Genetic mapping

Plant Genetic Resources (3502-470)

23 May 2025

Table of contents

Note: These notes need to be revised and made more concise.

1 Motivation

The phenotypic characterisation of PGR in field trials opens the possibility to identify genes, which influence interesting traits. Several types of methods are available for identifying underlying genes. They include linkage or QTL mapping based on crosses of few or multiple parents, or genome-wide association studies. Genome-wide association studies are a frequently used approach in genetic mapping because they use information of large numbers of accessions characterized in projects and are greatly facilitated by the sequencing or large-scale genotyping of genebank accessions.

In the following, genome-wide association studies are introduced as one approach for the characterization of plant genetic resources.

Identifiying large amount of associations problem that arises frequently in modern genomics data. The goal of genome-wide association studies (GWAS) is to unterstand the genetic of quantitative traits. In the following, we assume you have a basic understanding of the nature and characteristics of quantitative traits.

Studying the genetics of natural variation involves:

- Understanding the genetic architecture of traits of ecological and agricultural importance
- Identifying the genomic regions that control genetic variation

As genomics datasets become more common and sample sizes grow, the need for efficient statistical tests increases. GWAS tests for associations at many variants instead of some candidate genes and therefore hypothesis-free instead of hypothesis-driven with respect to the number and types of genes controlling phenotypic variation.

Figure 1 shows the key principle of GWAS: A set of plants (i.e., genebank accessions) are phenotyped. Phenotypic variation is indicated as the different colors. The same material is genotyped and the indivduals are ordered by their phenotypes. Associations between phenotypes and genotypic variation are identified using a statistical test.

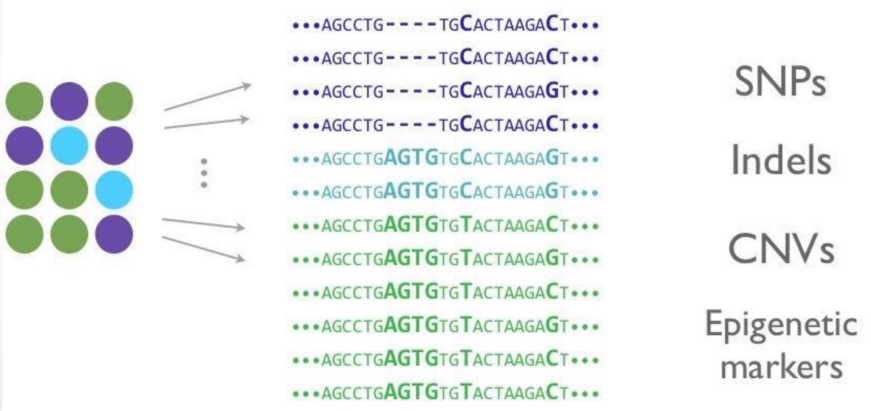


Figure 1 *– Principle of GWAS. Source:*

In quantitative traits, both genes and the environment determine the phenotype. For this reason, a good experimental design needs to be employed to be able to differentiate between the two factors, and to allow the identification of the genotype on the phenotypic vcariation.

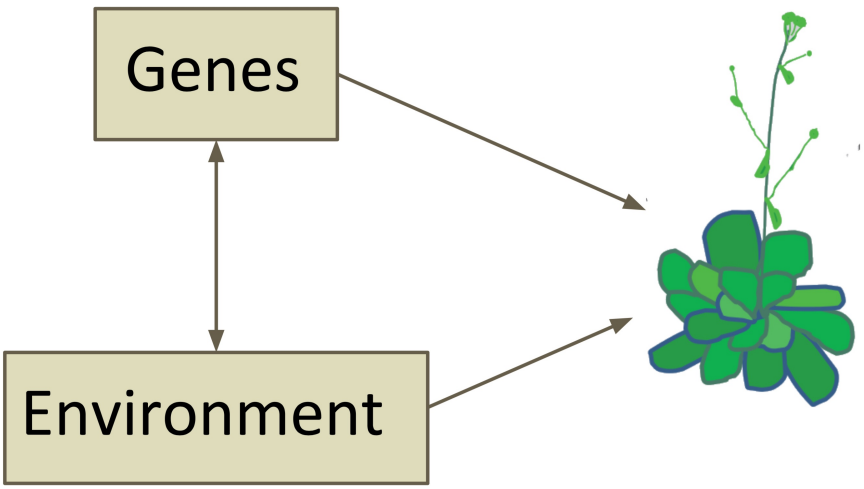


Figure 2 *– Genes and environment determine phenotype.*

In particular, Genotype by Environment (GxE) provide a challenge because different traits are influenced by different degrees by the environment or by the genes.

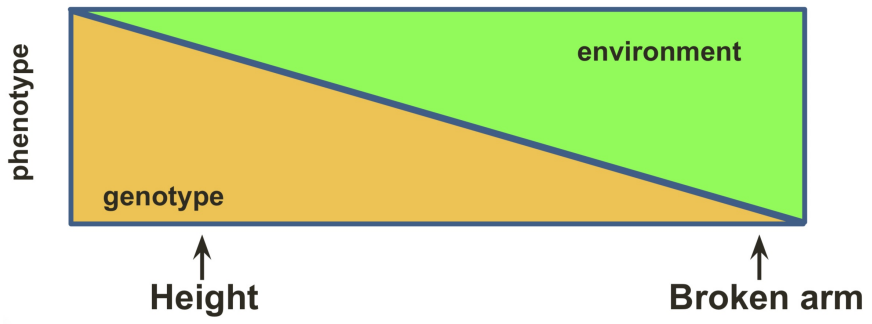


Figure 3 *– The range of genotype x environment (GxE) interactions.*

Figure 4 shows an example of a GWAS from the model plant *Arabidopsis thaliana* which allows to identify a genomic region that contains genetic variation for a trait in leaves.

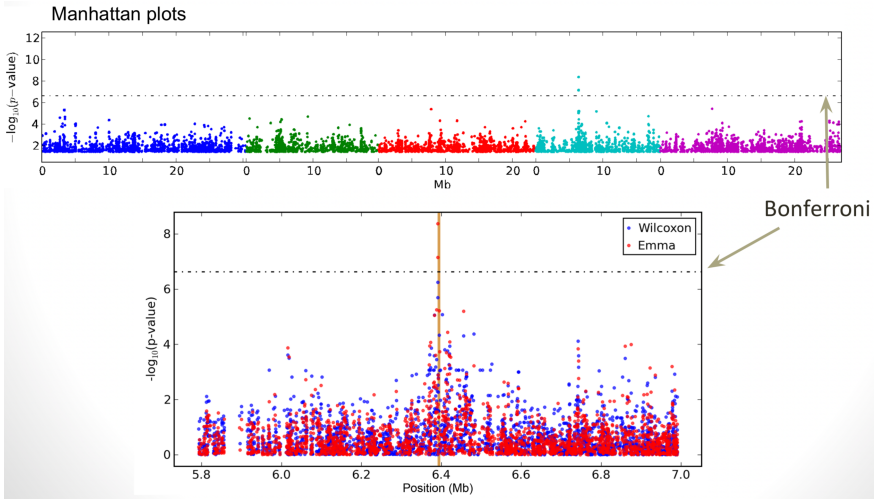


Figure 4 *– Example of a simple GWAS: Sodium concentration measured in* A. thaliana *leaves. Source: Ümit Seren (Ref: Atwell et al 2010?)*

1.1 Multiple testing problem

In GWAS a large number of marker tests are conducted, which leads to a multiple testing problem. Using a 5% significance threshold, we would expect 5% of the markers that have true marker effects of 0 to be significant in a statistical test of a null hypothesis. Bonferroni correction is a method for reducing the proportion of false positive. By assuming markers are independent we can obtain a conservative bound on the probability of rejecting the null hypothesis for one or more markers.

$$1 - P(T_1 \leq t, \ldots, T_m \leq t | H_0) \leq \alpha \quad (1)$$

for a given significance threshold α. It should be noted, however, that the Bonferroni correction is overly conservative for a large number of tests and therefore leads to a high proportion of false negatives.

Other common methods include adjusted Bonferroni correction depending on rank, and permutations. However because of the statistical problems with the Bonferroni correction, other statistical approaches such as the use of the false discovery rate (FDR) have been proposed.

1.2 GWAS in plants

GWAS are important in plant research. In contrast to GWAS in humans, which are often highly critized, GWAS in plants have the advantage that appropriate experimental design reduces the environmental variance and the error. In addition, GWAS can be easily complemented by linkage mapping that is based on the analysis of crosses and their offspring.

A disadvantage of GWAS studies in plants is that they are often based on small sample sizes. Therefore the power to identify and estimate the effect of variants with small effect is difficult. For example the classical study by Atwell et al. (2010) was based on 107 individuals only.

Meanwhile, for model plants like *A. thaliana* and model crops like rice or maize, numerous resources are available that allow to download and (re-)analyse data using GWAS.

1.3 Linkage disequilibrium

Neighboring markers will tend to be inherited together, causing linkage disequilibrium (LD) between the two markers (Figure 5).

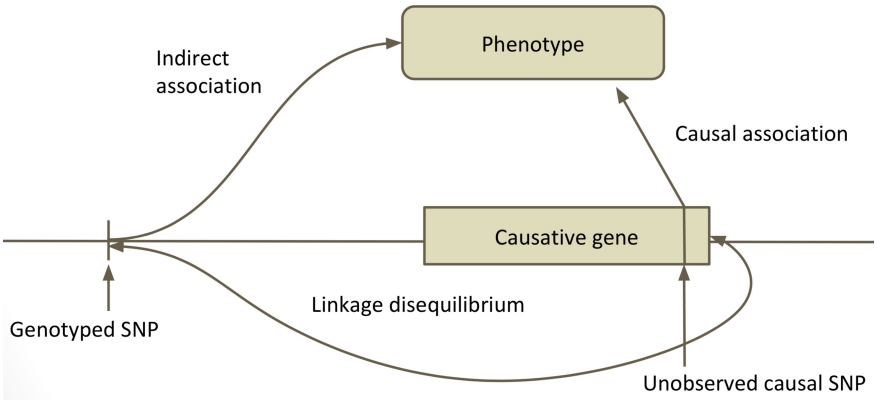


Figure 5 *– Importance of linkage disequilibrium*

Since LD causes correlations between markers, in a given population we expect a lot of redundancy in the genotypes. LD is voth and advantage and disadvantage at the same time GWAS. The advantage is that linked markers are sufficient to identify an association with a QTL, on the other hand it makes is difficult to identify the causal variant if too many markers in the vicinity are linked to the causal variant.

Because of LD there is an important relationship between marker numbers and sample size to consider because both determine the resolution of GWAS. This relationship is strongly influenced by the genome-wide level of LD.

1.4 Population structure

In natural populations, individuals tend to cluster in sub-populations according to their geographic origin. Methods for the inference of population structure like principal components analysis (PCA) (Figure 6) allow to identify subpopulations.

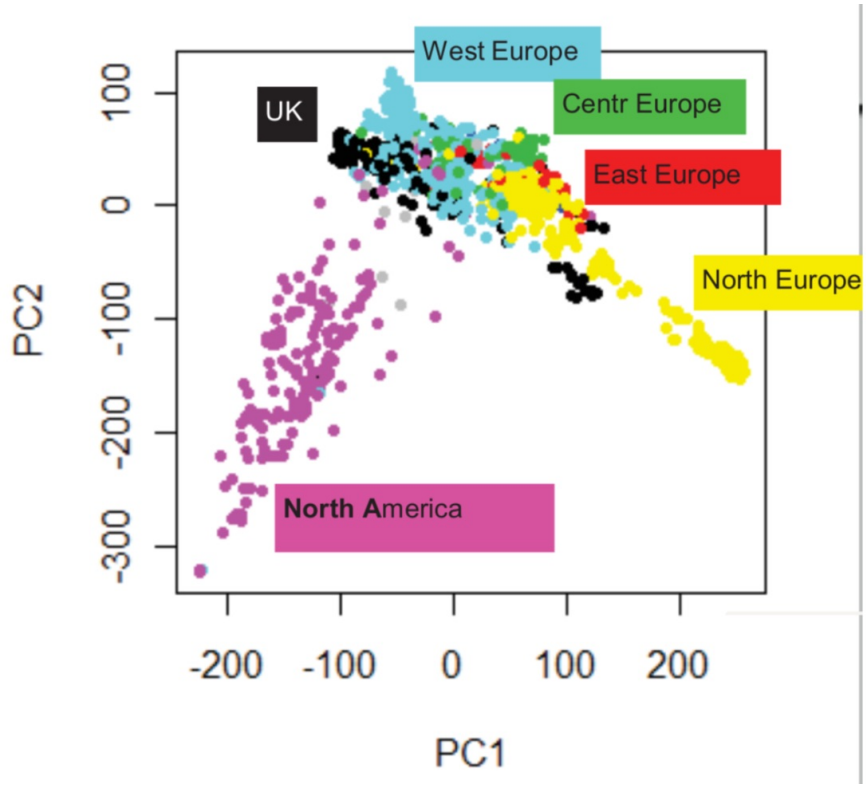


Figure 6 *– Principle components analysis (PCA).*

Confounding due to population structure may arise if it correlates with the trait in question (Figure 7)

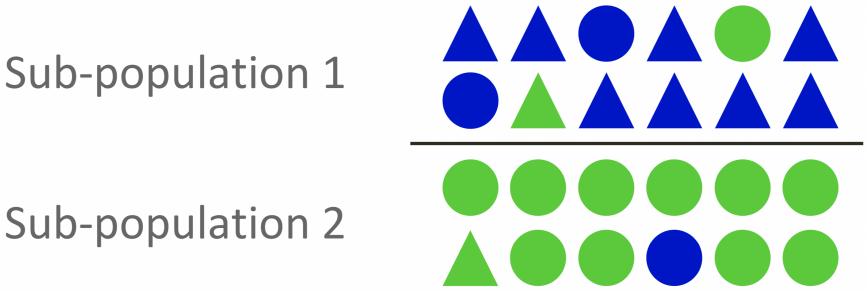


Figure 7 *– Confounding when trait variation correlates with structure.*

Any variant which is fixed for different alleles in each sub- population will show an association.

Humans Genetic marker for skin color may also be associated with malaria resistance because the trait is correlated with population structure.

Arabidopsis thaliana Flowering time is correlated with latitude (Figure 8).

Disease resistance is not correlated with latitude.

Population structure is reflected in long-range LD: A strong population structure in a sample also leads to long-range LD in the sample (Figure 9).

Implications for association studies:

- Test statistics is inflated
- High false positive rate

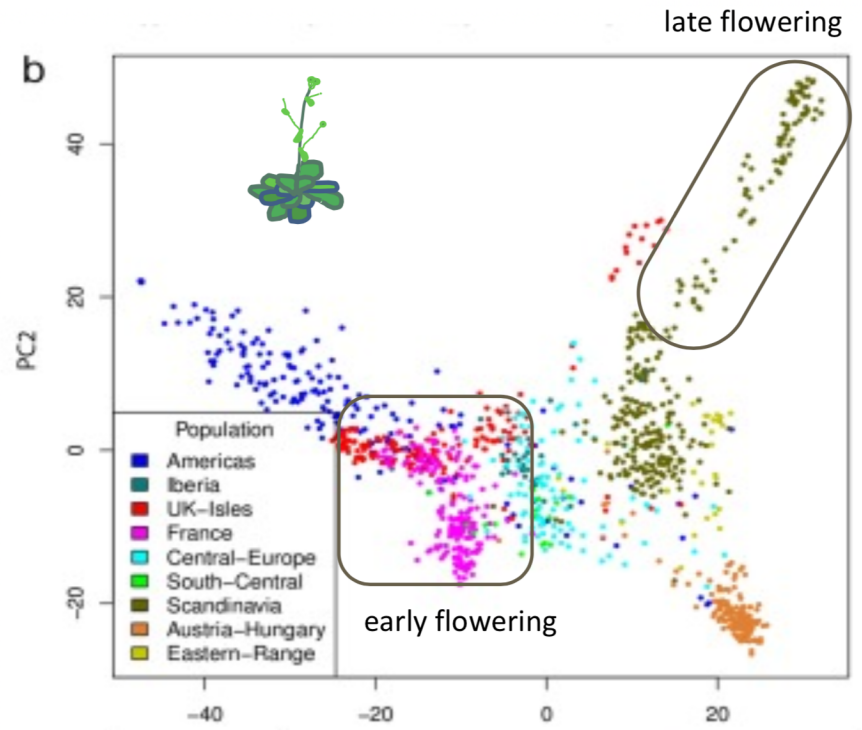
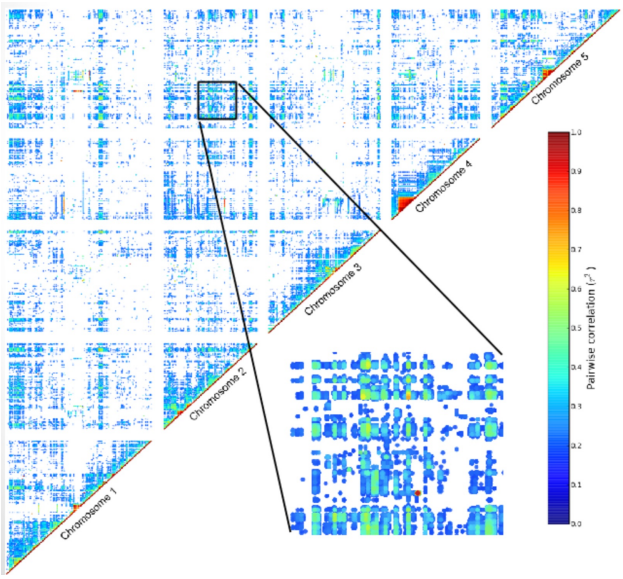


Figure 8 *– Arabidopsis flowering time latitude*



Linkage disequilibrium in *A. thaliana*, 214K SNPs and 1307 accessions.

Figure 9 *– Population structure in LD.*

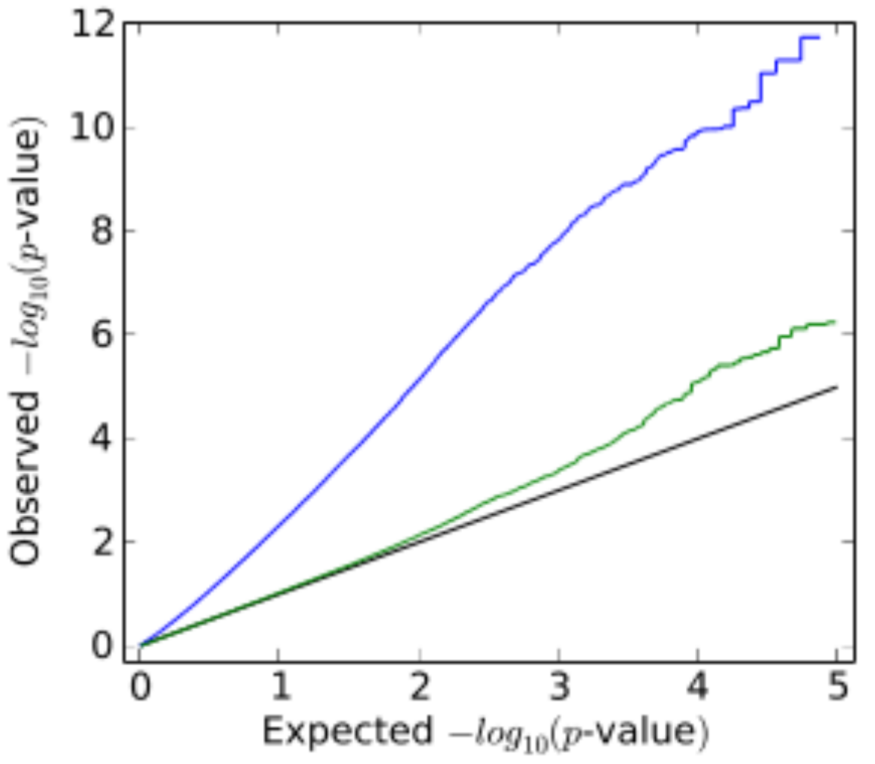


Figure 10 *– Expected vs. observed test statistic*

For this reason GWAS studies require a control to reduce the confounding of marker-trait associations with the population structure (Figure 11)

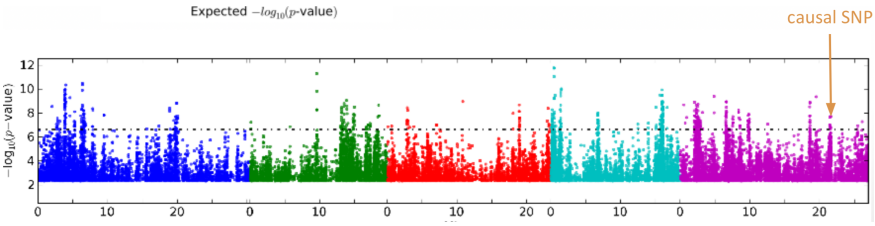


Figure 11 *– Confounding signals without control.*

1.5 Correcting for population structure

Several approaches to correct for population structure were described.

Genomic control: Scale down the test-statistic so that its median becomes the expected median. Heavily used, but does not solve the problem (Devlin and Roeder, 1999).

Structured association: Pritchard et al. (2000)

PCA approach: Accounting for structure using the first n principle components of the genotype matrix (Price et al., 2006). However, when population structure is very complex, e.g., in *A. thaliana*, too many PCs are needed.

Mixed model approach: Model the genotype effect as a random term in a mixed model, by explicitly describing the covariance structure between the individuals (Kang et al., 2008; Yu et al., 2006).

2 Methods for GWAS based on linear models

2.1 Linear model

A linear model generally refers to linear regression models in statistics.

$$y_i = \sum_{j=1}^{P} \beta_j x_{ij} + \epsilon_i \quad (2)$$

and

$$Y = X'\beta + \epsilon \quad (3)$$

where

- Y consists of the phenotype values or case-control status for N individuals
- X is the $N \times P$ genotype matrix, consisting of P genetic variants, e.g., SNPs
- β is a vector of P effects for the genetic variants
- ϵ is the error or noise term.

The linear model can be tested with several tests.

The t-test and the F-test assume that the underlying distribution is Gaussian. For a single SNP this means that the conditional phenotype distribution is Gaussian. This is obviously not true for many traits. For traits that are not normally distributed, non-parametric tests can be used. For biallelic SNPs, which are coded as binary markers (i.e., encoded as 0 and 1), the Wilcoxon rank sum test or a Fisher's exact test can be used. For more general markers (or heterozygous genotypes encoded as 0, 1, 2) a Kruskal-Wallis, Wilcoxon rank sum test or the Spearman rank correlation can be used.

2.2 Linear mixed model (LMM)

It is necessary to consider that a simple linear model and non-parametric tests do not account for population structure.

For this reason, a linear mixed model can be used in which the population structure is modeled as fixed effect and the SNP effect as random effect.

The simplest version of the model is

$$Y = X\beta + u + \epsilon, u \sim N(0, \sigma_g K), \epsilon \sim N(0, \sigma_e I) \quad (4)$$

where

- Y typically consists of the phenotype values, or case-control status for N individuals
- X is the $N \times P$ genotype matrix, consisting of P genetic variants, e.g. SNPs
- u is the random effect of the mixed model with var(u) = $\sigma_g K$
- K is the $N \times N$ kinship matrix inferred from genotypes
- β is a vector of P effects for the genetic variants
- ϵ is a $N \times N$ matrix of residual effects with var(ϵ) = $\sigma_e I$

The model was proposed for association mapping by Yu et al, 2006.

Kinship

The kinship measures the degree of relatedness and is in general different from the covariance matrix. It is estimated usng either a pedigree (family relationships) or (nowadays) using genome-wide genotype data. For the estimation of kinship from pedigree data, it is usually assumed that the ancestral founders of the population are unrelated. However, they can be sensitive to confounding by cryptic relatedness. Alternatively, the kinship can be estimated from genotype data. It needs to be considered that genotype data may be incomplete. In addition, weights or scaling of genotypes can impact the kinship. In the model plant *Arabidopsis thaliana* an identity-by-state (IBS) matrix works quite well to estimate the kinship (Atwell et al., 2010; Zhao et al., 2005).

2.3 Implementations of the linear mixed model (LMM)

There are different implementations of the LMM that differ mainly by the algorithmic complexity and the speed of computations.

Original implementation: EMMA (Kang et al., 2008)

Problem: $O(PN^3)$ - 1 GWAS in 1 day (500k individuals)

Approximate methods $O(PN^2)$:

- GRAMMAR (Aulchenko et al., 2007) http://www.genabel.org/packages/GenABEL
- P3D (Zhang et al., 2010) http://www.maizegenetics.net/#!tassel/c17q9
- EMMAX (Kang et al., 2010) http://genetics.cs.ucla.edu/emmax/

Exact methods:

- FaST LMM (Lippert et al., 2011) http://mscompbio.codeplex.com/
- GEMMA (Zhou et al., 2012) http://www.xzlab.org/software.html

This is too slow for large samples (>20,000 individuals), i.e. exactly the sample sizes where one might expect to see most gains.

- BOLT-LMM (Loh et al., 2015), $O(PN)$ https://data.broadinstitute.org/alkesgroup/BOLT-LMM/?

LMM reduce the test statistic inflation (Figure 13).

LMM also reduce false positives. Figure 14.

2.4 Advanced Mixed Models

The mixed-model performs pretty well, but GWAS power remain limited and need to be improved:

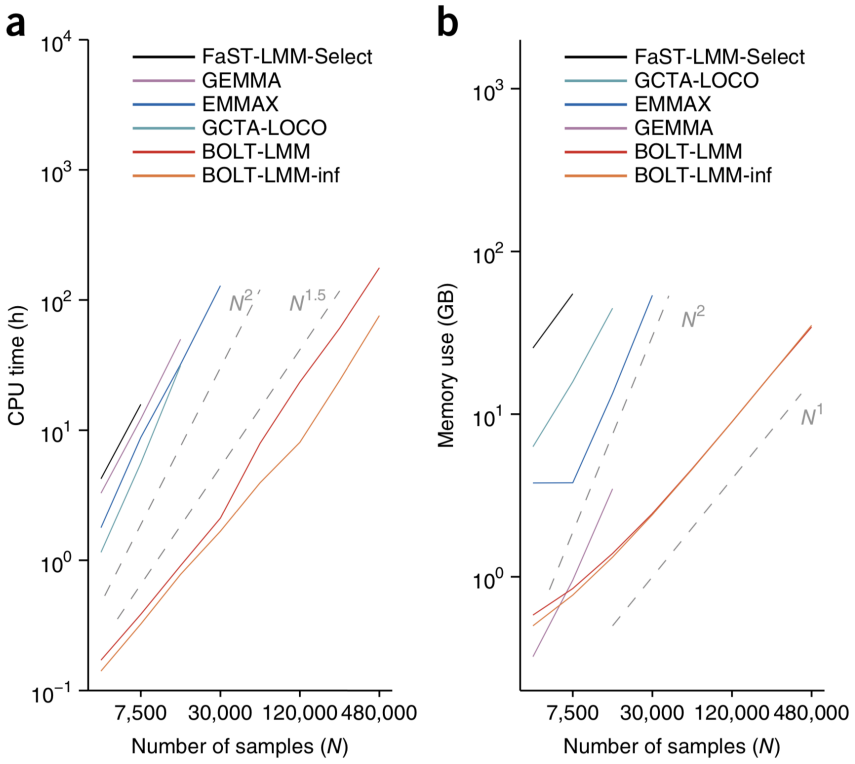


Figure 12 *– Bolt-LMM performance*

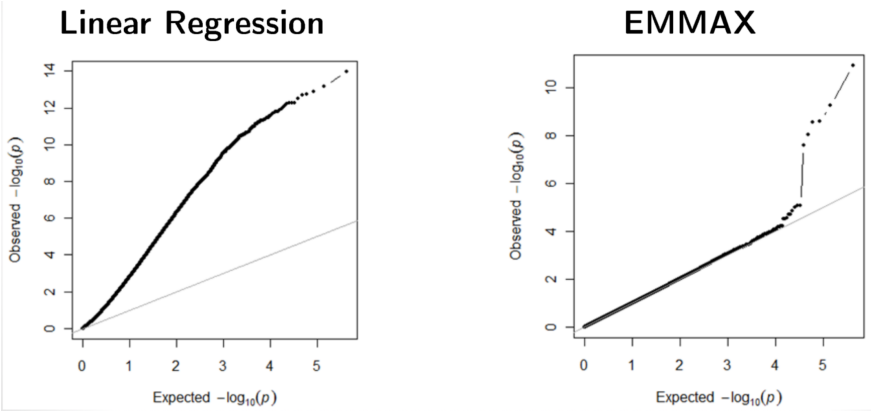


Figure 13 *– Test statistic inflation.*

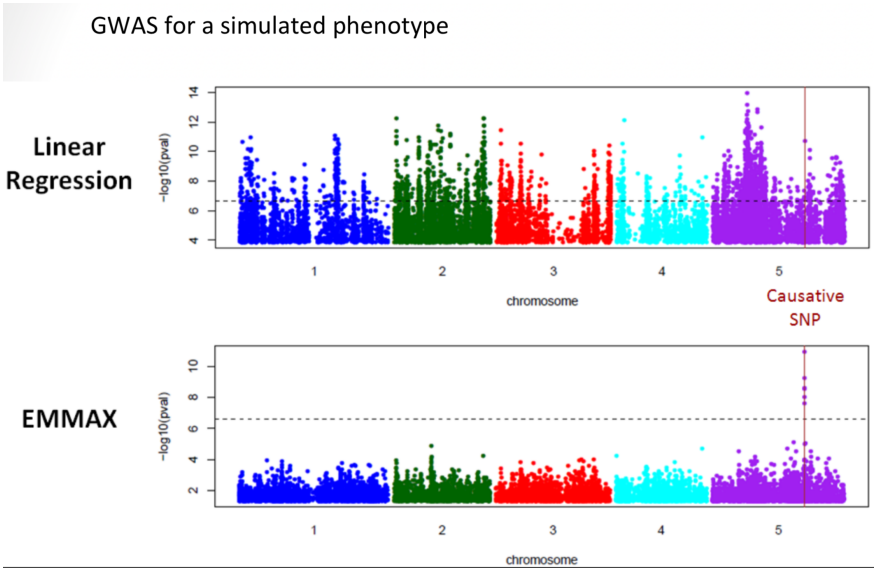


Figure 14 *– False positives.*

Multi Locus Mixed Model (MLMM),

- Single SNP tests are wrong model for polygenic traits
- Increase in power compared to single locus models
- Detection of new associations in published datasets
- Identification of particular cases of (synthetic associations) and/or allelic heterogeneity or linkage between causative polymorphisms.
- Combining correlated traits in a single model should thus increase detection power

Multi Trait Mixed Model (MTMM), Korte et al. (2012)

Traits are often correlated due to pleiotropy (shared genetics). When multiple phenotypes consists in a single trait measure in multiple environments, plasticity can be studies through the assessment of GxE interaction.

2.5 Caveats & Problems

Accounting for population structure does not alway work:

Accounting for population structure does not alway work:

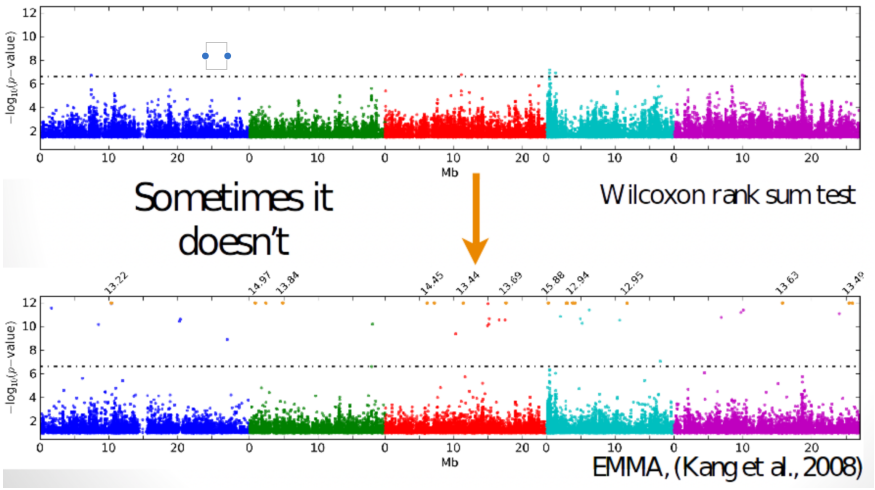


Figure 15 *– Problems with population structure as confounding factor*

Sometimes it is difficult to decide which peaks are significant. One solution the the permutation of p-values.

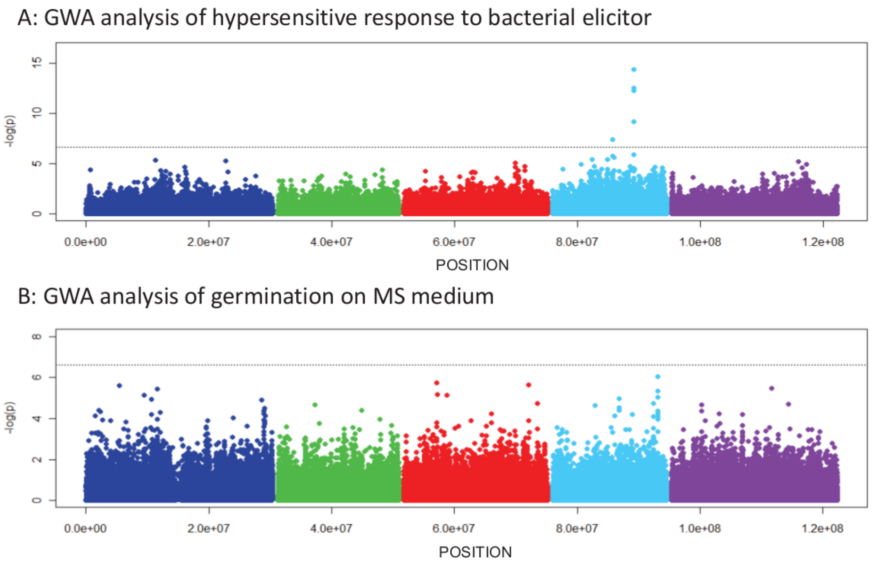


Figure 16 *– Difficult identification of peaks depending on the environment.*

Peaks are complex and make it difficult to pinpoint causative site

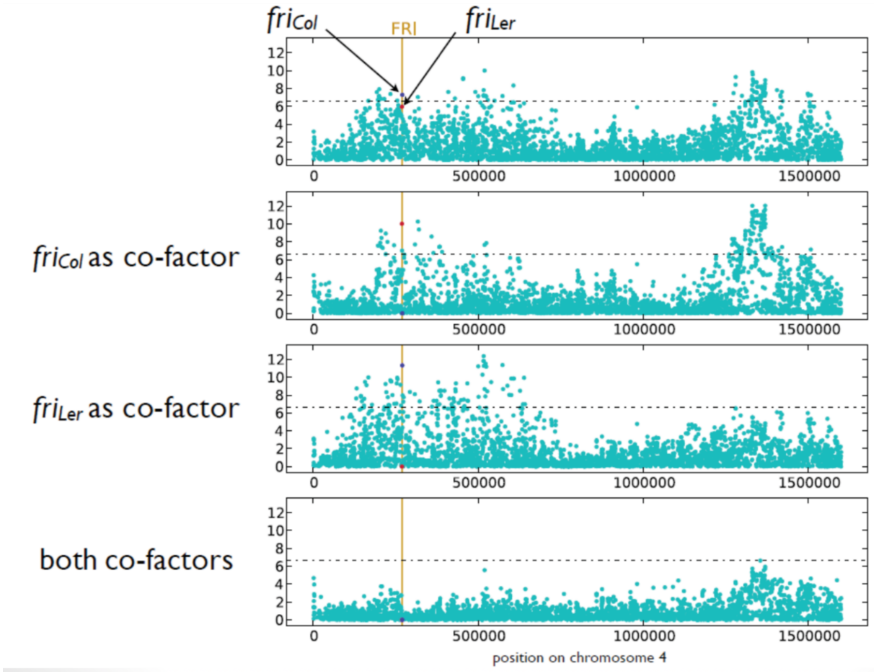


Figure 17 *– Complex peaks*

Condition under which GWAS will be positively misleading:

- More than one causal factor
- Epistasis
- Correlation between causal factors and unlinked non-causal markers

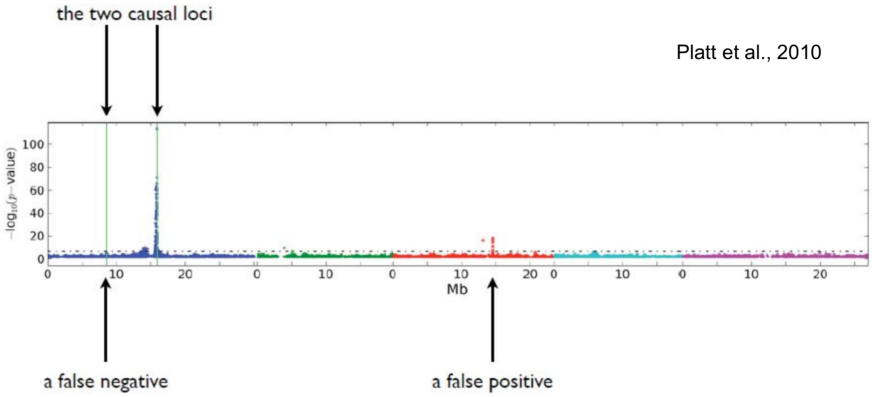


Figure 18 *– More than one causal factor*

Different associations for different subsets (e.g., flowering time at 10°C).

- Highly heritable, easy to measure, polygenic trait
- 925 worldwide accessions
- Flowering time greatly varies in different populations

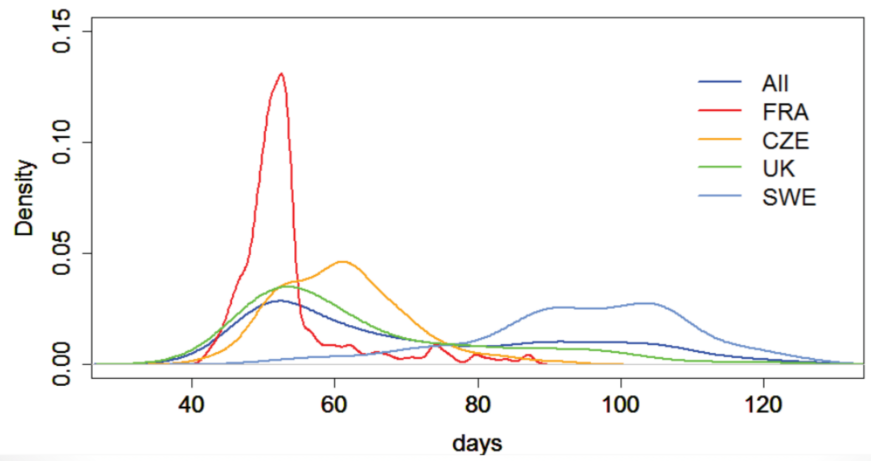


Figure 19 *– Flowering time variation in different subgroups.*

Significance and effect size differ dramatically in different subsets for the following reasons:

- False positives
- Effect depends on genetic background (Epistasis)
- Differences in allele frequency of the causal marker
- Artefact of LMM

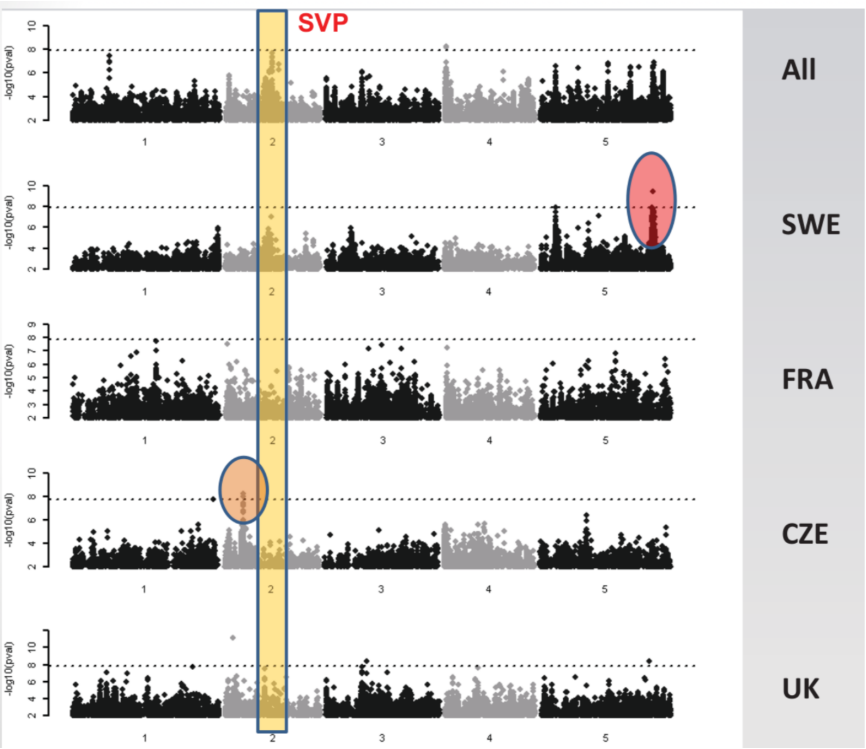


Figure 20 *– Different associations in different populations*

3 Examples of GWAS in plant genetic resources

There are multiple examples where interesting genes GWAS were conducted in plant genetic resources to identify novel useful genetic variation after phenotypic evaluation.

In the following, two examples are shown.

The barley collection in the German genebank at IPK is one of the largest barley collections worldwide. It has been genotyped using a method genotyping-by-sequencing (GBS), which is a reduced representation sequencing method because only about 2% of the genome is being sequenced. Using GBS, a total of 22,626 IPK genebank accessions were genotyped, of which 19,778 were domesticated barleys (*Hordeum vulgare*) and 1,140 wild barleys (*Hordeum spontaneum*) (Figure 21). Overall a, total of 171,263 SNPs were identified by GBS.

Table 1 | Number of segregating SNPs detected by GBS in wild and domesticated barley samples

	Number of SNPs	SNPs with MAF ≥ 1%	SNPs with MAF ≥ 5%
All samples ($n = 22,626$)	171,263	23,908	15,683
Domesticated barleys ($n = 19,778$)	76,102	22,356	15,872
Wild barleys ($n = 1,140$)	127,408	46,392	20,511

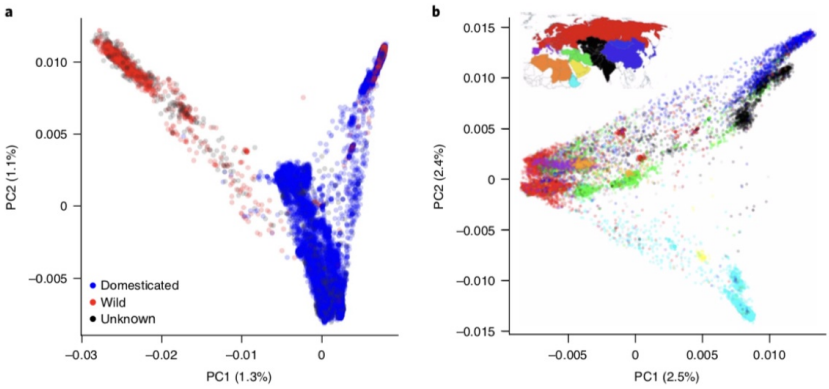


Figure 21 – *Summary of the genotyping of the IPK barley genebank accessions. Source: Milner et al. (2018)*

A population structure analysis with PCA of both wild and domesticated barley shows that a strong differentiation and within cultivated barley a strong differentiation between genetic regions (Figure 21 a and b). A comparison of a PCA of the complete IPK collection and the International Barley Core Collection further shows that the IBCC is a good representation of the total diversity of cultivated barley.

The IPCC was then used for large-scale GWAS analysis using a variety of domestication and improvement traits. Domestication traits were row type genes (2 or 6 row) and hull adherence, whereas agronomic traits included flowering time and disease resistance. The GWAS analyses detected several major QTL genes, of which some were known before, but also uncovered novel genes controlling these traits, which may be useful for further utilization in breeding (Figure 23).

One example of such a gene is a gene for awn roughness, which is very different between wild and domesticated barley (Figure 24 a). Rough awns evolved allow grains to adhere to the fur of animals, for example, and therefore contribute to the geographic distribution (and ultimately, fitness) of wild barley. However, this trait is superfluous in domesticated barley, because seeds are not shattering anymore and the rough awns cause pain to the farmers during the harvest. A detailed analysis of the best hit in the

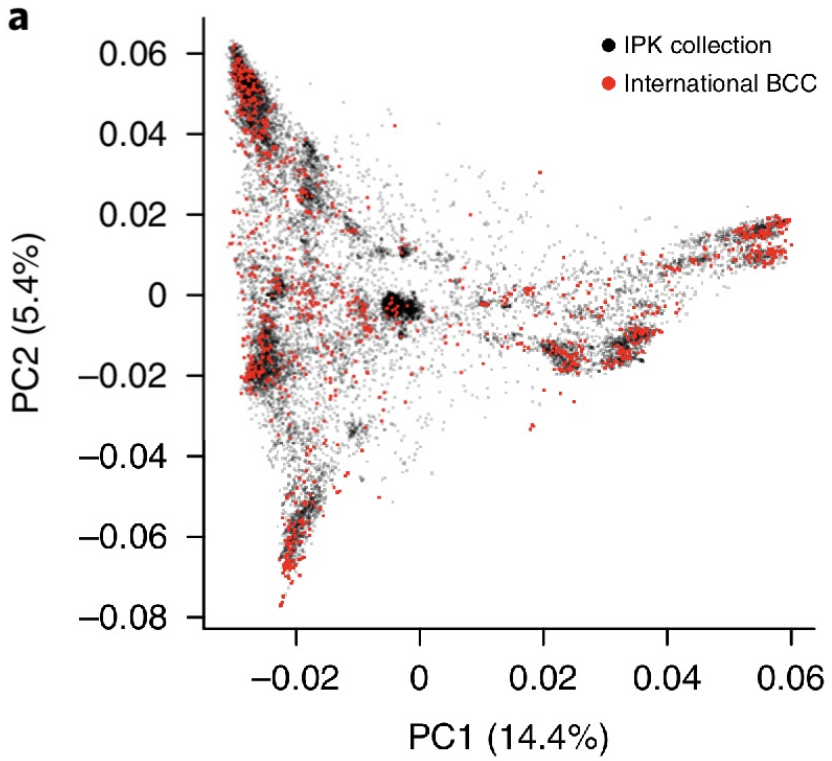


Figure 22 *– PCA analysis of the IPK cultivated barley and the International Barley Core Collection (IPCC). Source: Milner et al. (2018)*

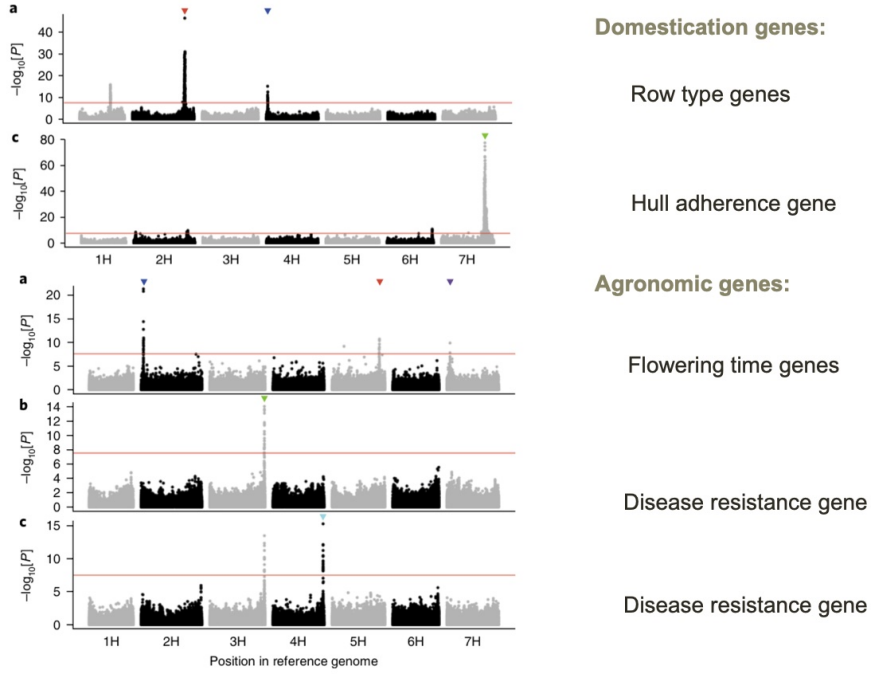


Figure 23 *– Results of a GWAS for various domestication and agronomic traits. Source: Milner et al. (2018)*

GWAS analysis and in a related type of analysis called bulk segregant analysis identified the Rough awn 1 gene as causal gene for this trait. The causal mutation is a splice-site mutation, which changes the processed messenger RNA (mRNA) of this gene (Figure 24 b and c). This gene was validated as causal gene in a mutagenized barley population in which individuals with soft awns carried different mutations that also disrupted the funtion of the Rough awn 1 gene (Figure 24 d and e).

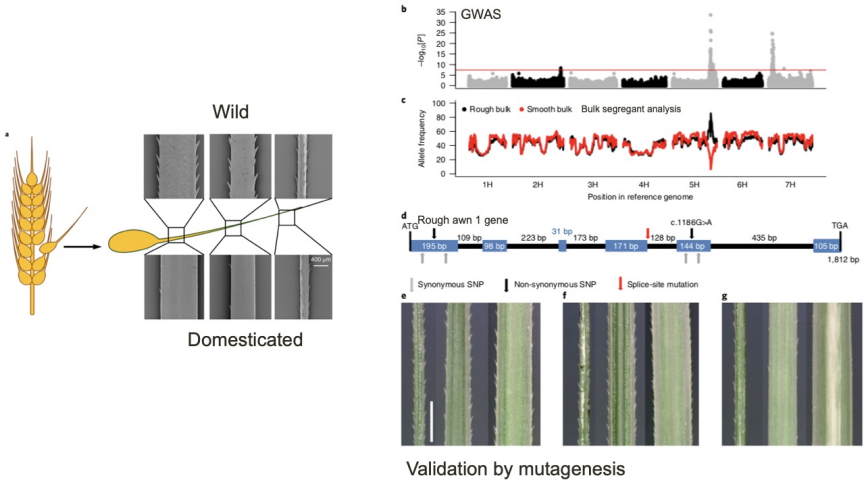


Figure 24 *– Functional analysis of the Rough awn 1 gene of barley. a) Phenotype of the trait in wild and domesticated barley. b) and c) Mapping of the gene in the IPCC using GWAS and Bulk segregant analysis. d) Annotation of the gene and location of the natural mutation (red) and the mutation induced by mutagenesis (black). e) Validation of the gene by phenotypic analysis of various mutants. Source: Milner et al. (2018)*

A similar study was carried out for an ancient crop of the Americas, amaranth. There are three types of grain amaranth (*Amaranthus caudatus*, *A. cruentus* and *A. hypochondriacus*) that were all domesticated from the same wild species, *Amaranthus hybridus*. A key domestication traits are white seeds in comparison to red seeds of wild amaranth. The whole genome resequencing of a set of wild and domesticated amaranths, and the subsequent GWAS, Bulk segregant analysis and QTL analysis of the trait revealed the same set of two genomic regions associated with the trait. The genomic region with the stronger statistical support harbors a so-called MYB-like transcription factor gene. Closely related genes (homologs) of this gene in other species were shown to be involved in controlling seed or fruit color by regulating the anthocyanin or betalain pathways. In the case of amaranth, further functional validation is required to test the hypothesis that the Amaranth MYB-like gene is a causal gene for seed color.

These examples (and many others in the scientific literature) show that GWAS of plant genetic resources is a powerful approach to identify many genes that may reveal the history of domestication or can be used in modern plant breeding programs using approaches like marker-assisted selection.

4 Key concepts

□ Confounding effect of population structure □ Linear (mixed) model □ Allelic heterogeneity

5 Summary

- Genetic mapping with GWAS and related methods is a powerful approach to understand the genetics of phenotypic variation in plant genetic resources
- GWAS is based on the association between genotypic and phenotypic variation
- Population structure present in the analysed population may cause confounding and many false positive associations
- Multiple methods for GWAS are available.
- Methods based on linear mixed models are particular powerful because they can correct for population structure
- GWAS is challenging because epistatic interactions, allelic heterogeneity and GWAS on subsamples interfere with a simple interpretation of GWAS results

6 Further reading

- Cortes et al. (2021) - An accessible and timely review of GWAS in plants

7 Study questions

1. What are the key differences between GWAS and QTL linkage methods for genetic mapping?
2. What are the advantages and disadvantages of each of the two methods?
3. Why doese population structure in a sample cause many false positive associations if it is not corrected by the model used for analysis?
4. If there is a perfect correlation between a phenotypic trait of interest and population structure: Would GWAS be able to identify the causal gene, and would other methods be able to identify the gene controlling the trait?
5. Why is allelic heterogeneity a problem in GWAS studies?
6. Why is it useful and interesting to carry out GWAS in plant genetic resources?
7. Which strategies can be used to further utilized significant trait-marker analysis from GWAS?

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, Meaux J de, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M. 2010. Genome-wide association study of 107 phenotypes in

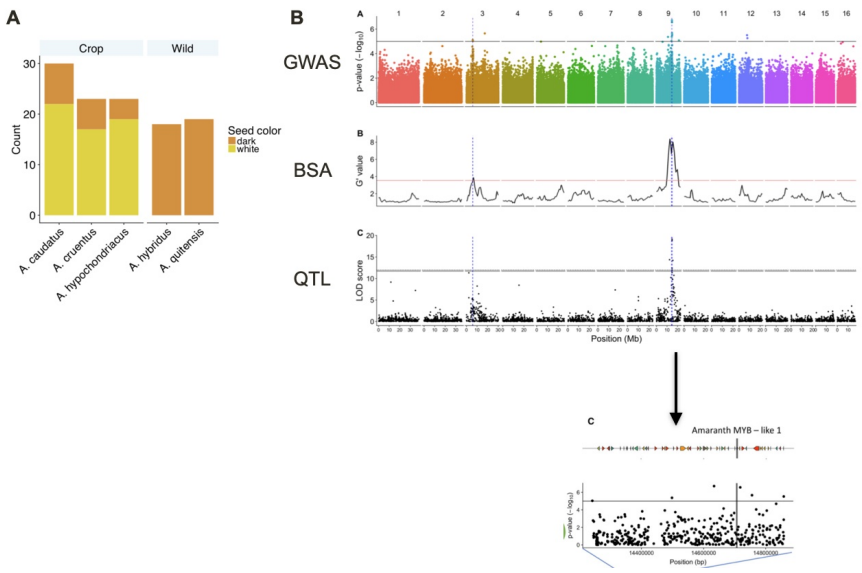


Figure 25 – *Genetic analysis of seed color in wild and domesticated amaranth species. (A) Seed color is predominant in the domesticated amaranth species* A. caudatus*,* A. cruentus*,* A. hypchondriacus*, whereas dark (red) seed color is the only seed color in the wild amaranth species* A. hybridus *and* A. quitensis*. (B) Genetic mapping of the genes causing seed color by GWAS, BSA and QTL (linkage) mapping approaches and identification of a candidate gene for seed color. Source: Stetter et al. (2020)*

Arabidopsis thaliana inbred lines. *Nature* **465**:627–631. doi:10.1038/nature08800

Cortes LT, Zhang Z, Yu J. 2021. Status and prospects of genome-wide association studies in plants. *The Plant Genome* **14**:e20077. doi:https://doi.org/10.1002/tpg2.20077

Devlin B, Roeder K. 1999. Genomic Control for Association Studies. *Biometrics* **55**:997–1004. doi:10.1111/j.0006-341X.1999.00997.x

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**:1709–1723. doi:10.1534/genetics.107.080101

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics* **44**:1066–1071. doi:10.1038/ng.2376

Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo G, Xu D, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao Y, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N. 2018. Genebank genomics highlights the diversity of a global barley collection. *Nature Genetics*. doi:10.1038/s41588-018-0266-x

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**:904–909. doi:10.1038/ng1847

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association Mapping in Structured Populations. *The American Journal of Human Genetics* **67**:170–181. doi:10.1086/302959

Stetter MG, Vidal-Villarejo M, Schmid KJ. 2020. Parallel Seed Color

Adaptation during Multiple Domestication Attempts of an Ancient New World Grain. *Molecular Biology and Evolution* **37**:1407–1419. doi:10.1093/molbev/msz304

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**:203–208. doi:10.1038/ng1702

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. 2005. An Arabidopsis Example of Association Mapping in Structured Samples. *PLoS Genetics* **preprint**:e4. doi:10.1371/journal.pgen.0030004.eor