenance partially explain structure

Based on 2064 LD-pruned SNPs

There are several way to describe and measure structure



Genome-wide association study

$$y = \underbrace{\text{SNP} + \text{Q (or PCs)}}_{\text{General Linear Model (GLM)}} + e$$

(fixed effect) (fixed effect)

$$Y = \underbrace{\text{SNP} + \text{Q (or PCs)}}_{\text{General Linear Model (GLM)}} + \text{Kinship} + e$$

(fixed effect) (fixed effect) (random effect)

Mixed Linear Model (MLM)

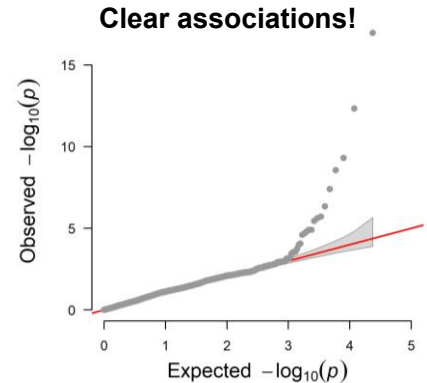
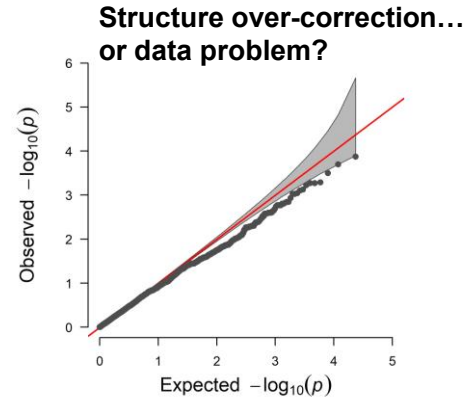
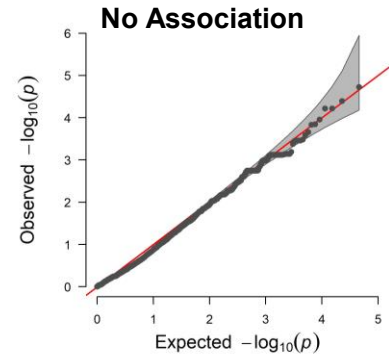
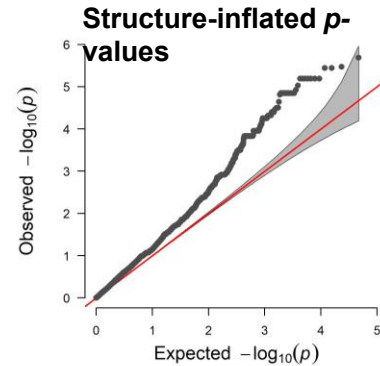
Yu et al, 2005 Nature Genetics

Several other methods including stepwise regressions like MMLM (Multi-Locus Mixed Linear Model), FarmCPU (Fixed and Random) and BLINK

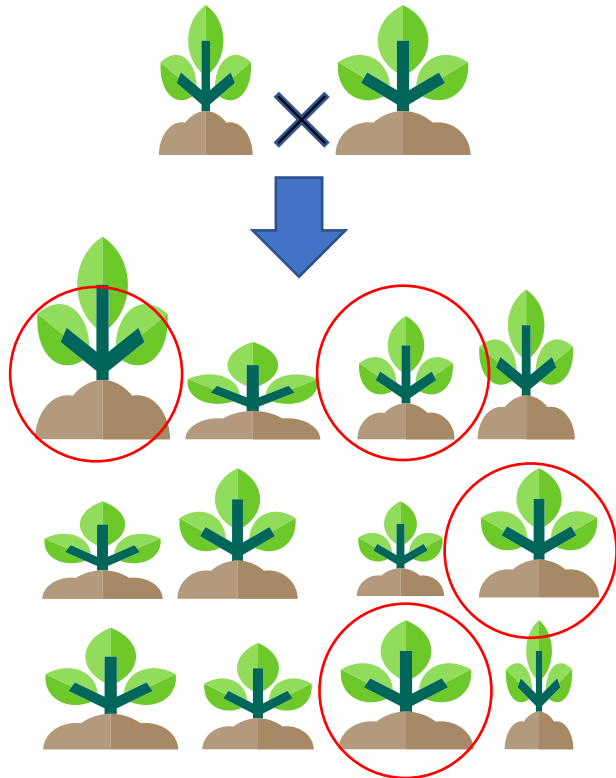
MMLM: Segura et al, 2012 Nature Genetics
FarmCPU: Liu et al. 2016, PLoS Genetics
BLINK: Huang et al., 2019, Gigascience

Evaluating the fit of the model (QQ-plots)

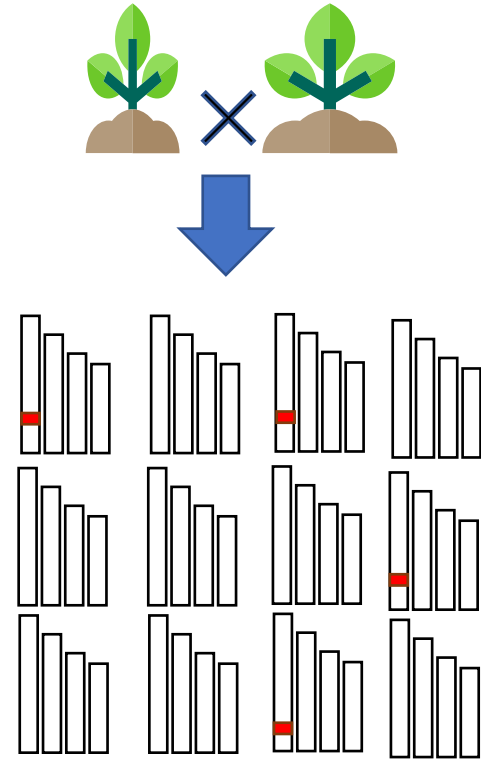
- The QQ-plot is a graphical representation of the deviation of the observed p-values from the null-hypothesis (no-association).
- P-values are sorted from the smallest to the largest and plotted against expected p -values from a theoretical χ^2 -distribution.
- If the observed values correspond to the expected values, all points are on or near the middle line between the x-axis and the y-axis.



Use of mapping information for breeding



Selection based on traits



Selection based on markers

Once a marker-trait association is discovered, it can be used to accelerate the development of new varieties with improved traits

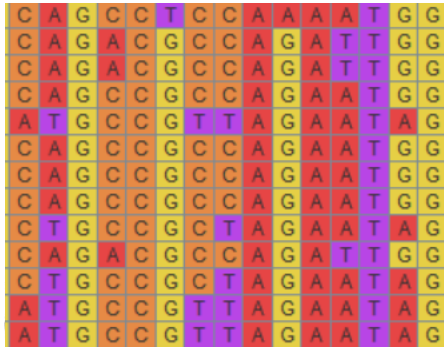


Modifying genes

Can we use this approach to study adaptation?

Genome wide association study

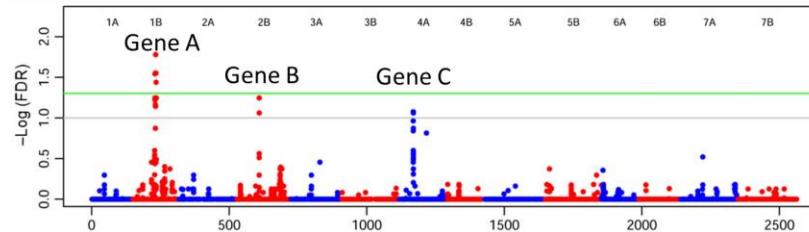
Genetic diversity



Phenotypic diversity



Association analysis

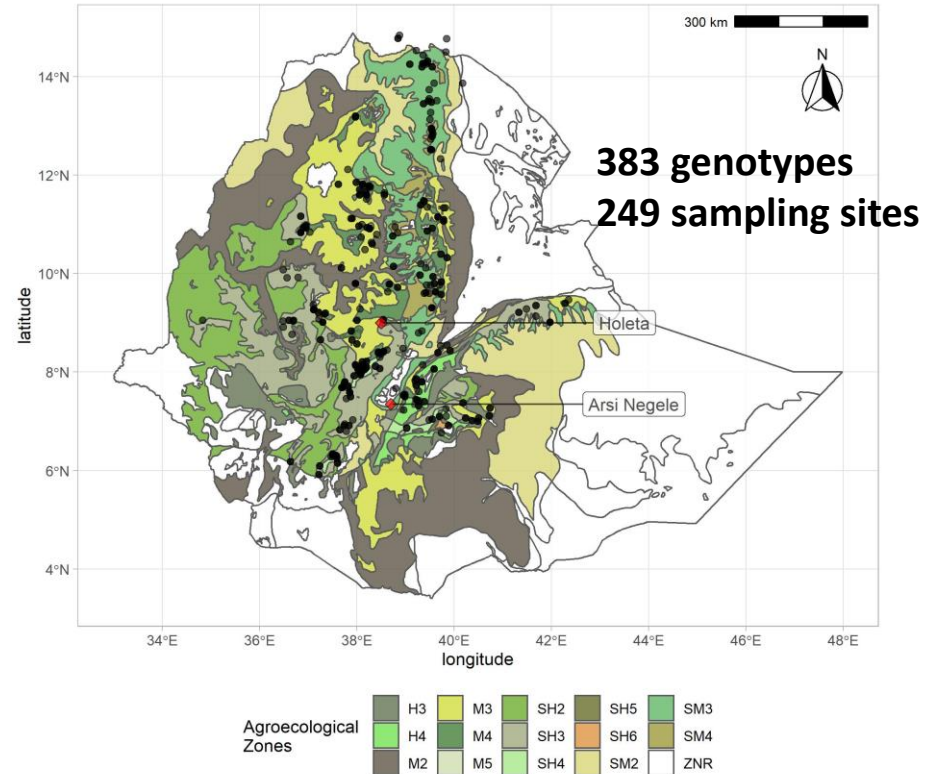


Genomic loci with breeding relevance

Let's Go back to the Barley diversity panel

**From the largest genebank of Africa
Ethiopian Biodiversity Institute (EBI)**

- **249 georeferenced landraces,**
(sampling sites)
- **383 genotypes** (resulting from
purification)
- Panel sided with **40 improved
varieties.**



Some international databases worth to mention



Remote sensing raw datasets: high-quality satellite imaging data are available from 1979



WorldClim



Interpolated climate data: here you can freely download bioclimatic indicators and projected climate data at different RCPs and different horizons



ISRIC
World Soil Information

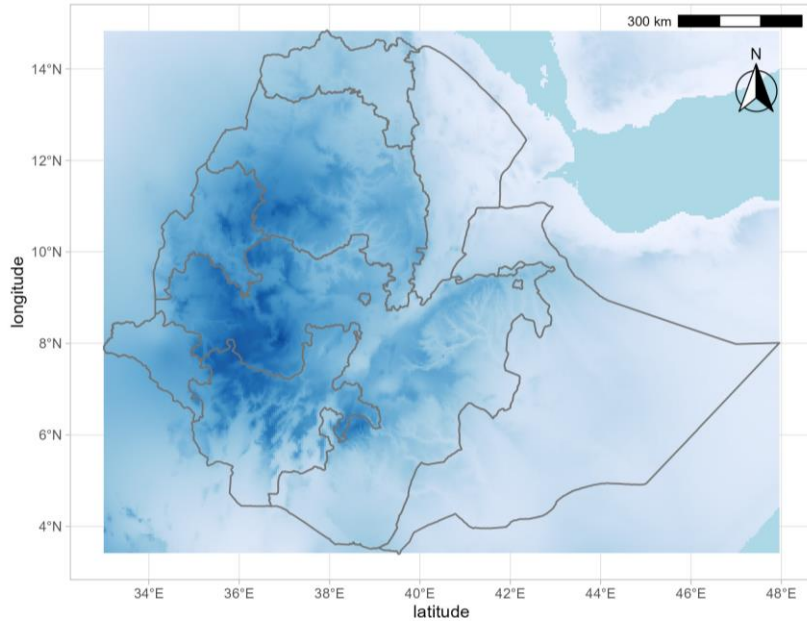


Interpolated soil datasets: soil pH, mineral content and a lot more

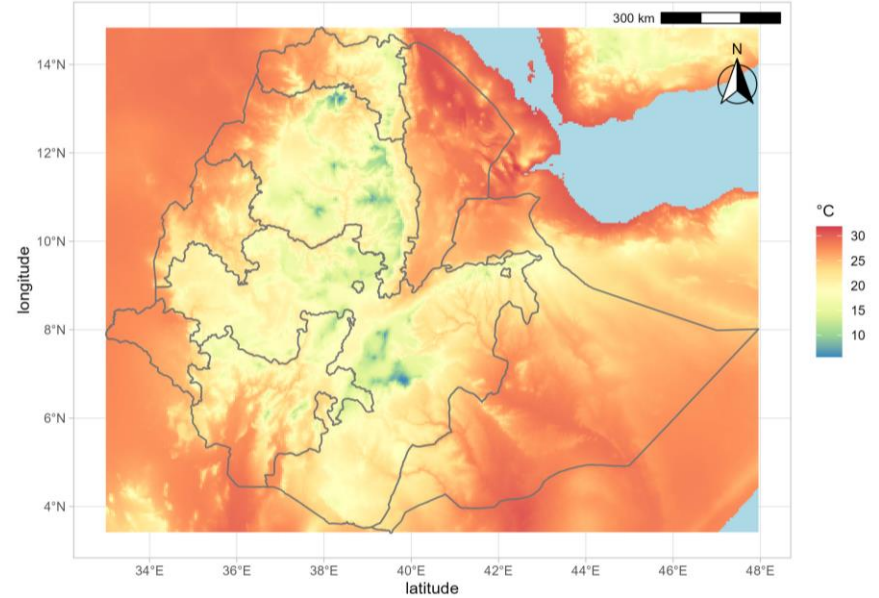
Historical climate data



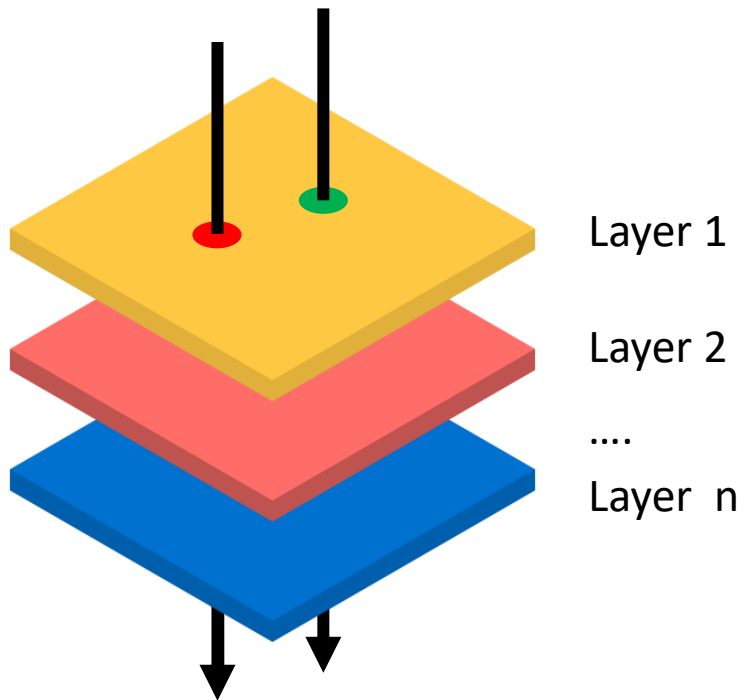
Annual precipitation









Annual mean temperature



Extracting climatic features at sampling points



		Layer 1 	Layer 2 	...	Layer n 
sampling point 1					
sampling point 2					
...					
sampling point n					

Climatic characterization

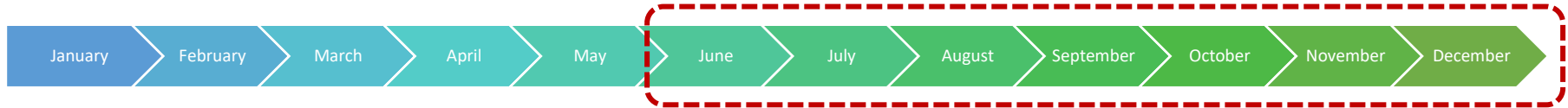
Historical climate data derived from the fifth generation of **European Centre for Medium- Range Weather Forecasts (ECMWF)** atmospheric reanalysis version (**ERA5**) data.



SPATIAL RESOLUTION $0.25^{\circ} \times 0,25^{\circ}$



Meher



For optimal spatial and temporal adjustment,

- **Temperature** data of 150 weather station of NMA +
- **Rainfall** RainFARM

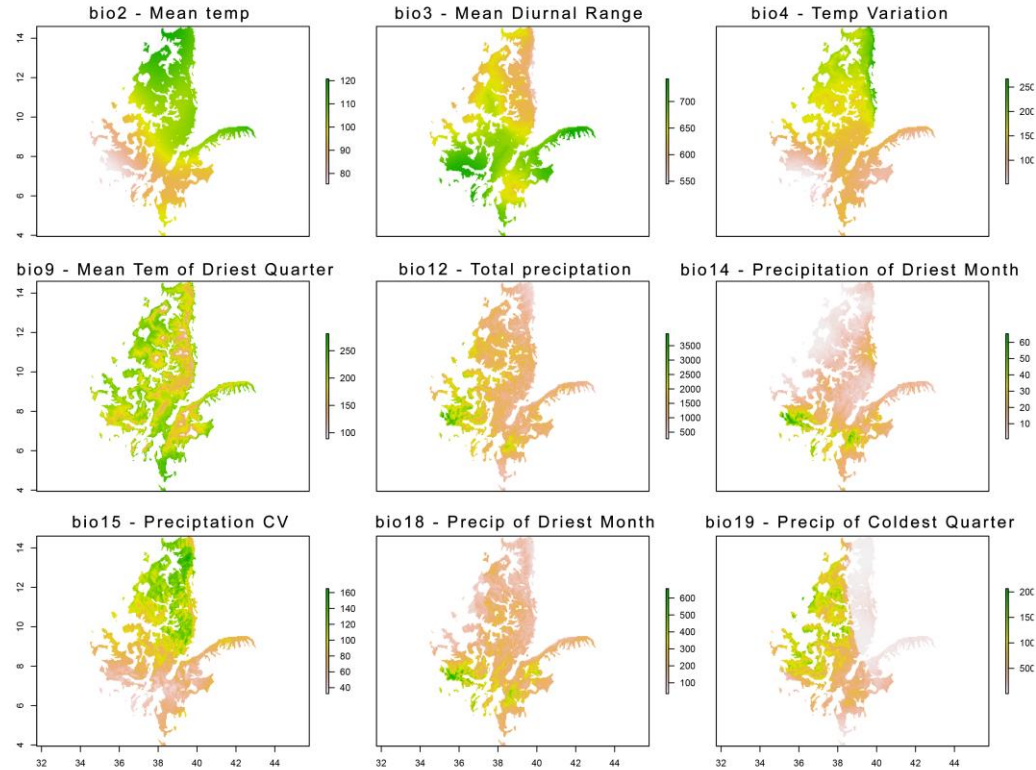
Deriving bioclimatic indicators

derive 19 bioclimatic variables
(historical climate)

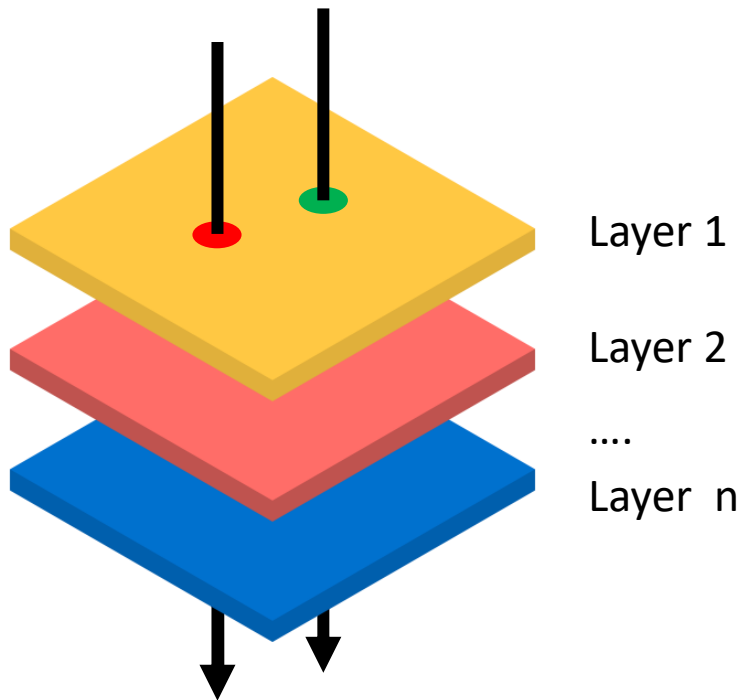
Keep data only on the agro-
ecologies > 10 hits







Extract bioclimatic variables at
each sampling point

Measure multicollinearity with
Variance Inflation Factor (VIF)



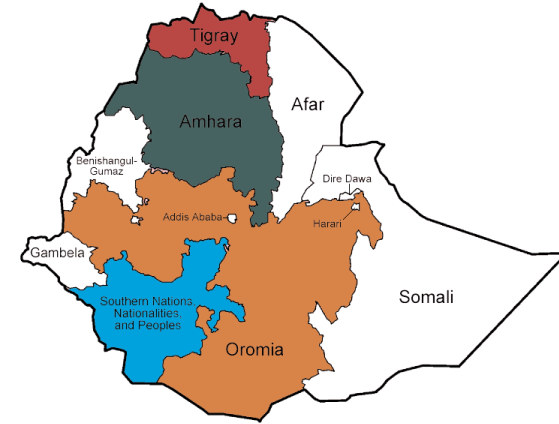
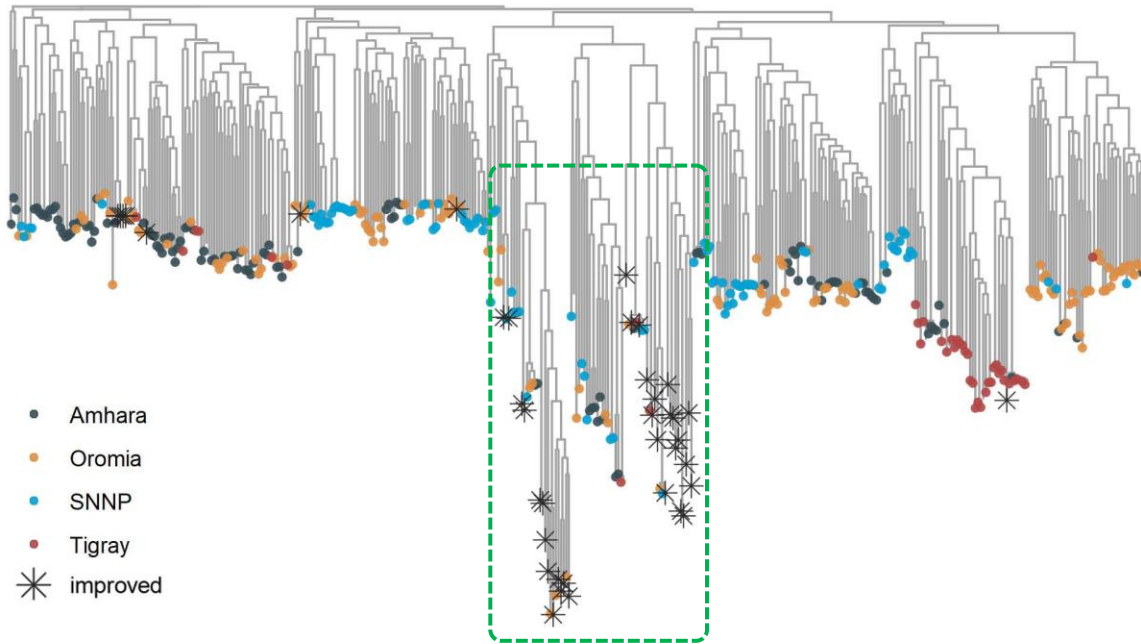
Extracting climatic features at sampling points



		Layer 1 	Layer 2 	...	Layer n 
sampling point 1					
sampling point 2					
...					
sampling point n					

Is there any evidence that is suggesting that possibly

Genetic diversity and structure

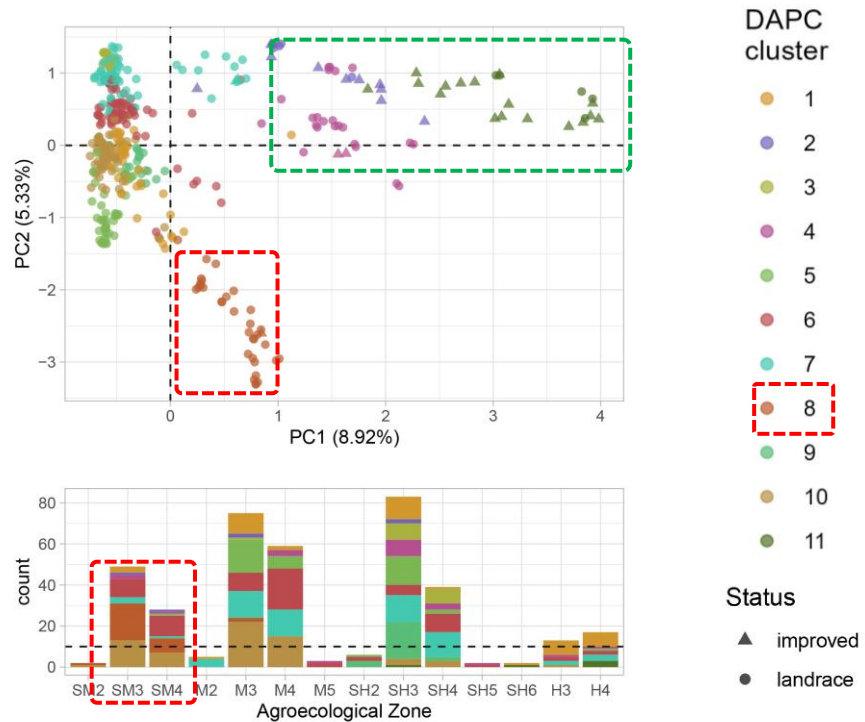
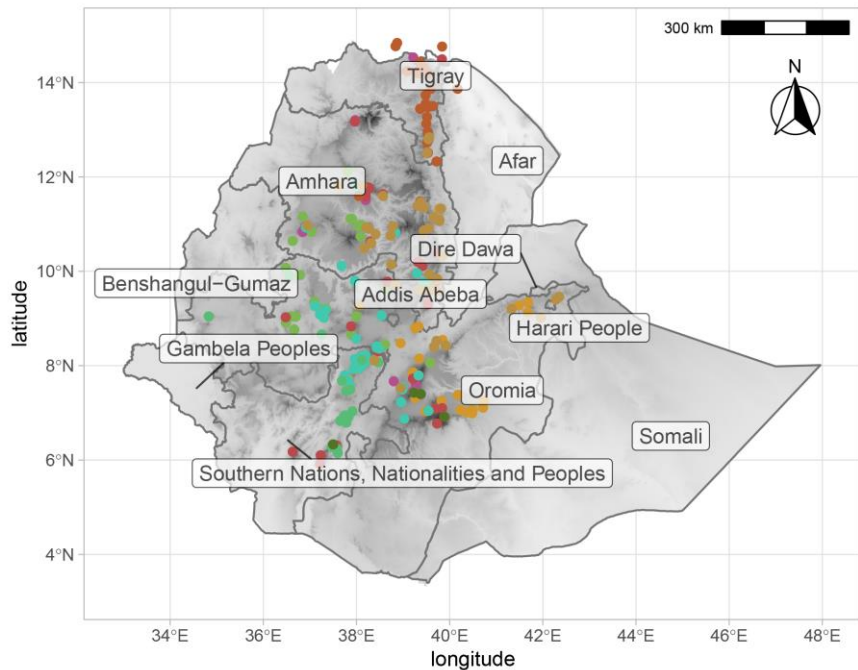


- 2064 K LD-pruned SNP to study structure
- Most **improved materials** set apart
- Regional provenance partially explain structure

Based on 2064 LD-pruned SNPs

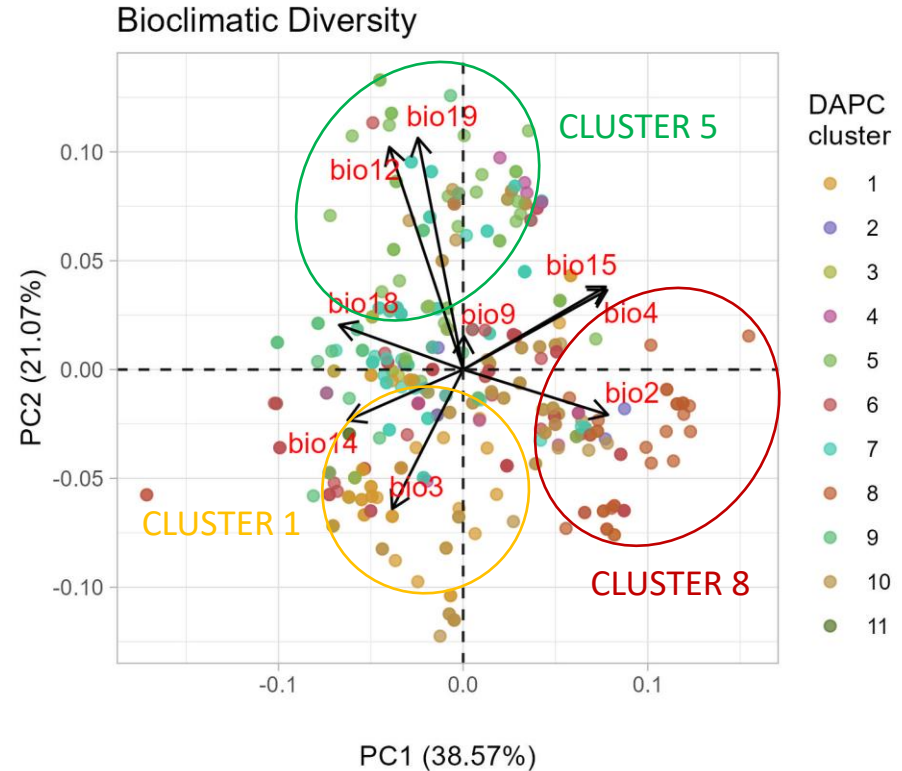
Genetic diversity

Based on 2064 LD-pruned SNPs



Bioclimatic and genetic diversity

- **383** georeferenced genotypes (landraces)
- Colors represent DAPC genetic clusters
- Precipitation and temperature patterns can separate some genetic clusters

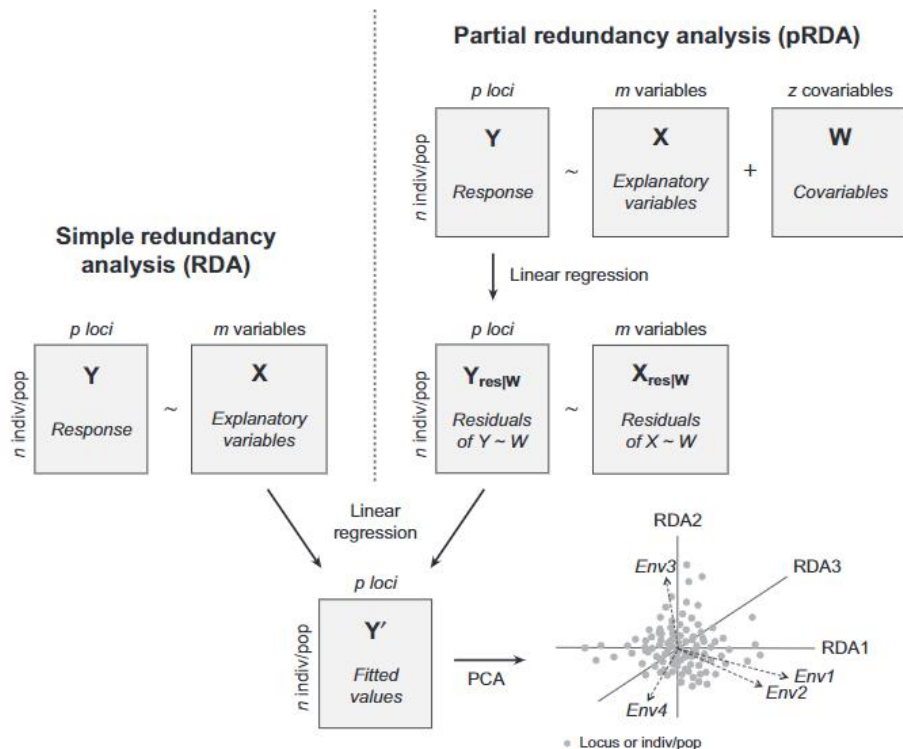


Redundancy analysis RDA

simple RDA: compute the fitted values (Y'); then conduct a PCA of the Y' matrix

partial RDA: partial linear regression is used to adjust the linear effects of X on Y accounting for the covariables W (population structure, spatial eigenvectors, ect)

- **RESPONSE (Y):** LD pruned-SNPs
- **EXPLANATORY (X):** bioclimatic, geographical, phenotypic



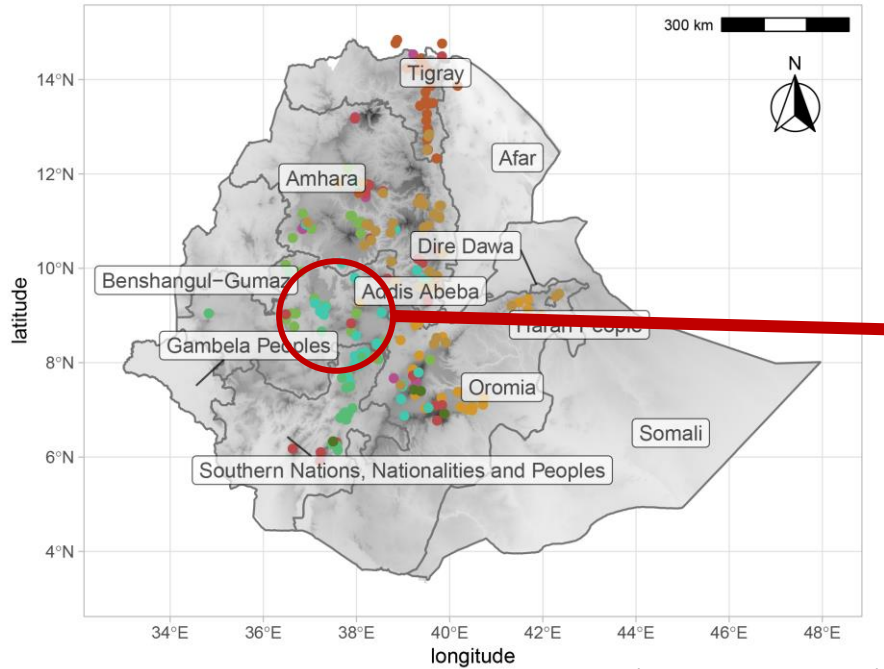
pRDA variance partitioning

- RDA models linear relationships among sets of **explanatory variables** (e.g. bio vars, geo, genetic PCs) and **response variables** (genomic variation, i.e. LD-pruned SNPs).

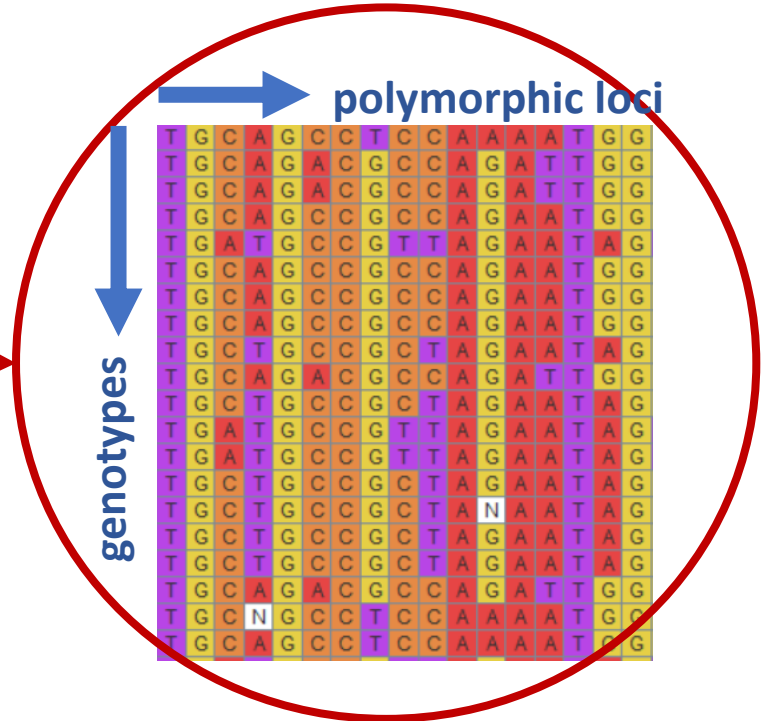
pRDA model	Inertia	R^2	adj R^2	$p(>F)$	Proportion of explainable variance	Proportion of total variance
Full model: $G \sim clim. + geog. + struct.$	247.1	.22	.19	.001***	1.00	.22
Pure climate: $G \sim clim. (geog. + struct.)$	54.7	.05	.03	.001***	.22	.05
Pure structure: $G \sim struct. (clim. + geog.)$	118.1	.11	.10	.001***	.48	.11
Pure geography: $G \sim geog. (clim. + struct.)$	13.6	.01	.01	.001***	.06	.01
Confounded climate/structure/geography	60.8				.25	.05
Total unexplained	860.5					.78
Total inertia	1107.6					1.00

*** $p \leq .001$.

Distribution of genetic diversity



(Caproni et al., 2023)

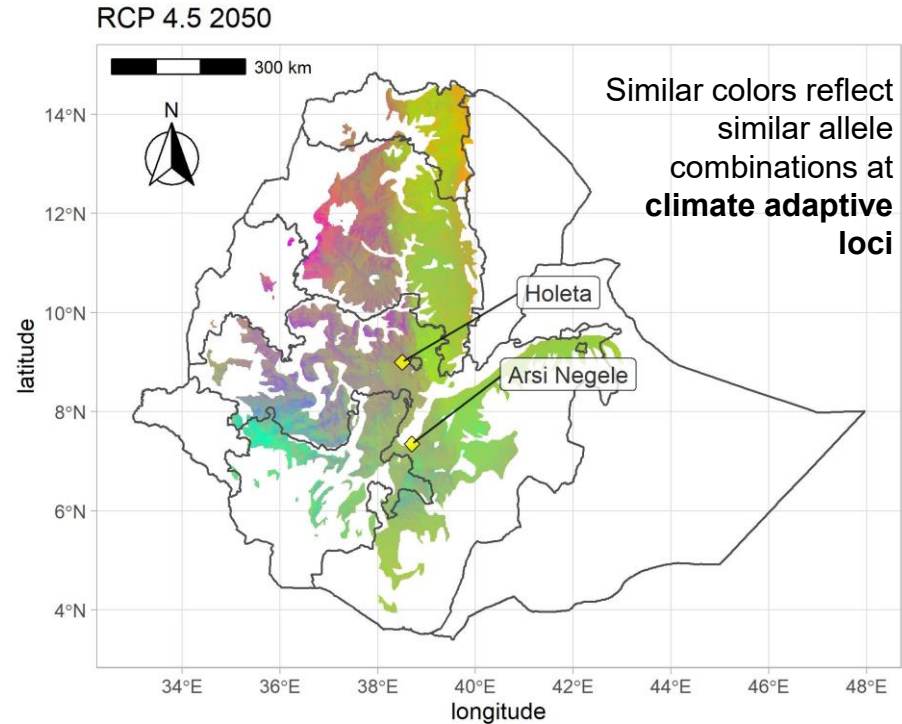
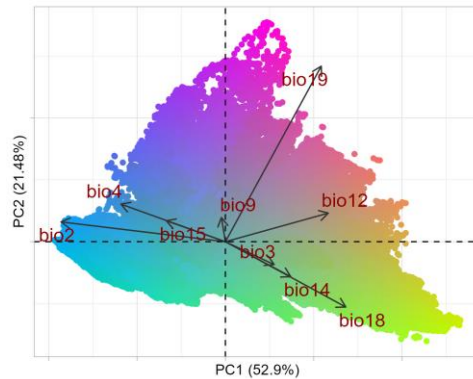


Modelling the genetic drivers of adaptation

Machine Learning tree-based approach (Gradient Forest)

- **RESPONSE VARIABLES** = LD pruned-SNPs
- **PREDICTORS** = Non-collinear historical bioclimatic vars + MEMs (geography)

PCA is used to summarize the allelic turnover across the landscape



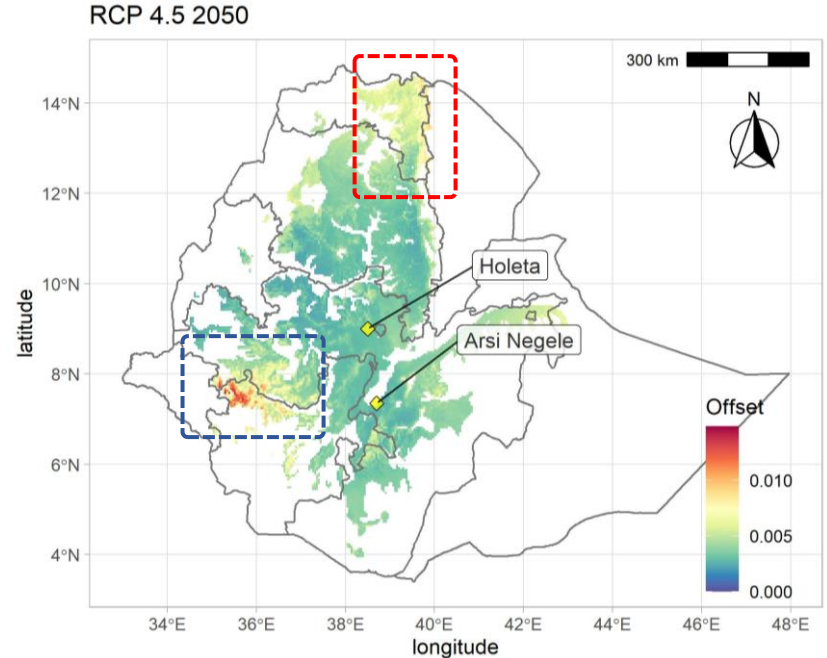
Estimating the level of predicted maladaptation (genomic vulnerability)

Calculated as the Euclidean distance (ED) between the allelic turnover under the historical and projected climate



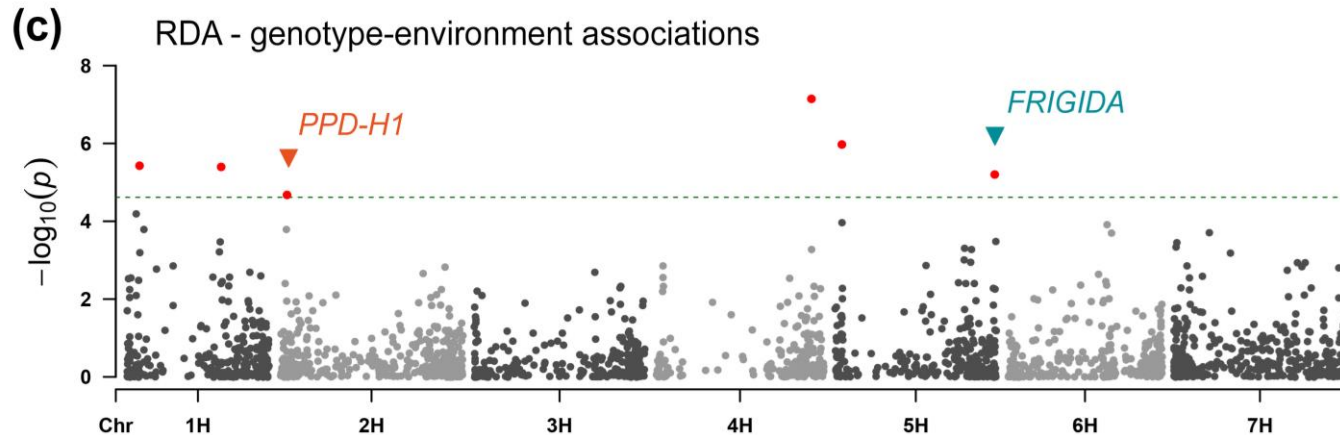
Vulnerable areas:

- **Western Tigray**
- **Area bordering Oromia and SNNP**



Genotype-environment associations (partial-RDA)

- **explanatory variables** noncollinear bioclimatic indicators
- **response variables** LD-pruned SNP (2064)



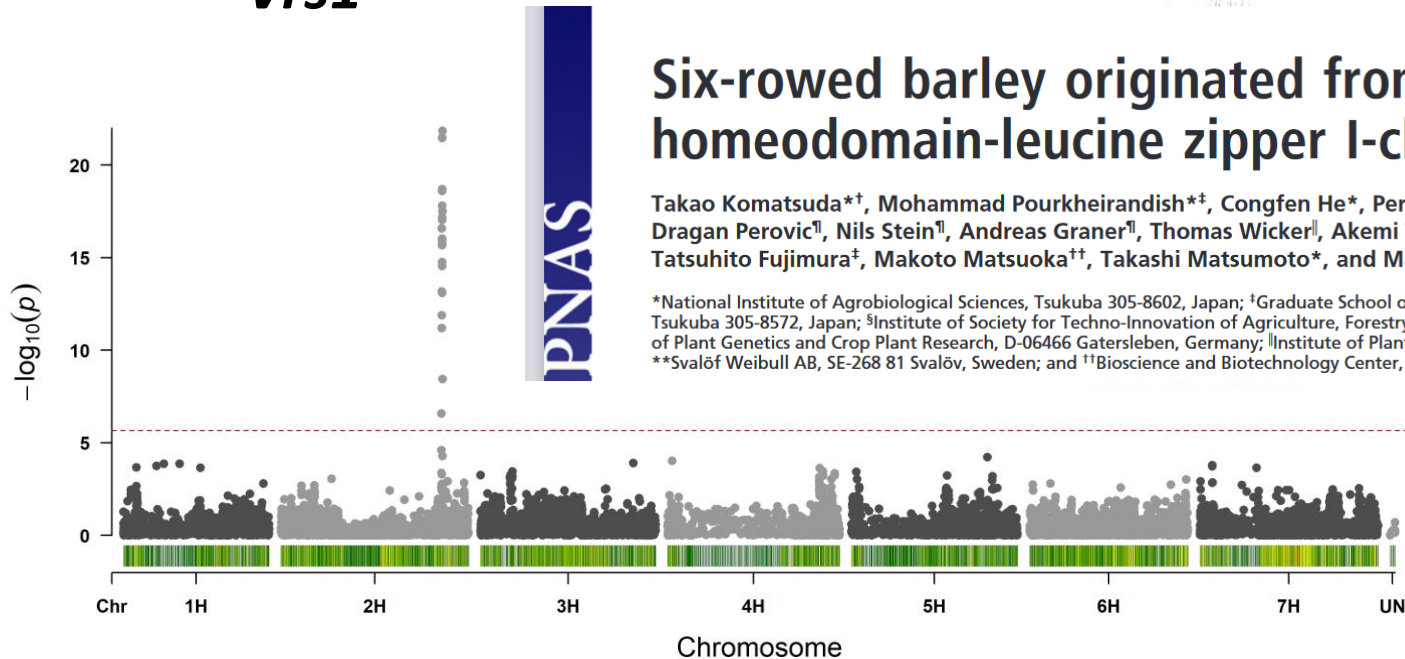
END

Spare slides

GWAS: proof of concept

Mapping lateral spikelet fertility

Vrs1



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^{**}, Tatsuhiro 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