Understanding genotype-phenotype relationships using the model plant *Arabidopsis thaliana*

Arthur Korte

01.03.2021
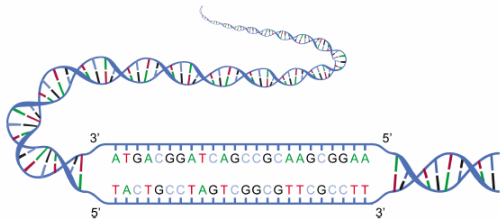
@arthurkorte

arthurkorte
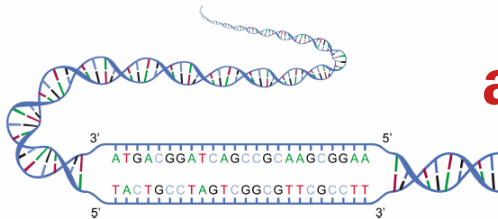
Giersch - Summer School & International Conference

Genotype ⇌ **Phenotype**

Genotype → ← **Phenotype**

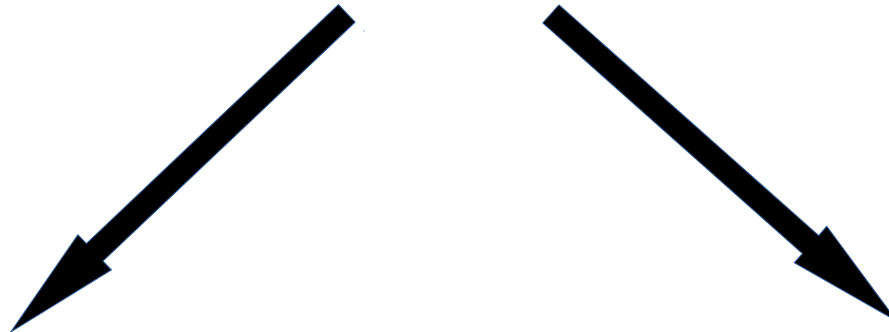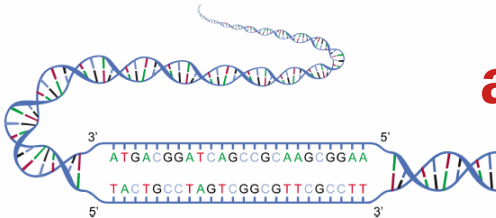**Genome-wide association studies (GWAS)**

Environment

Genotype **Phenotype**

Genome-wide association studies (GWAS)

Arabidopsis thaliana

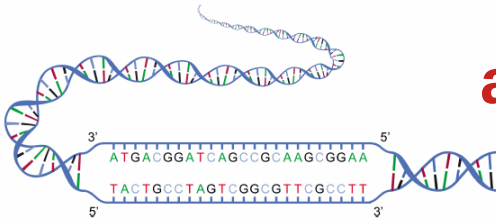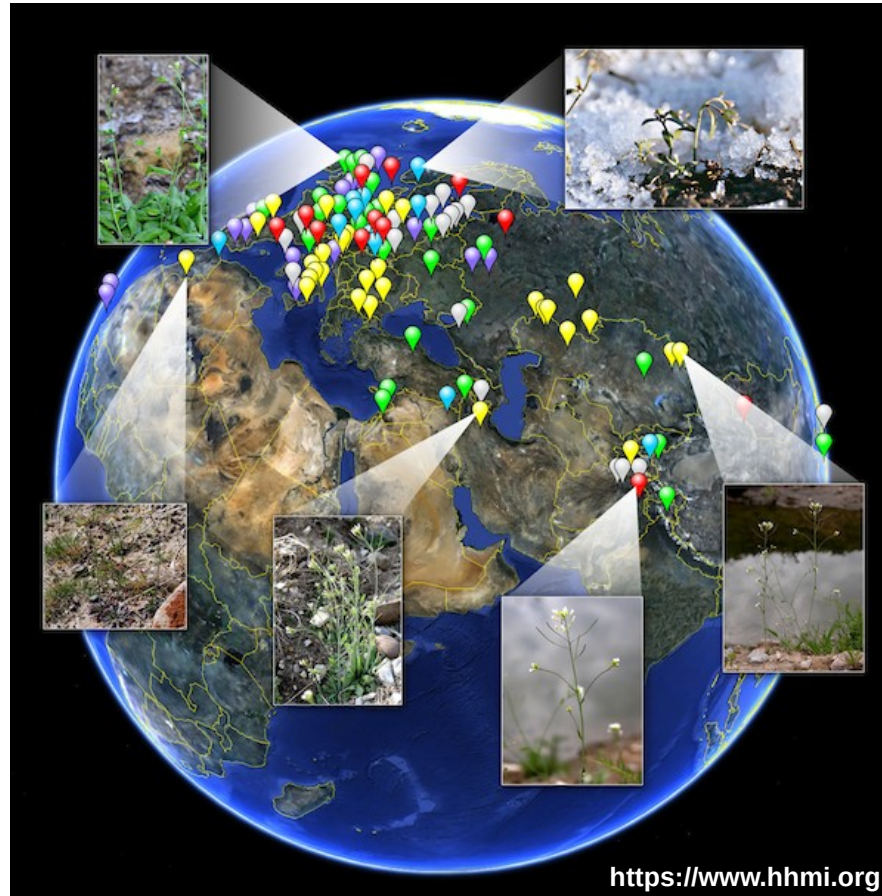

https://www.hhmi.org

**Arabidopsis is a prime model for Evolutionary Genomics
(small genome, adapted to the local environment, occurs as natural inbred)**

The 1001 Genomes Project

1001 Genomes
A Catalog of *Arabidopsis thaliana* Genetic Variation

Home   Data Providers   Accessions   Tools   Software   Data Center   About

- **Full genome sequence of 1,135 different natural ecotypes**

10M SNPs and 500k structural variants in a genome with ~125 M Bp
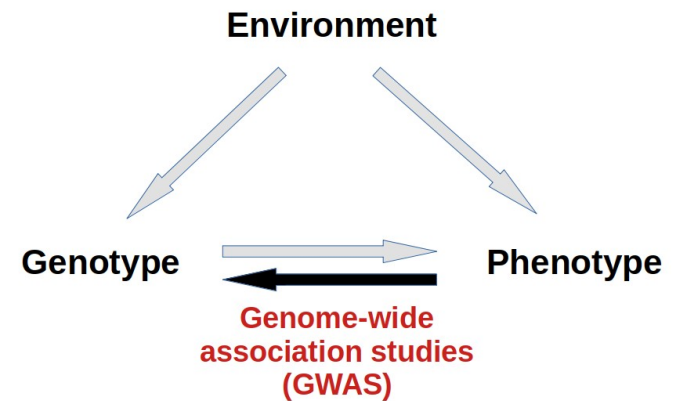
www.1001genomes.org
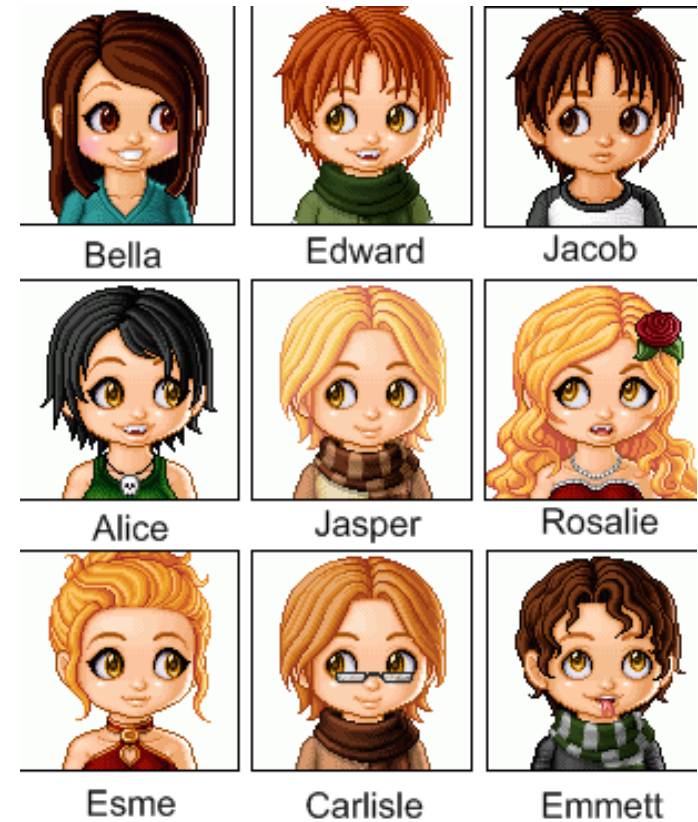
Arabidopsis thaliana

Genome-wide association studies (GWAS)
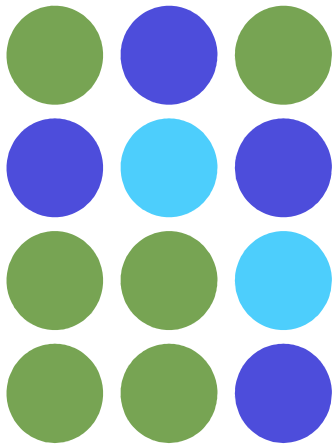

Associations between genotype and phenotype
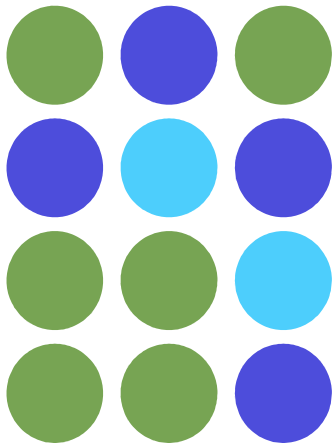
Genome-wide association studies (GWAS)

Associations between genotype and phenotype

Introduction: GWAS

```
•••AGCCTG----TGCACTAAGACT•••
•••AGCCTG----TGCACTAAGACT•••
•••AGCCTG----TGCACTAAGAGT•••
•••AGCCTG----TGCACTAAGACT•••
•••AGCCTGAGTGTGCACTAAGAGT•••
•••AGCCTGAGTGTGCACTAAGAGT•••
•••AGCCTGAGTGTGTACTAAGACT•••
•••AGCCTGAGTGTGTACTAAGAGT•••
•••AGCCTGAGTGTGTACTAAGACT•••
•••AGCCTGAGTGTGTACTAAGAGT•••
•••AGCCTGAGTGTGTACTAAGACT•••
•••AGCCTGAGTGTGTACTAAGACT•••
```

Introduction: GWAS

Introduction: GWAS

Calculate a statistical test for each SNP using linear mixed models

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

We get a p-value for each SNP that can be plotted in a Manhattan plot
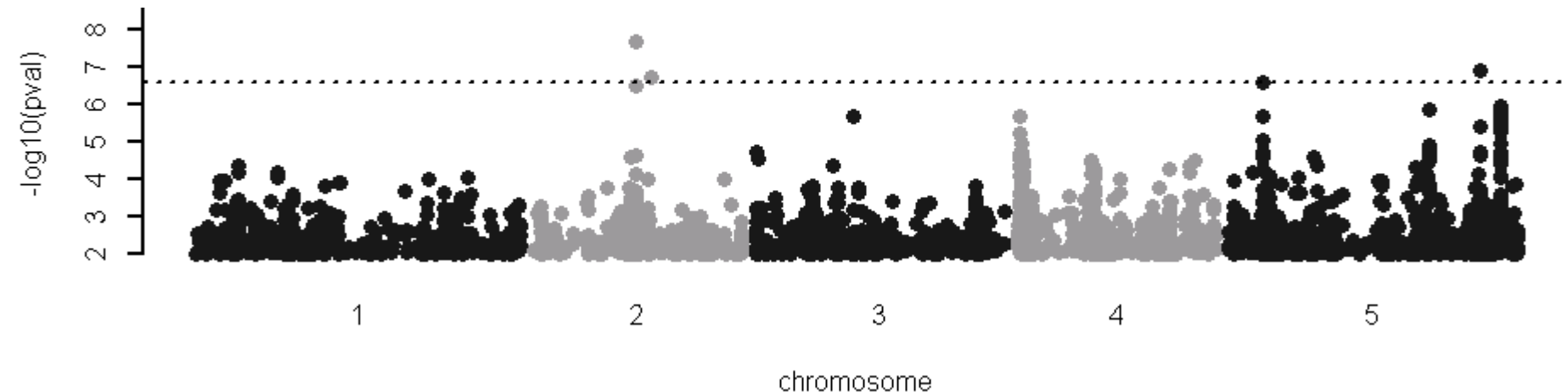
Goal: To identify markers that are significantly associated with a given trait permutation-based significance threshold (Freudenthal et al. 2019)

AraPHENO and AraGWAS

Seren *et al.* NAR (2017), Togninalli *et al.* NAR (2018) & Togninalli *et al.* NAR (2020)

GWAS on Covid-19 severeness
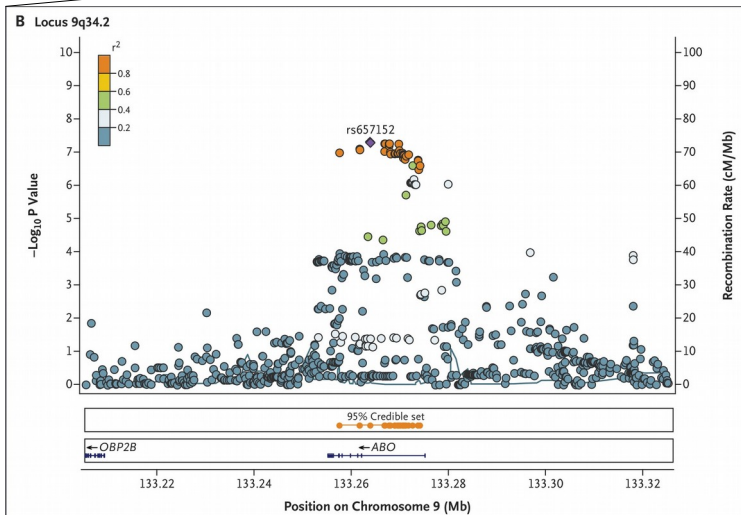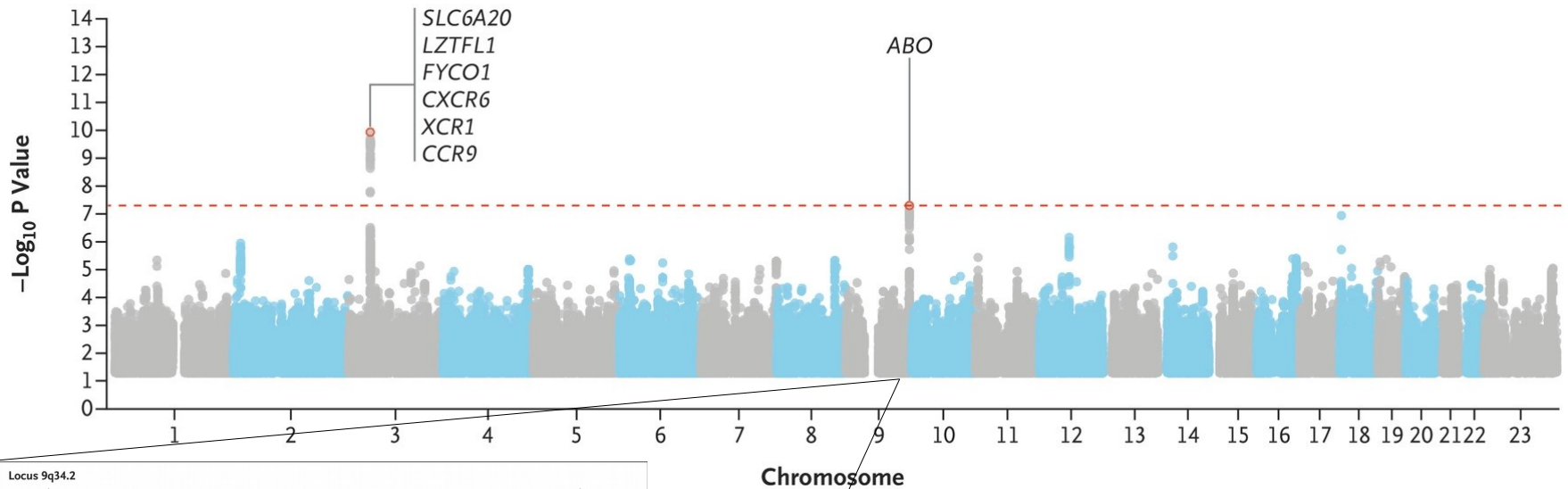
The Severe Covid-19 GWAS Group
The New England Journal of Medicine (2020)

Gene x environment interactions
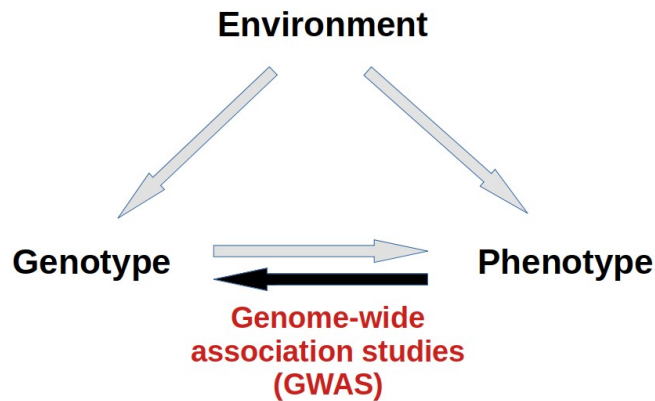
Environment

Genotype → **Phenotype**

Genome-wide association studies (GWAS)

GWAS : Genome - Phenotype

A mixed-model approach for genome-wide association studies of correlated traits in structured populations

Arthur Korte[1,*], Bjarni J. Vilhjálmsson[1,2,*], Vincent Segura[1,3,*], Alexander Platt[1,2], Quan Long[1], and Magnus Nordborg[1,2]

[1] Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria.

[2] Molecular and Computational Biology, University of Southern California, Los Angeles, California, United States of America.

[3] INRA, UR0588, F-45075 Orléans, France.

Abstract

Genome-wide association studies (GWAS) are a standard approach for studying the genetics of natural variation. A major concern in GWAS is the need to account for the complicated dependence-structure of the data both between loci as well as between individuals. Mixed models have emerged as a general and flexible approach for correcting for population structure in GWAS. Here we extend this linear mixed model approach to carry out GWAS of correlated phenotypes, deriving a fully parameterized multi-trait mixed model (MTMM) that considers both the within-trait and between-trait variance components simultaneously for multiple traits. We apply this to human cohort data for correlated blood lipid traits from the Northern Finland Birth Cohort 1966, and demonstrate greatly increased power to detect pleiotropic loci that affect more than one blood lipid trait. We also apply this to an *Arabidopsis* dataset for flowering measurements in two different locations, identifying loci whose effect depends on the environment.

Introduction

The Omnics cascade

Environment

Genotype → Phenotype

Genome-wide association studies (GWAS)

GWAS : Genome - Phenotype

GENOME
What can happen

TRANSCRIPTOME
What appears to be happening

PROTEOME
What makes it happen

METABOLOME
What is happening

PHENOTYPE

stable

instable

The Omnics cascade

Full genomes >1000 ecotypes
(1001 genomes consortium, 2016)

GENOME
What can happen

TRANSCRIPTOME
What appears to be happening

PROTEOME
What makes it happen

METABOLOME
What is happening
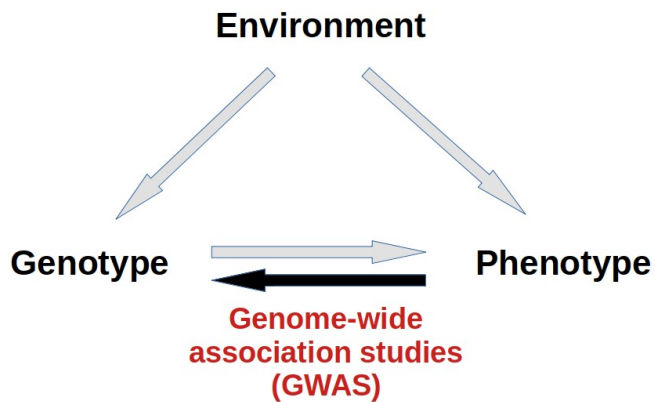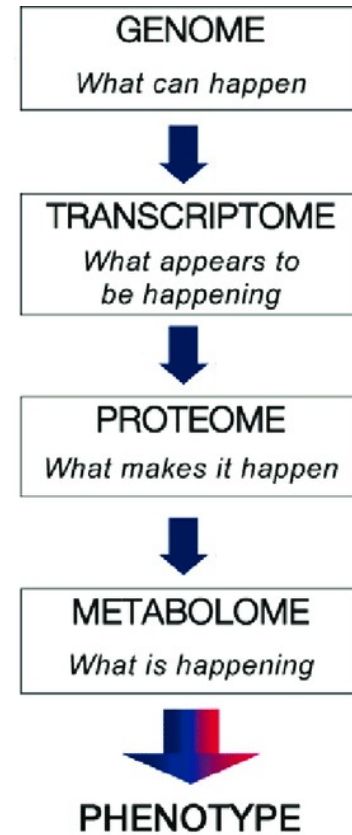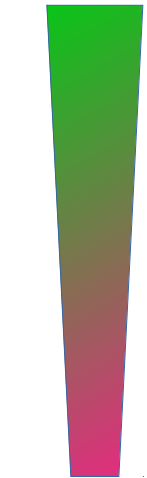
PHENOTYPE

stable

instable

> 500 different phenotypes

(Seren *et al.,*2017 & Togninalli *et al. 2020)
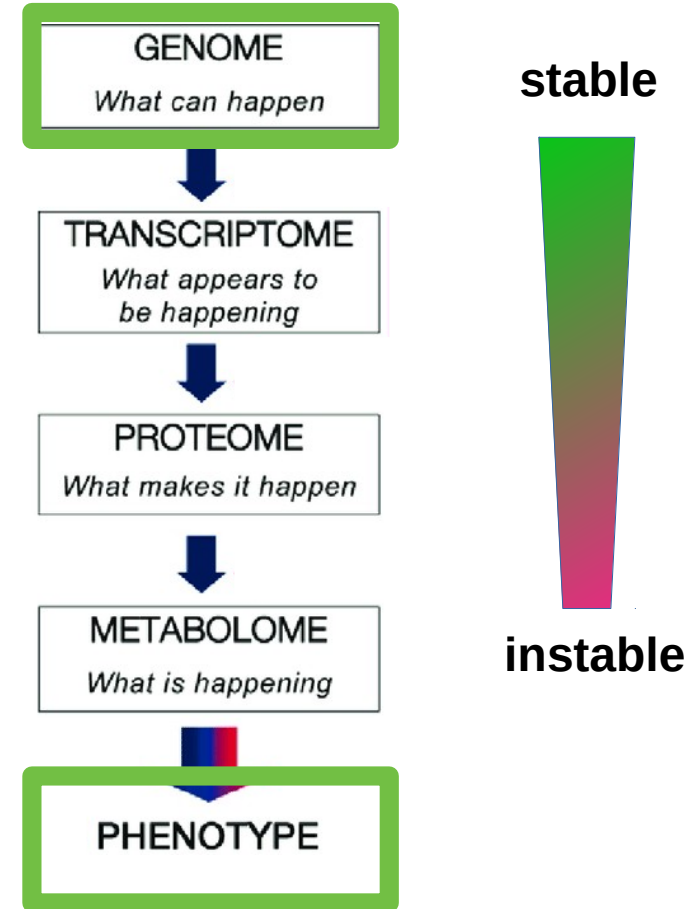
The Omnics cascade

Full genomes >1000 ecotypes
(1001 genomes consortium, 2016)

Complete transcriptomes (28 k genes)
for 80 ecotypes (Clauw et. al. 2016)
& for~ 700 ecotypes (Kawakatsu *et al.* 2016)

> 500 different phenotypes

(Seren *et al.,*2017 & Togninalli *et al.* 2020)

GENOME
What can happen

TRANSCRIPTOME
*What appears to
be happening*

PROTEOME
What makes it happen

METABOLOME
What is happening

PHENOTYPE

stable

instable

The Omnics cascade

Environment

Genotype → Phenotype

Genome-wide association studies (GWAS)

GWAS : Genome - Phenotype

TWAS : Transcriptome - Phenotype

GENOME
What can happen

TRANSCRIPTOME
What appears to be happening

PROTEOME
What makes it happen

METABOLOME
What is happening

PHENOTYPE

stable

instable

TWAS (Transcriptome – Phenotype) for Flowering Time

Using RNA expression to explain the phenotype (Kawakatsu *et al.* 2016)
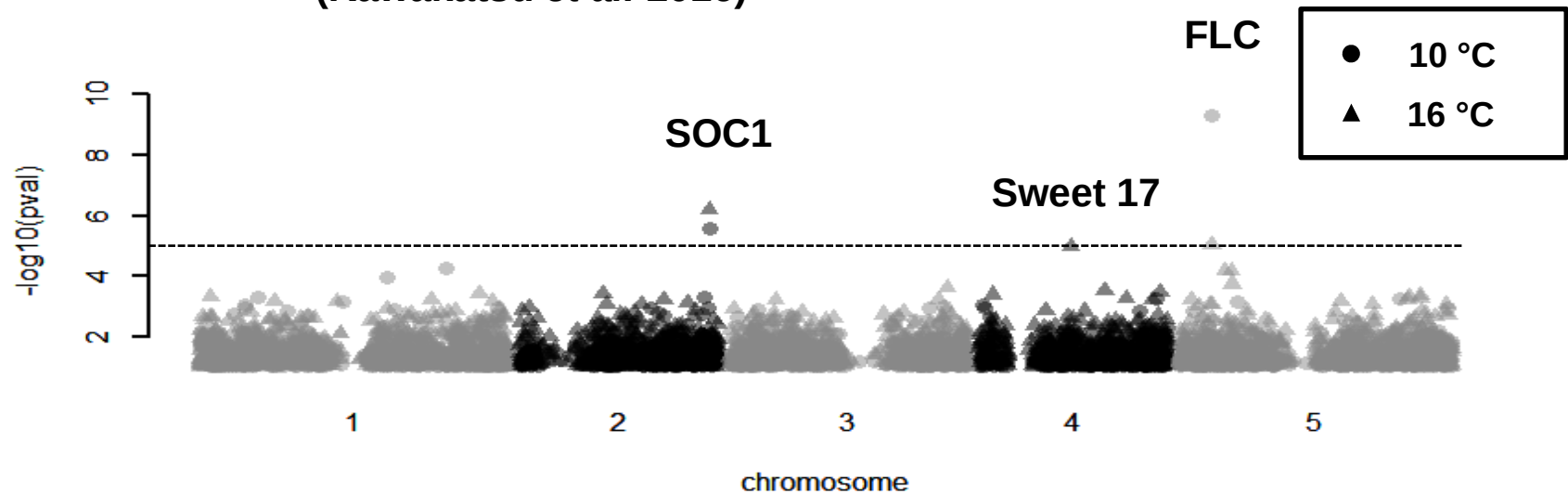
Beyond GWAS: Understanding the function of pathways
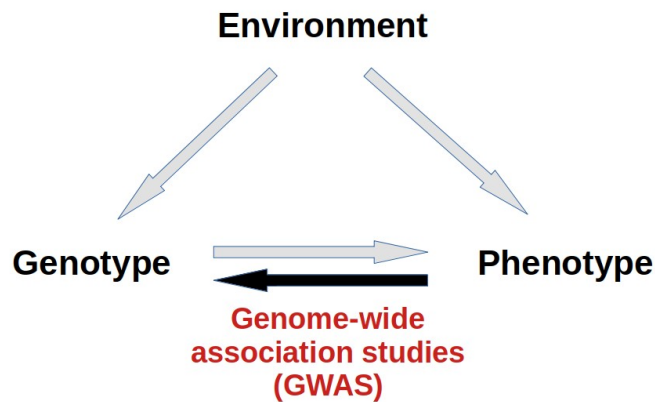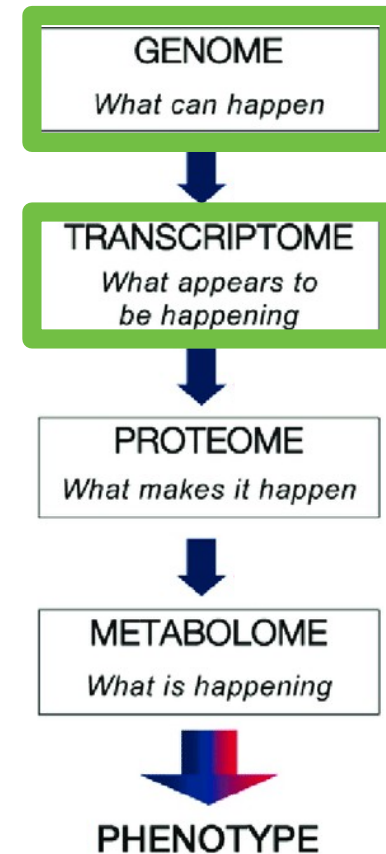
The Omnics cascade

Environment

Genotype → **Phenotype**

Genome-wide association studies (GWAS)

GWAS : Genome - Phenotype

TWAS : Transcriptome - Phenotype

eGWAS: Genome - Transcriptome

GENOME
What can happen

TRANSCRIPTOME
What appears to be happening

PROTEOME
What makes it happen

METABOLOME
What is happening

PHENOTYPE

stable

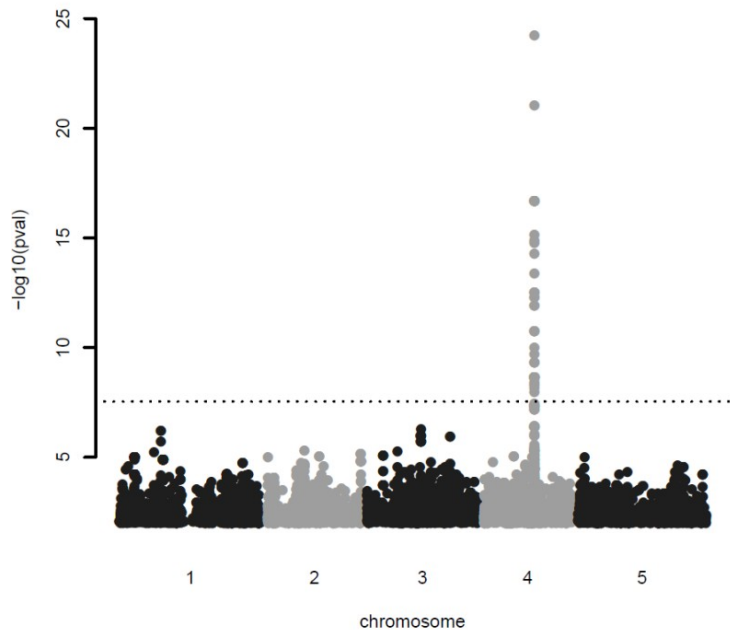**instable**

eGWAS (Genome - Transcriptome)

Data: 80 ecotypes grown under drought and control conditions (VIB Ghent), RNAseq for all of them.
**Use the expression values of each gene as molecular phenotype**

A SNP on Chromosome 4 affects the expression of the gene AT4G19480 How does this affect the phenotype ?

Clauw *et al.* (2016) Plant Cell

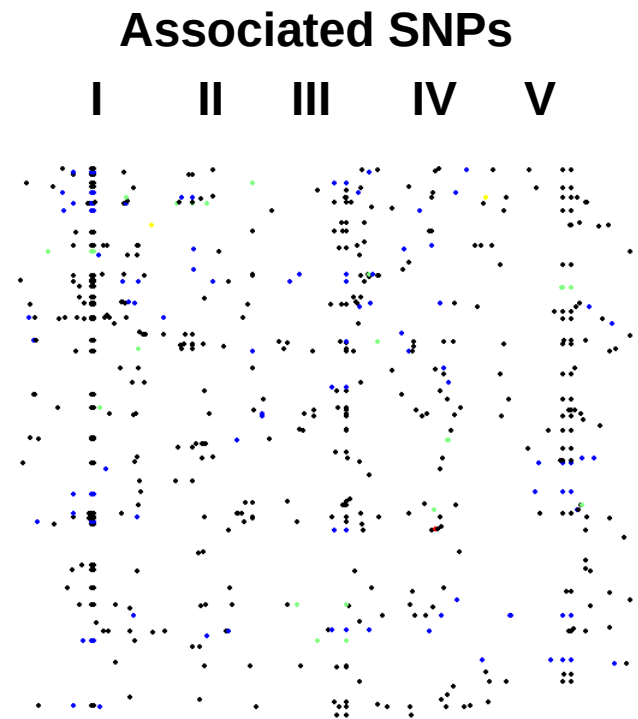# eGWAS, summary plot of 28k analyses

AT4G19480
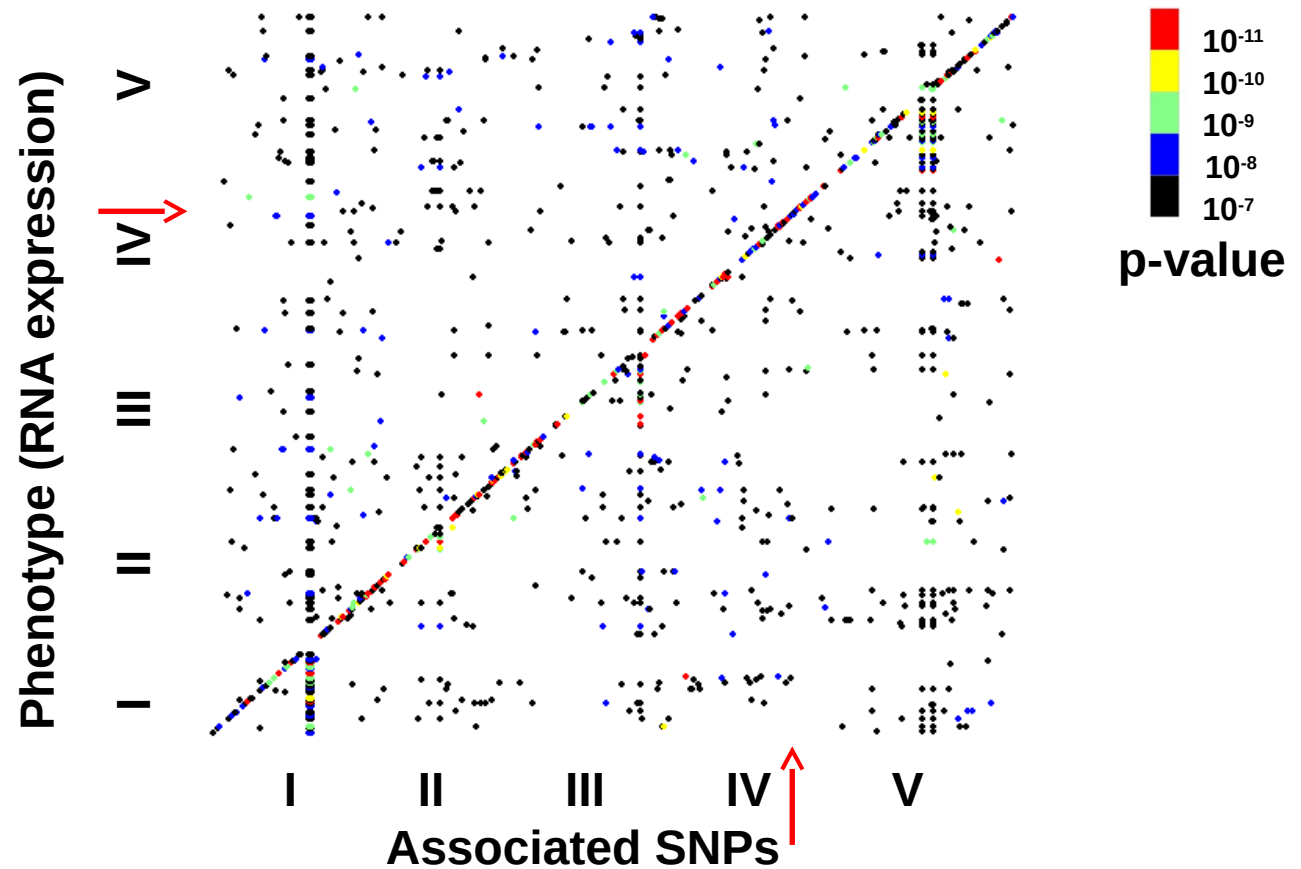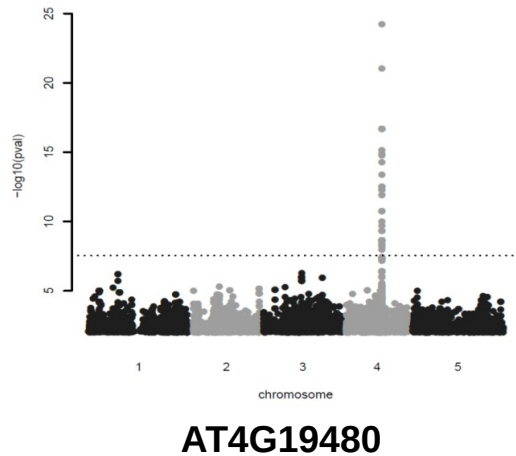
eGWAS with a multivariate model

Joint analyses of RNAseq data from control and drought-stressed samples
Method : Korte *et al.* (2012) NatGen

Associated SNPs

I   II   III   IV   V

Phenotype (RNA expression)

V   IV   III   II   I
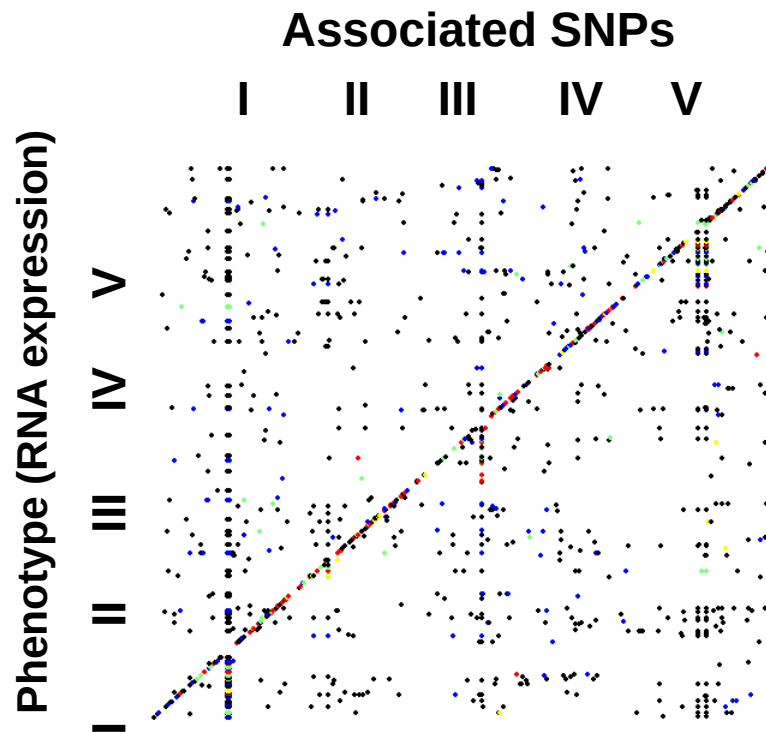
Environment independent

Associated SNPs

I   II   III   IV   V

G x E

Clauw *et al.* (2016) Plant Cell

Global genetic heterogeneitity in adaptive traits

Global genetic heterogeneity in adaptive traits

We measured flowering time in a growth chamber at 10°C for 888 European Arabidopsis ecotypes
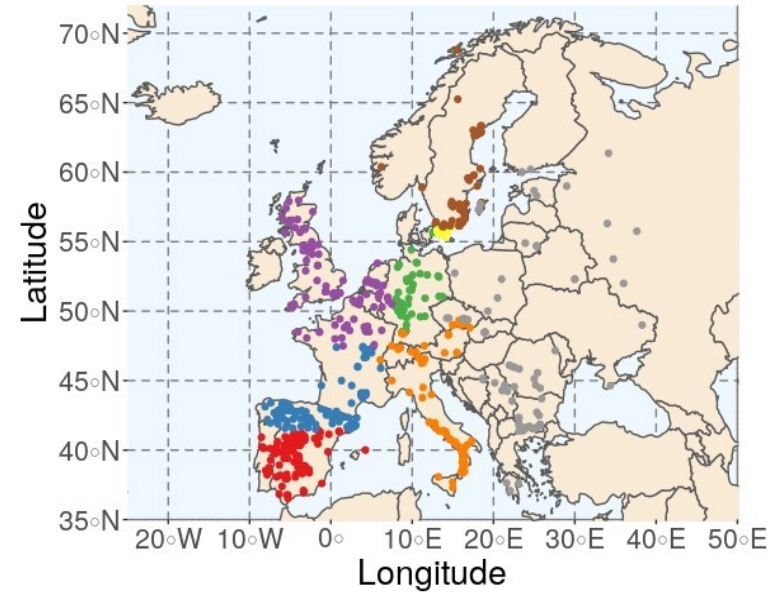
López Arboleda *et al*. BioRxiv

Global genetic heterogeneity in adaptive traits

**We measured flowering time at 10°C
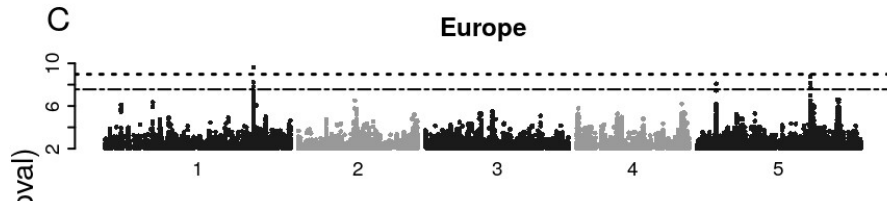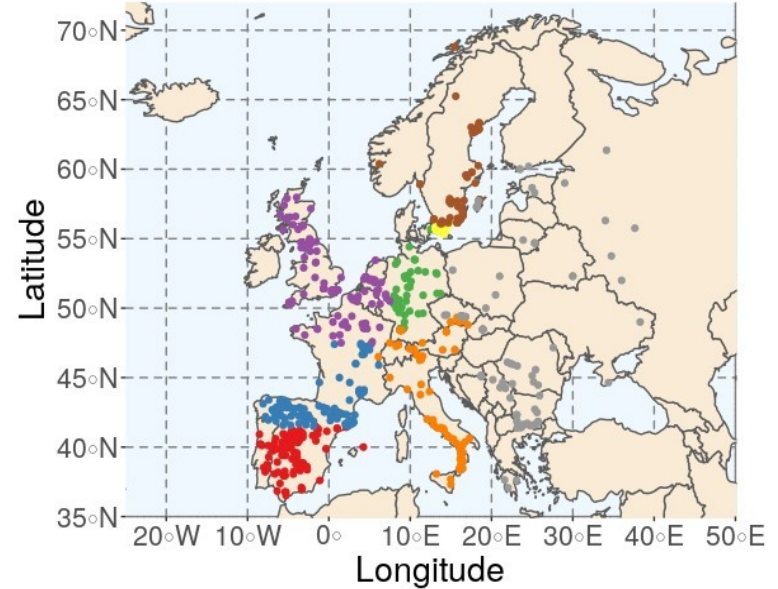In 888 European Arabidopsis ecotypes**

López Arboleda *et al*. BioRxiv

Global genetic heterogeneity in adaptive traits

**We measured flowering time at 10°C
In 888 European Arabidopsis ecotypes**

López Arboleda *et al*. BioRxiv
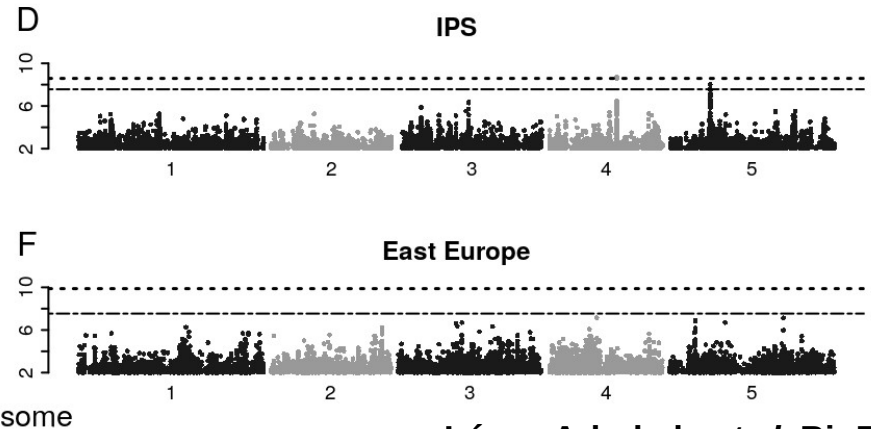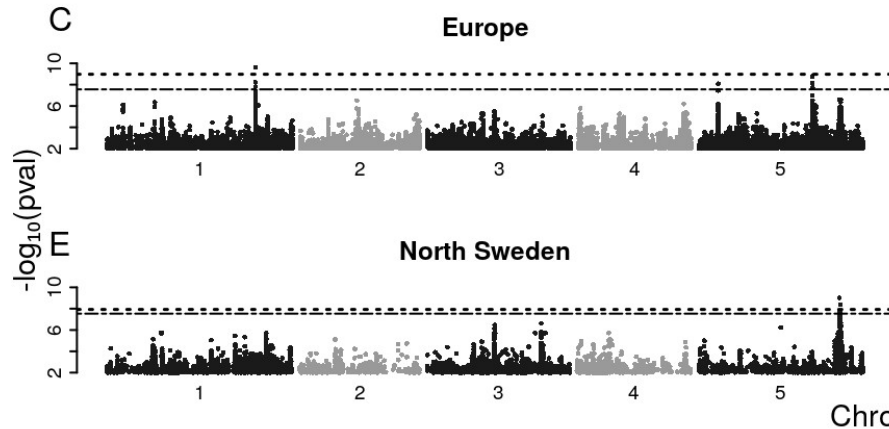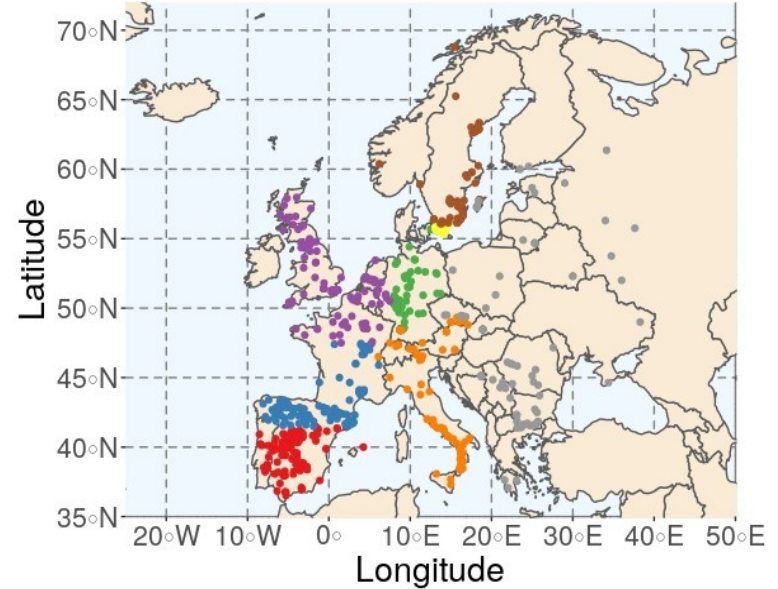
Genotype – Phenotype relationship

Is Flowering time optimized by different genes in different local subsets ?

Environment

Adaptation

Natural selection

Genotype

Phenotype

https://www.hhmi.org
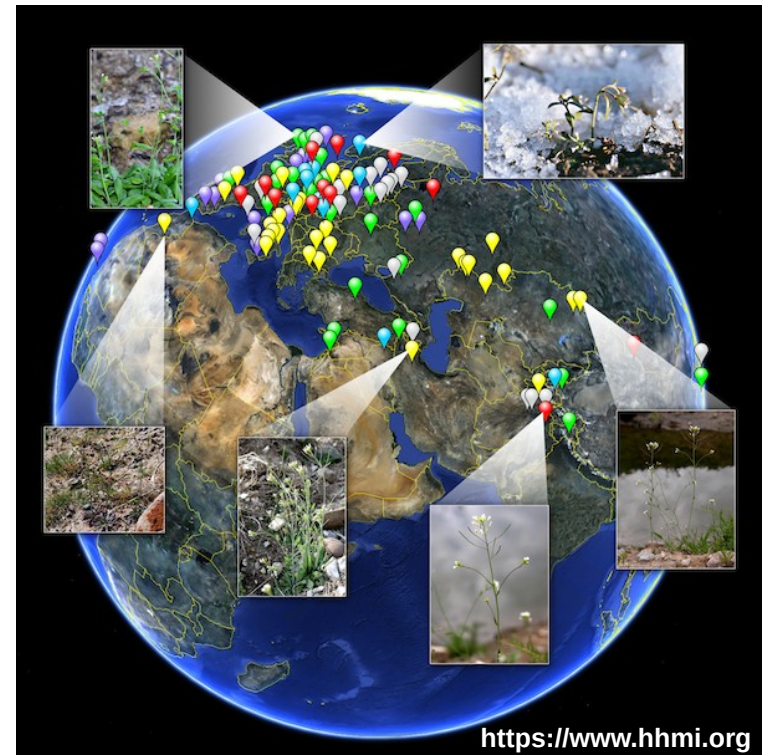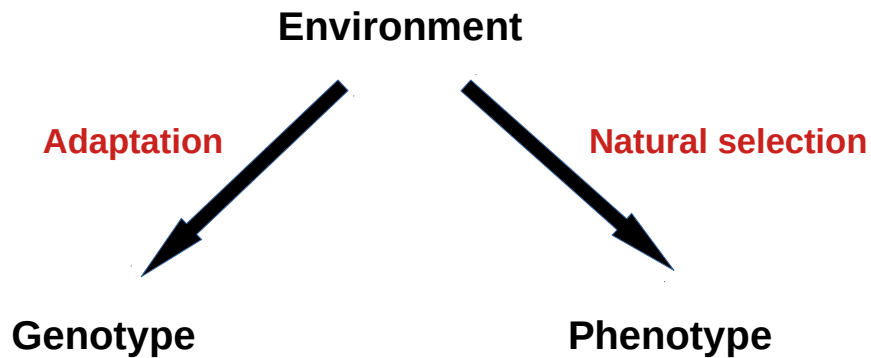
Different genetic architecture of flowering time in different subsets

Overlap of SNPs that are associated with a p-value <10^{-4} in the respective subset
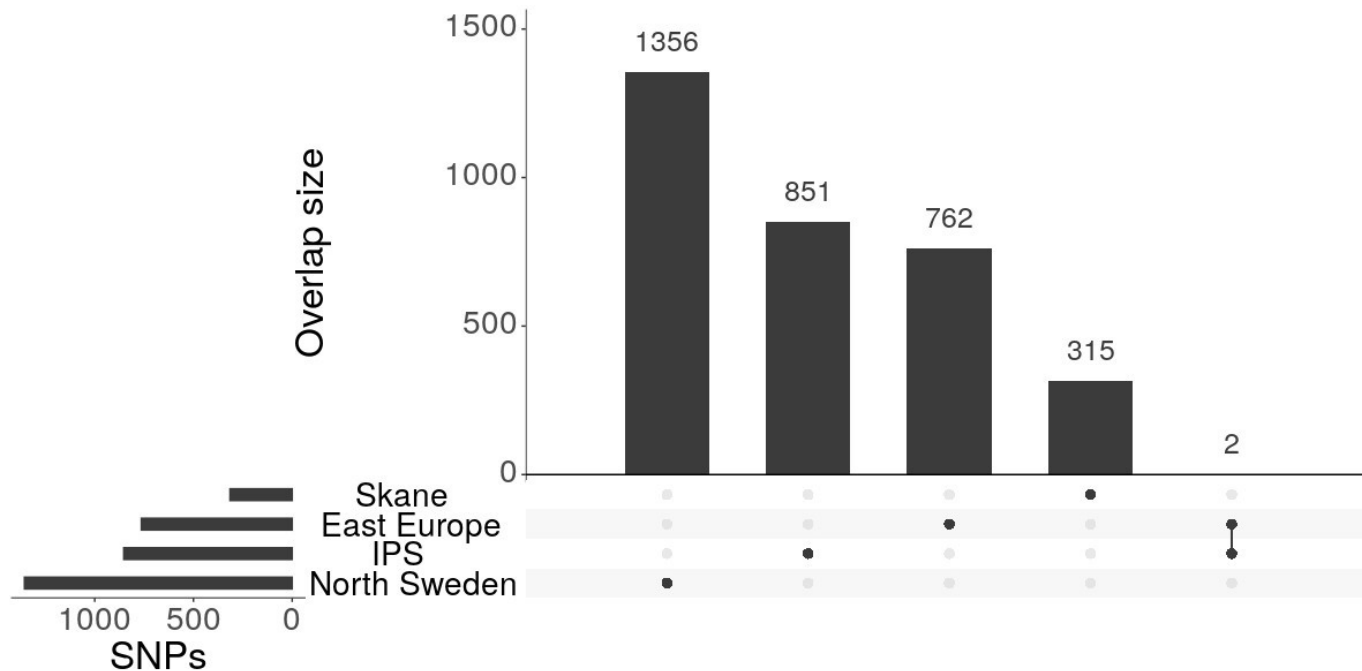
Different genetic architecture in many adaptive traits
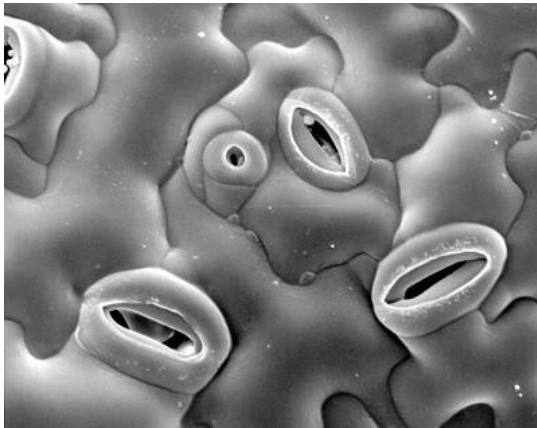
Stomata size was measured in a set of 330 European Arabidopsis ecotypes Dittberner *et al.* 2018

A

B

C Stomata size IP

E Stomata size SW

López Arboleda *et al*. bioRxiv
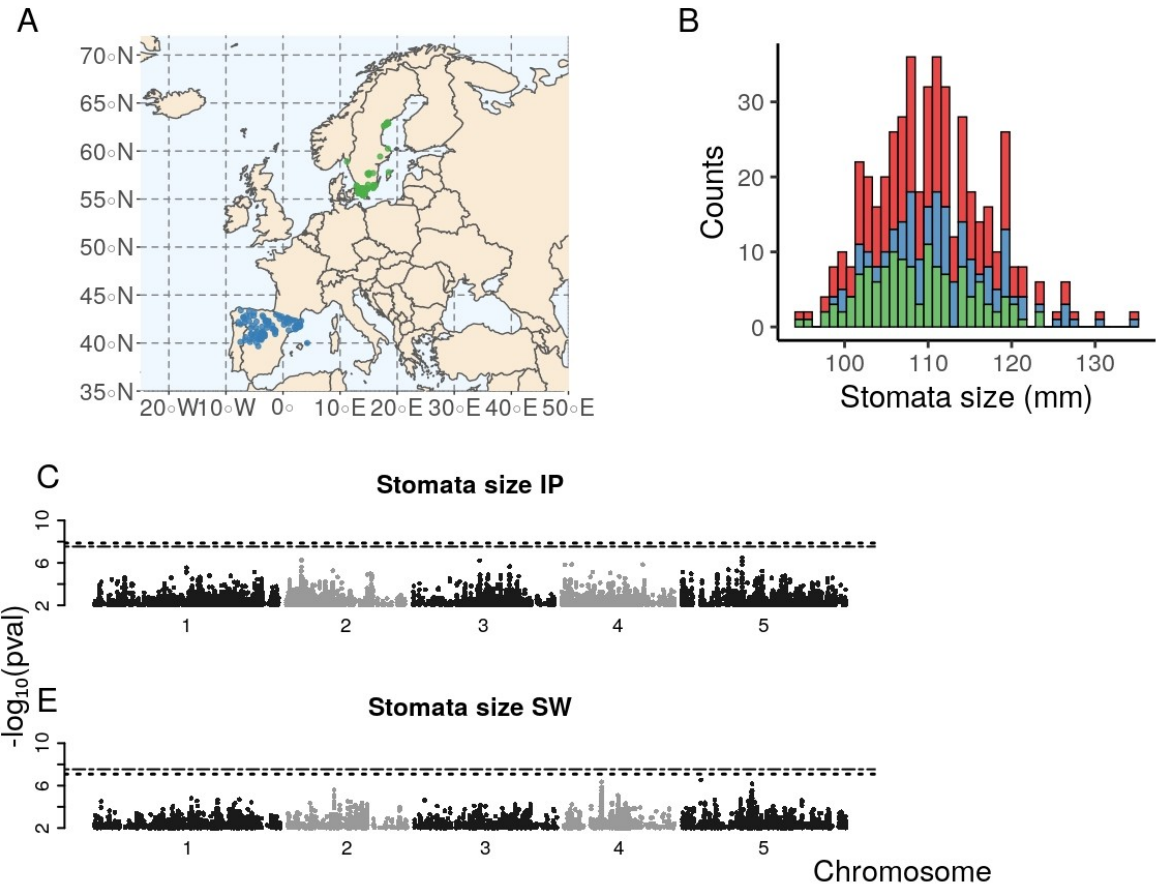
eGWAS in different local populations
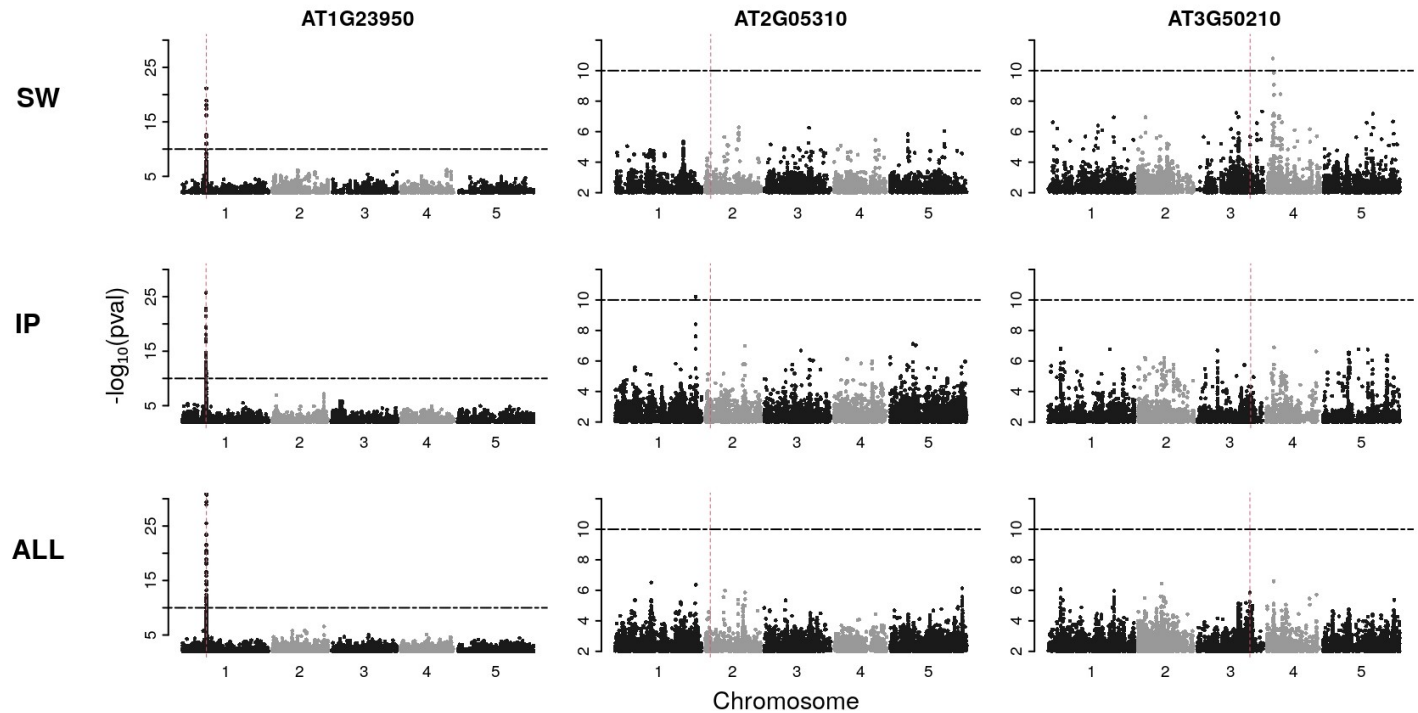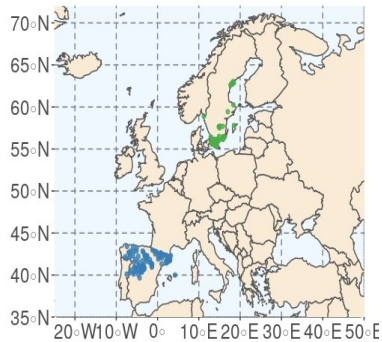
110 genes   **118 genes**   **64 genes**

eGWAS in different local populations

López Arboleda *et al*. bioRxiv
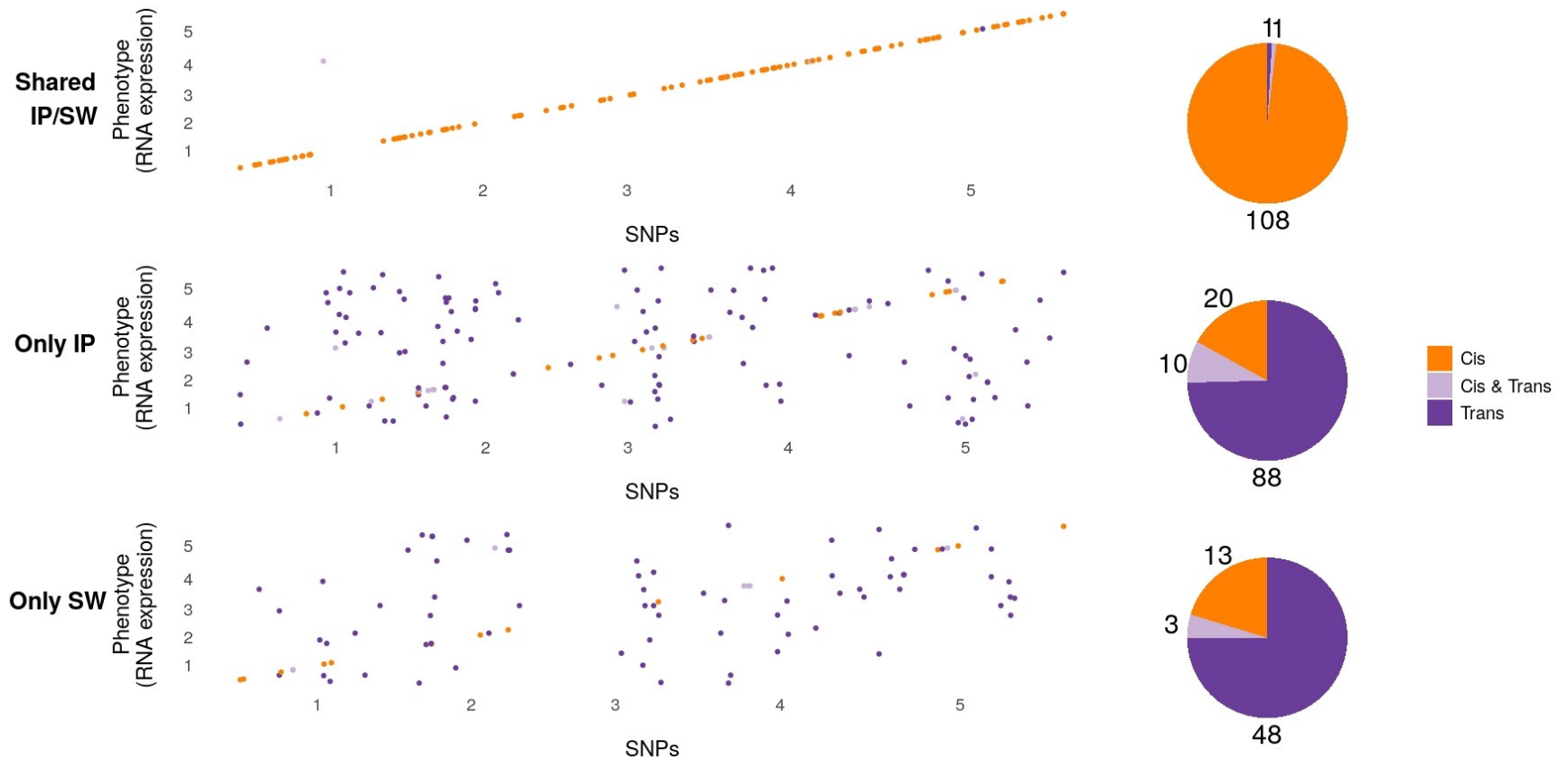
eGWAS in different non-local populations

López Arboleda *et al*. bioRxiv
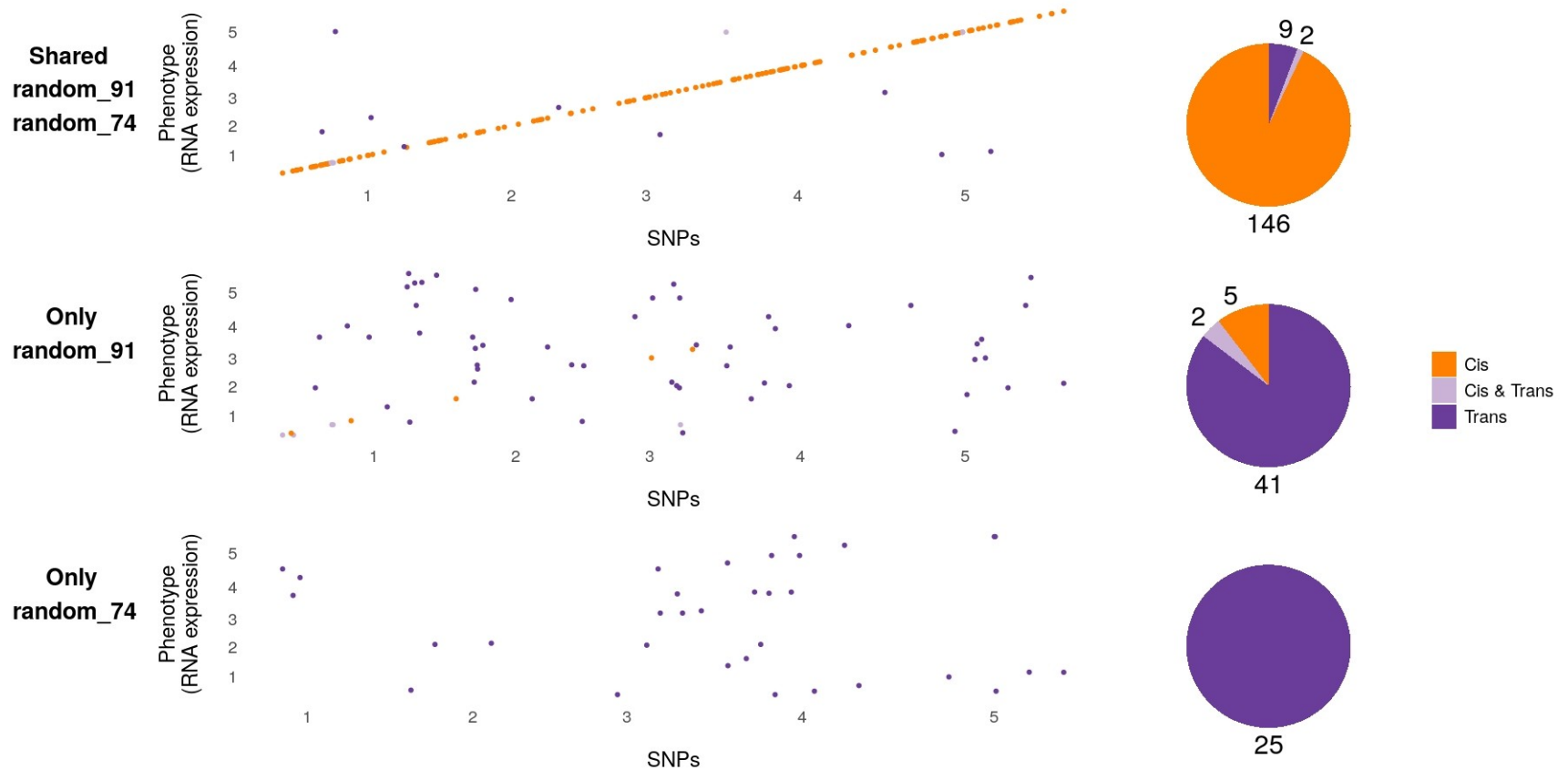
eGWAS in different local populations
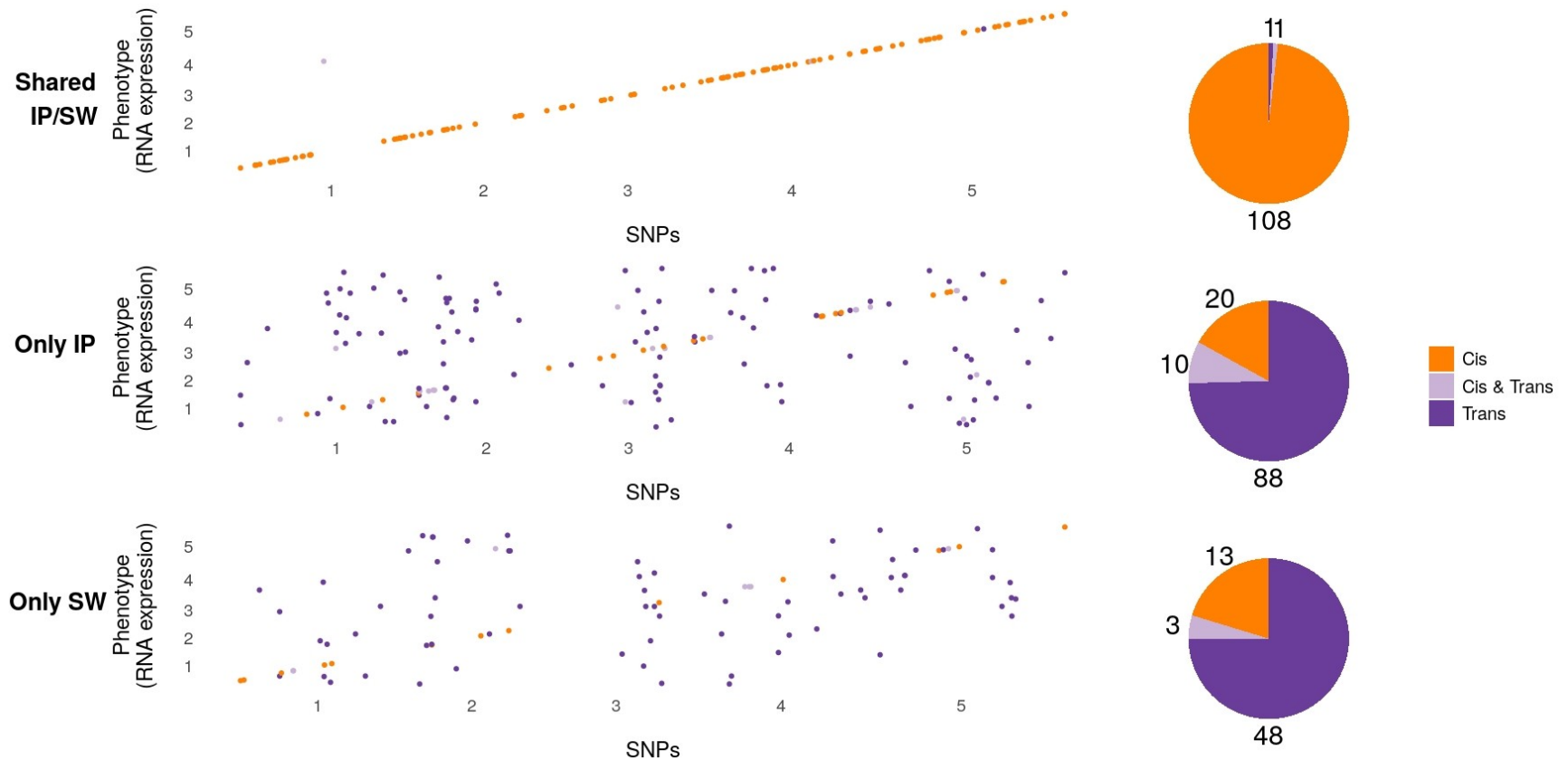
Summary I

What you conclude from GWAS will depend on where you sample
True for humans as well (e.g. genetic architecture for skin color differ dramatically between Iceland and Cape Verde Islands)

Summary II

Can we infer how local adaptation works ?

Putative local adaptive regulatory alleles are mostly regulated in *trans*

Genomic variation in the A.*thaliana* population (1135 accessions)

10,709,466 SNPs segregate in the population

1,854,599 SNPs are located in coding regions

28,148 SNPs lead to a premature STOP codon

Nearly 10,000 genes are knock-out in at least 1 accession

Co-occurence of premature Stop codons

Gene 1	Gene 2	gene1_count	gene2_count	Co - occurrence	Lower border	Upper border	P-value
AT1G66920	AT3G59750	341	251	122	100	157	0.03
AT4G36140	AT5G45150	312	356	266	138	196	1.7e-57
AT3G05685	AT3G63320	444	114	38	55	96	5.6e-16

Many genes are knocked-out together more (or less) often than expected

Co-occurence of premature Stop codons

	prematSTOP
Knocked-out together more often (over)	**2917**
Knocked-out together less frequently (under)	**138**

Co-occurence of premature Stop codons

	prematSTOP	SYN	NON_SYN
Knocked-out together more often (over)	2917	1216	1086
Knocked-out together less frequently (under)	138	46	59

More connections for functional variants compared to (potential) neutral variants

GWAS on the Co-occurence

Create a binary trait (0|1) which indicates if in a given accessions both genes contain functional premature STOP codons or if only one gene is knocked out

Y_AT2G15930_AT5G63630_NA
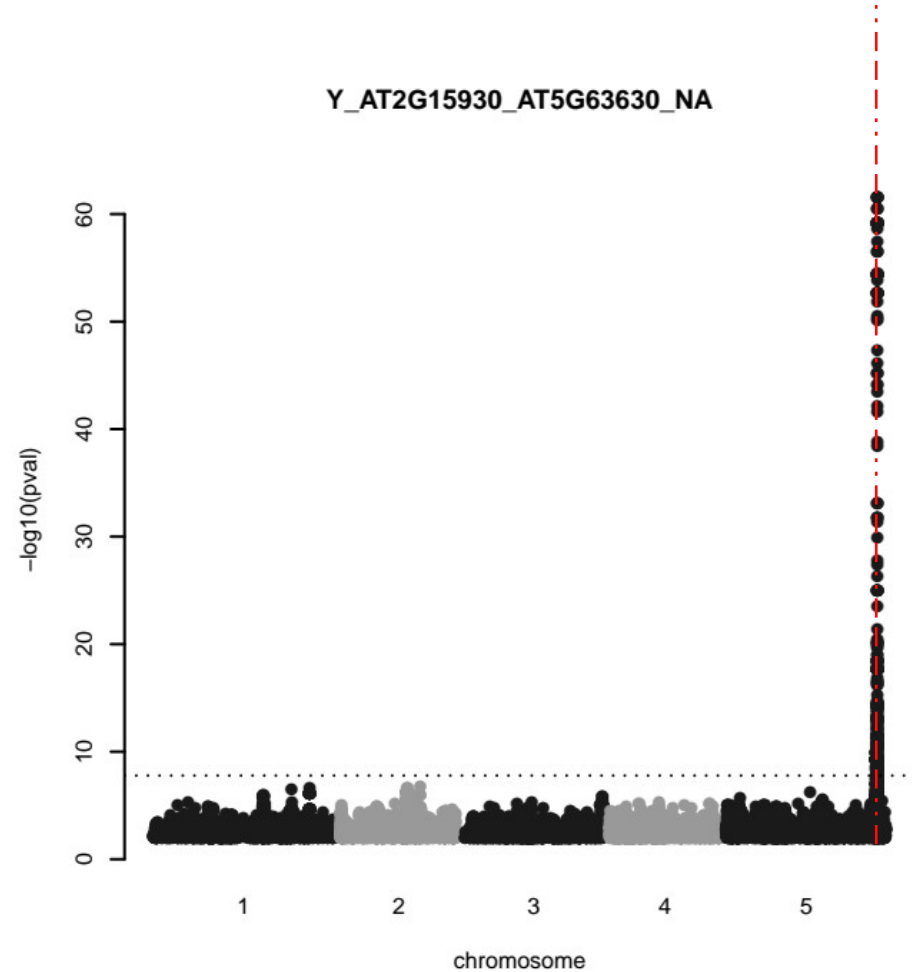
GWAS on the Co-occurence

Create a binary trait (0|1) which indicates if in a given accessions both genes contain functional premature STOP codons or if only one gene is knocked out

Signal in cis = expected (5569/5834 for over and 209/276 for under)

Y_AT2G15930_AT5G63630_NA

Anwar *et al*. In prep.

GWAS on the Co-occurence

Create a binary trait (0|1) which indicates if in a given accessions both genes contain functional premature STOP codons or if only one gene is knocked out
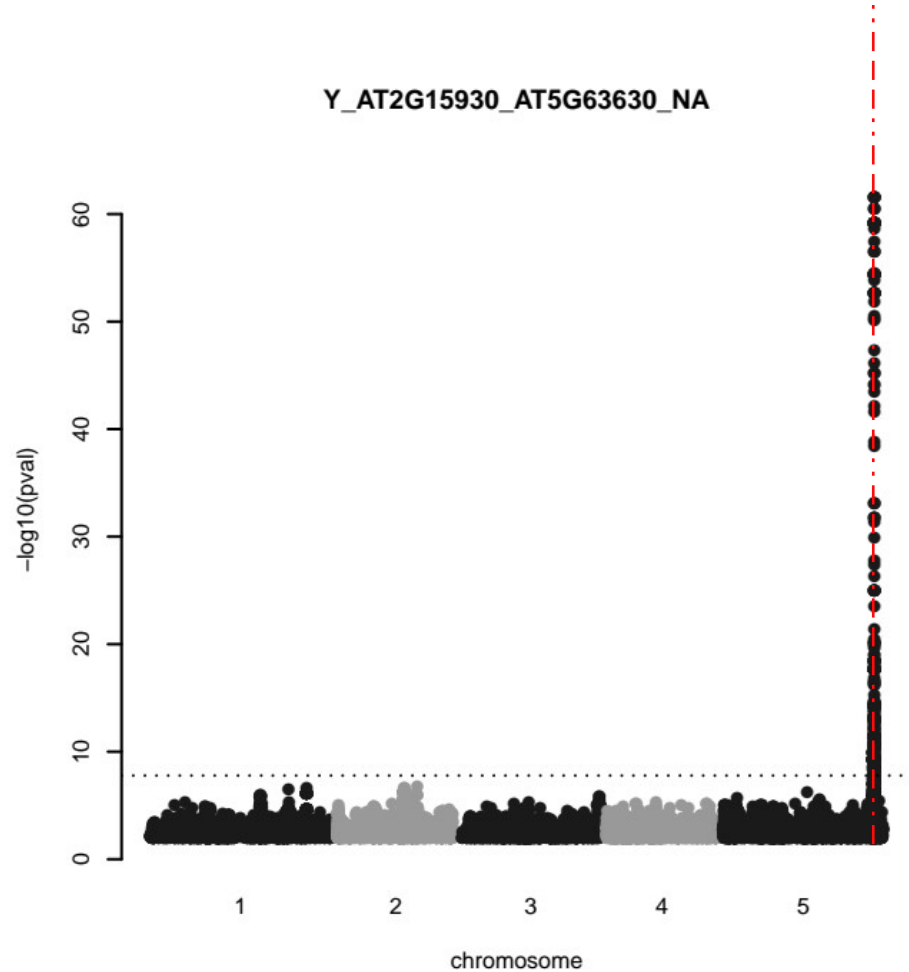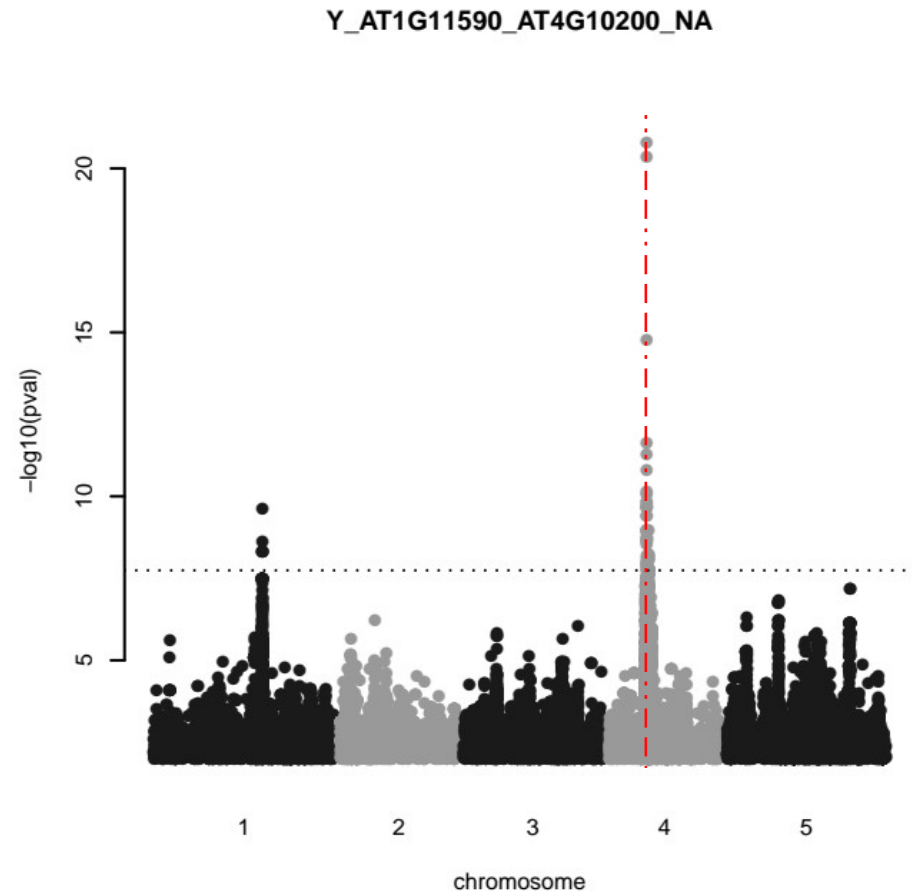

Y_AT1G11590_AT4G10200_NA
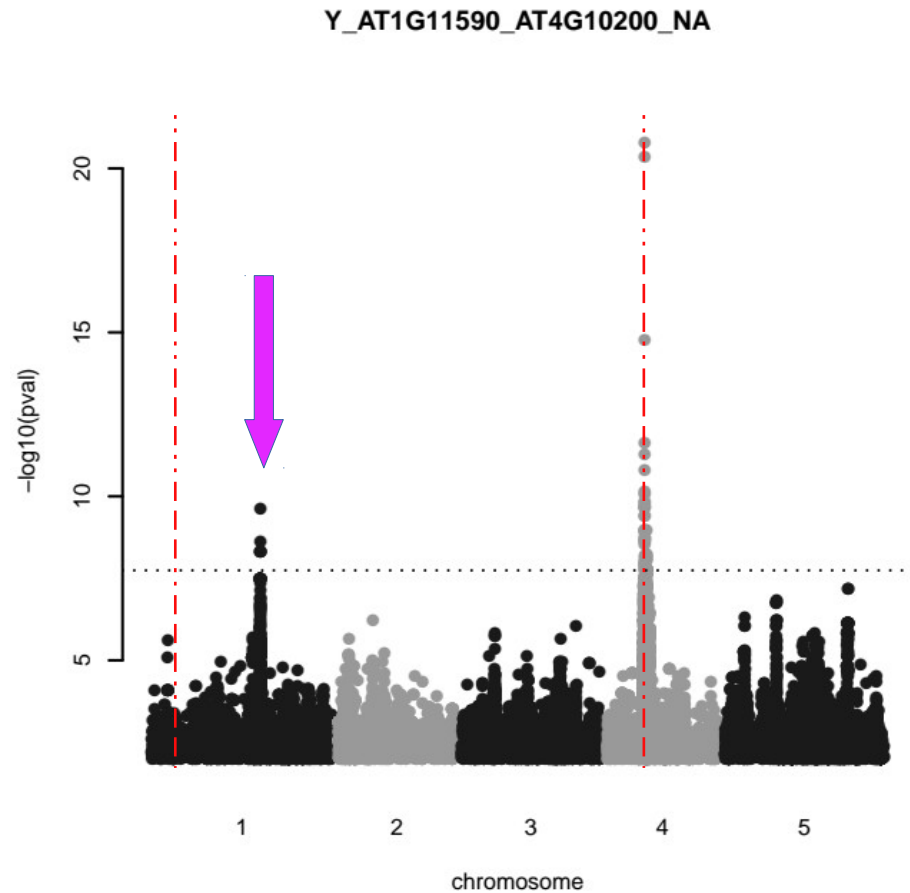
GWAS on the Co-occurence

Create a binary trait (0|1) which indicates if in a given accessions both genes contain functional premature STOP codons or if only one gene is knocked out

Additional signal in trans (212/5834 for over and 8/276 for under)

Y_AT1G11590_AT4G10200_NA

We can map genes that are responsible for the co-occurence

Anwar *et al*. In prep.

Network of genes that are knocked-out together
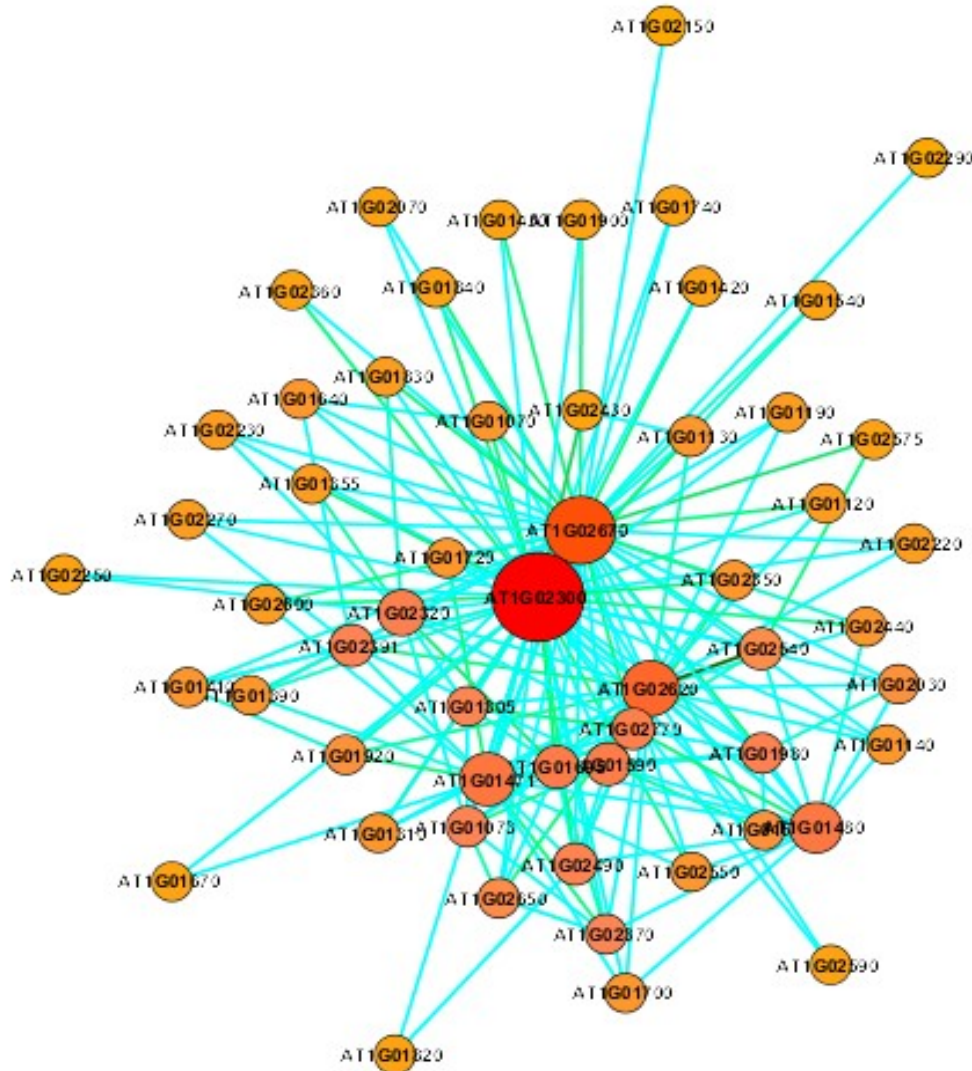
In collaboration with Pascal Braun (INET, HMGU)

Summary III

Tremendous amount of Natural Variation in signaling pathways

If we want to understand G-P, we need to understand the underlying gene regulatory networks

JMU Würzburg
Jan Freudenthal
Ammarah Anwar
William López Arboleda
Stephan Reinert
Laura Steinmann