Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

CTB Center for Computational and Theoretical Biology

Evolutionary Genomics
-
And why ML might have potential to analyze genomic data

@arthurkorte

arthurkorte

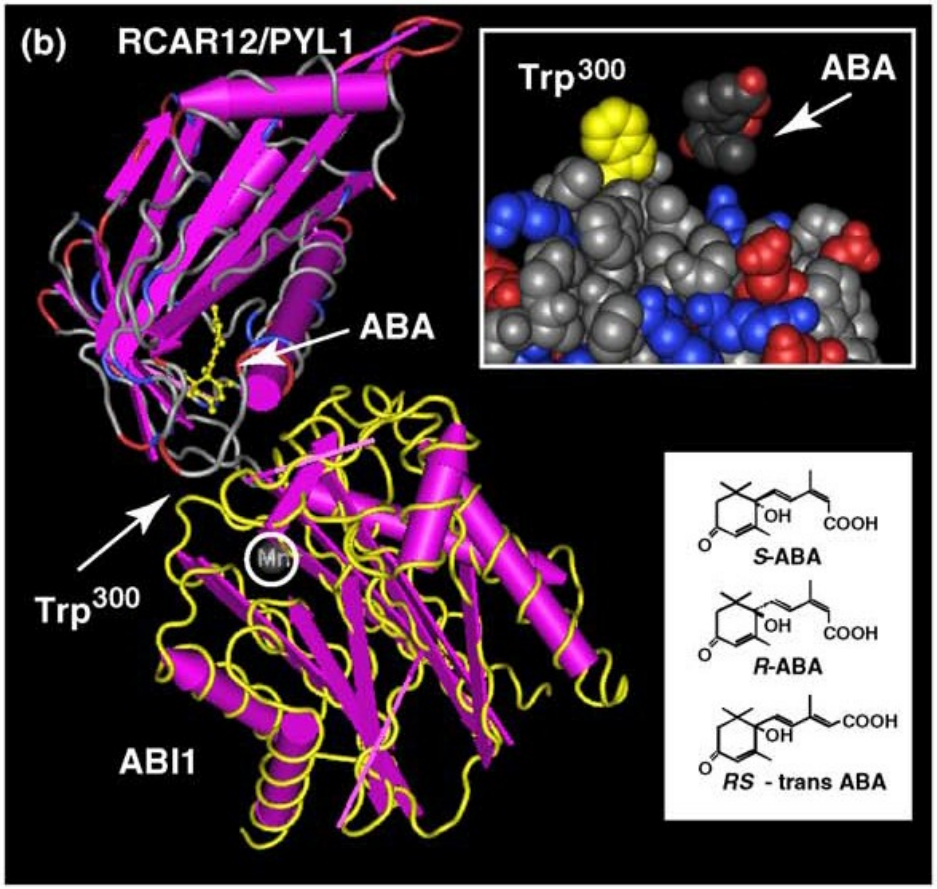
Arthur Korte

Sep 24th 2019

Andrew Kern, U Oregon

“We are still in the Wild West times of Machine Learning for Genomics”

PhD in Molecular Biology

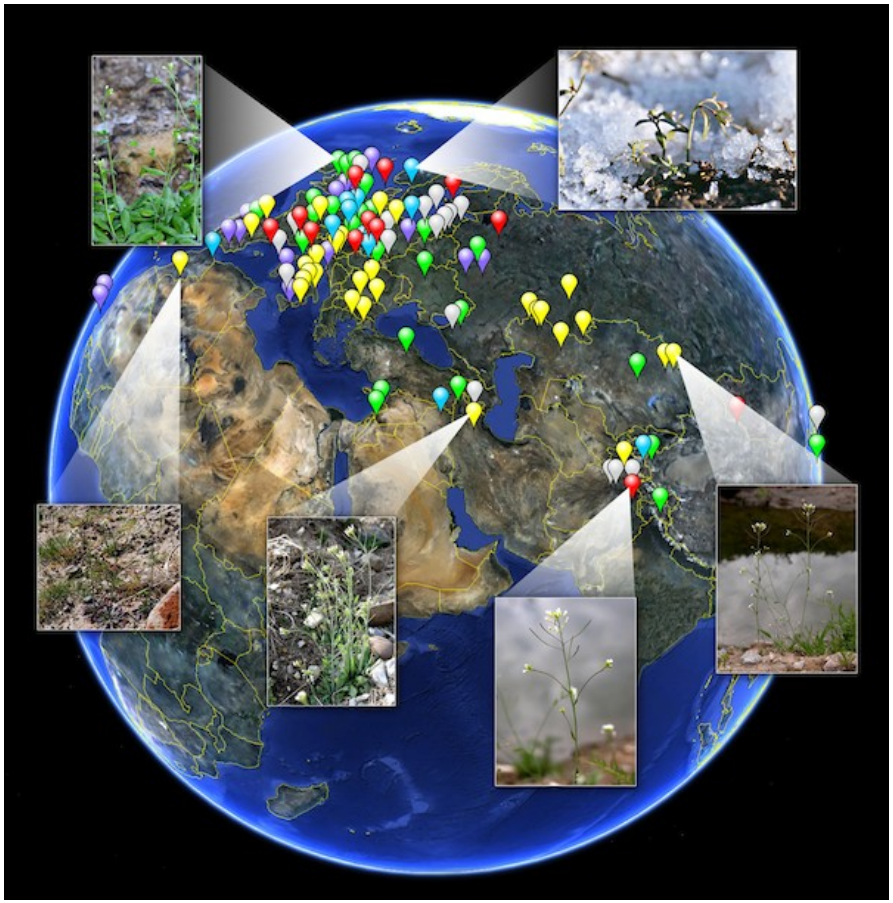


How do plants react to stress ?

Raghavendra *et al.* 2010, TiPS

Postdoc in Population Genetics

By Emmanuel Boutet - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1437488

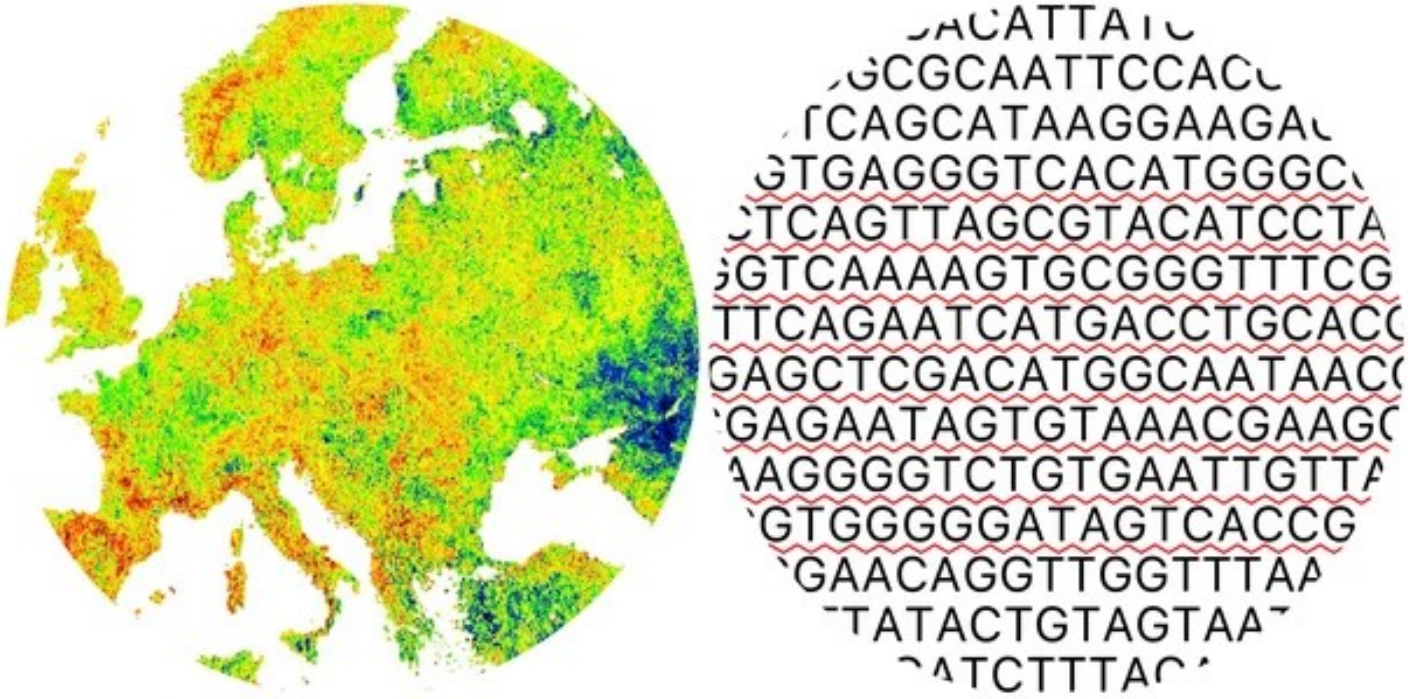


Credit: Photographs/images were contributed by Patrick Gooden, Kathleen Donohue and Google Earth. Graphic design: Jamie Simon, Salk Institute.

How do populations adapt to the environment ?

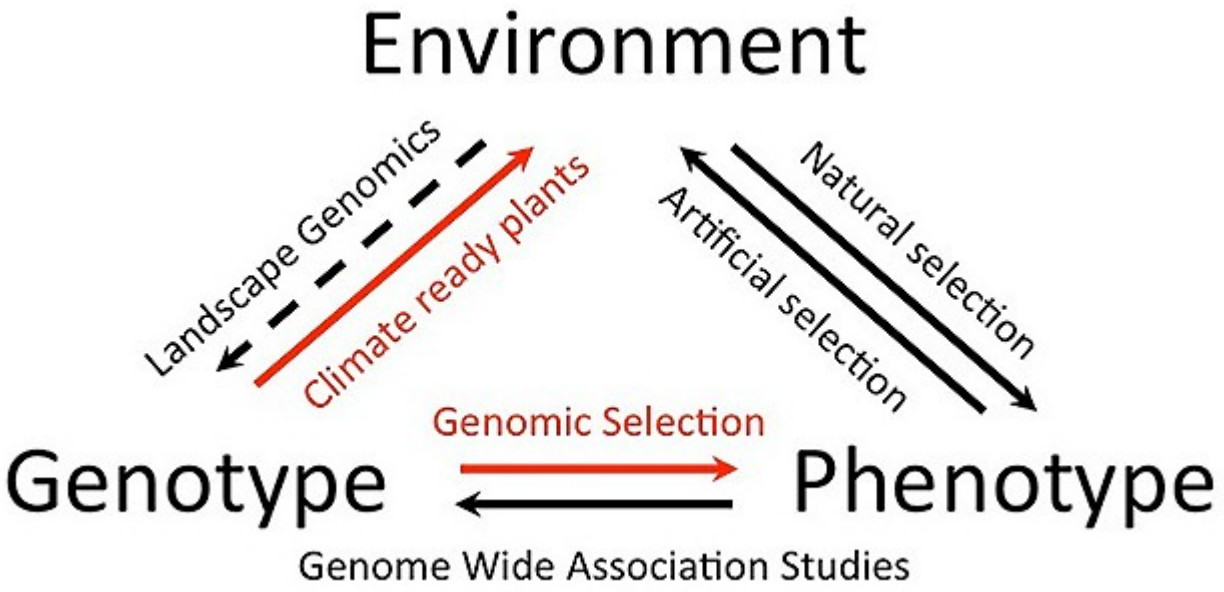
Quantitative Genetics
&
Computational Biology

How do plants adapt to drought

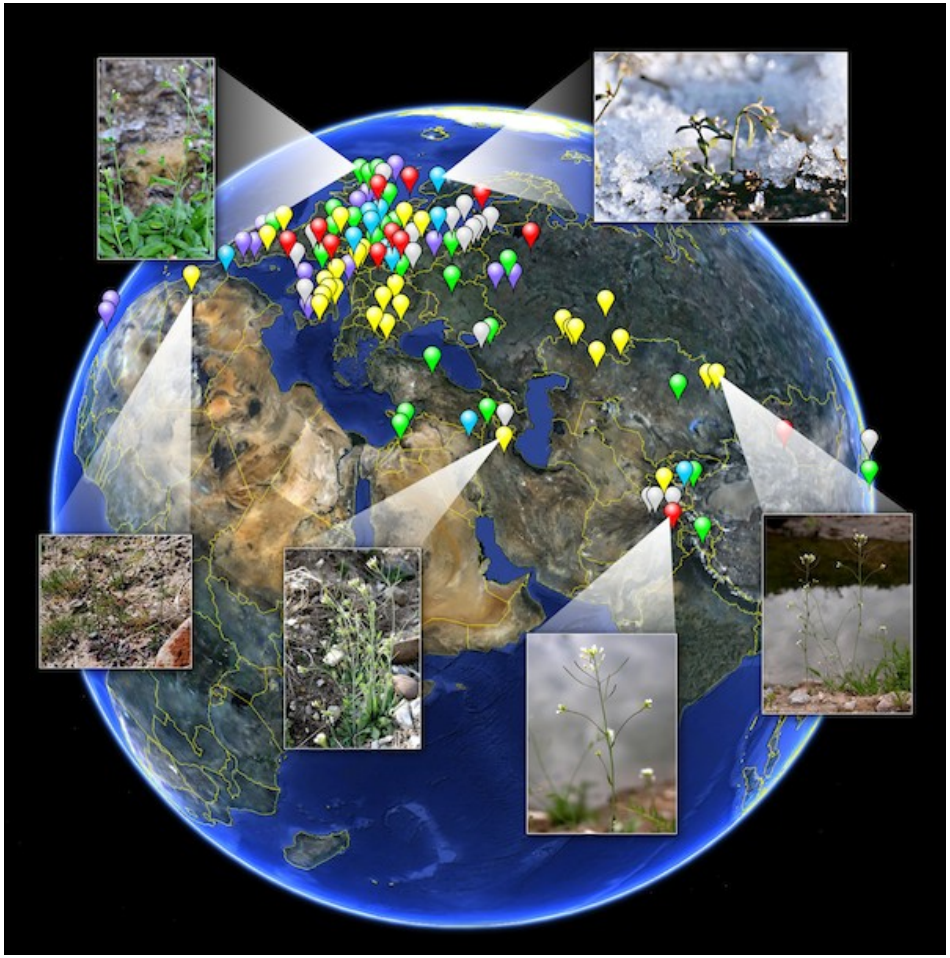


Monroe *et al.* (2019) eLife

Evolutionary Genomics



Arabidopsis thaliana



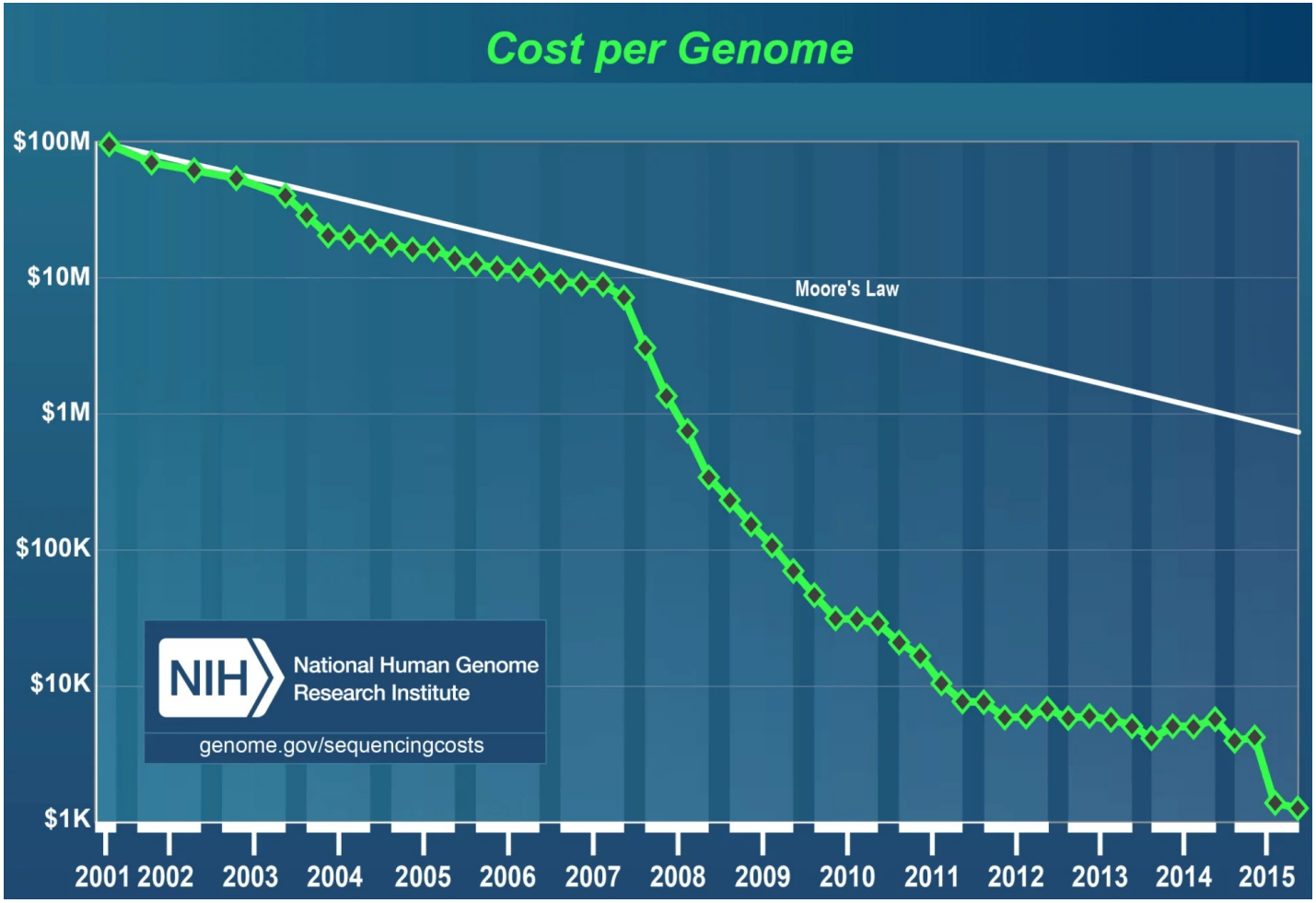
Genome-wide association studies (GWAS)

Associations between phenotype and genotype

Genome sequencing

1. DNA isolation
2. DNA sequencing
3. Genome assembly

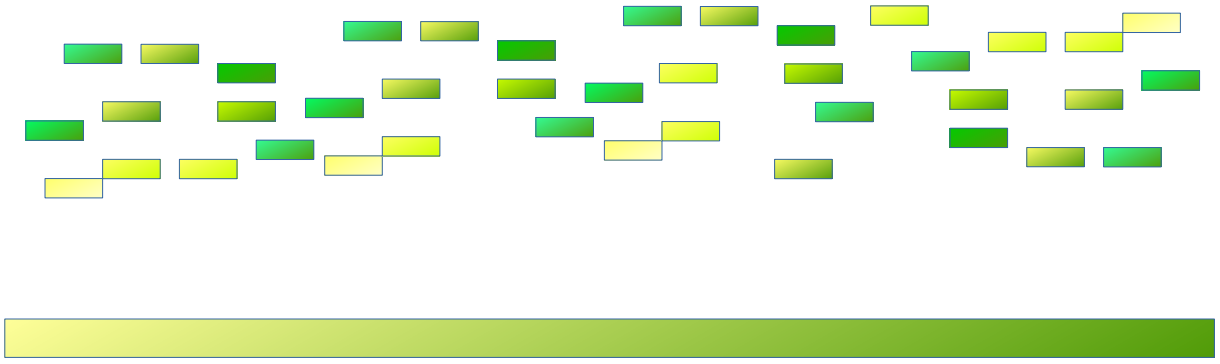
Sequencing costs



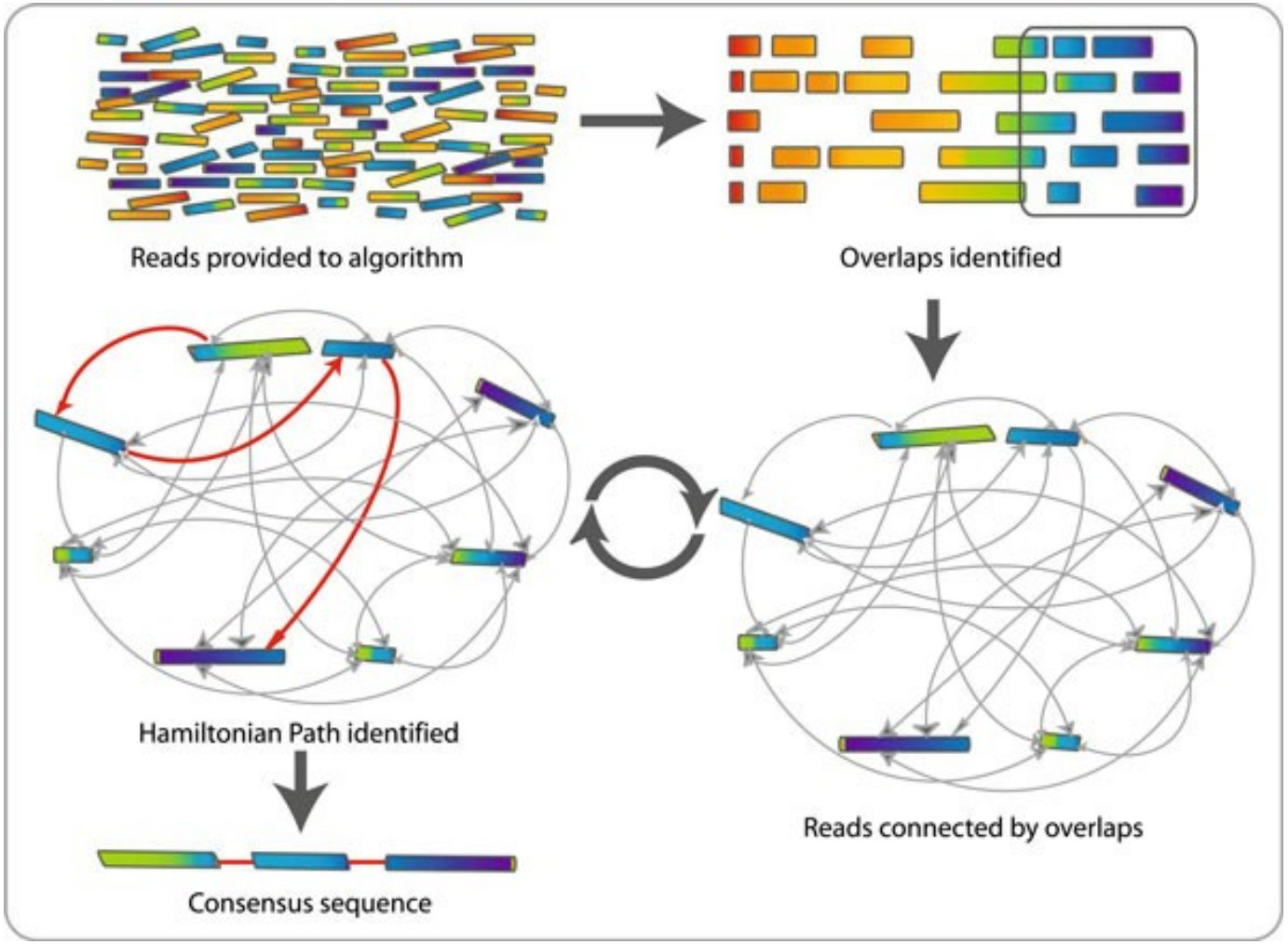
Genome assembly : A giant jigsaw puzzle

We want a continuos strand of DNA (*aka* whole chromosomes (e.g. Human 200 Mbp, Arabidopsis 20 Mbp)

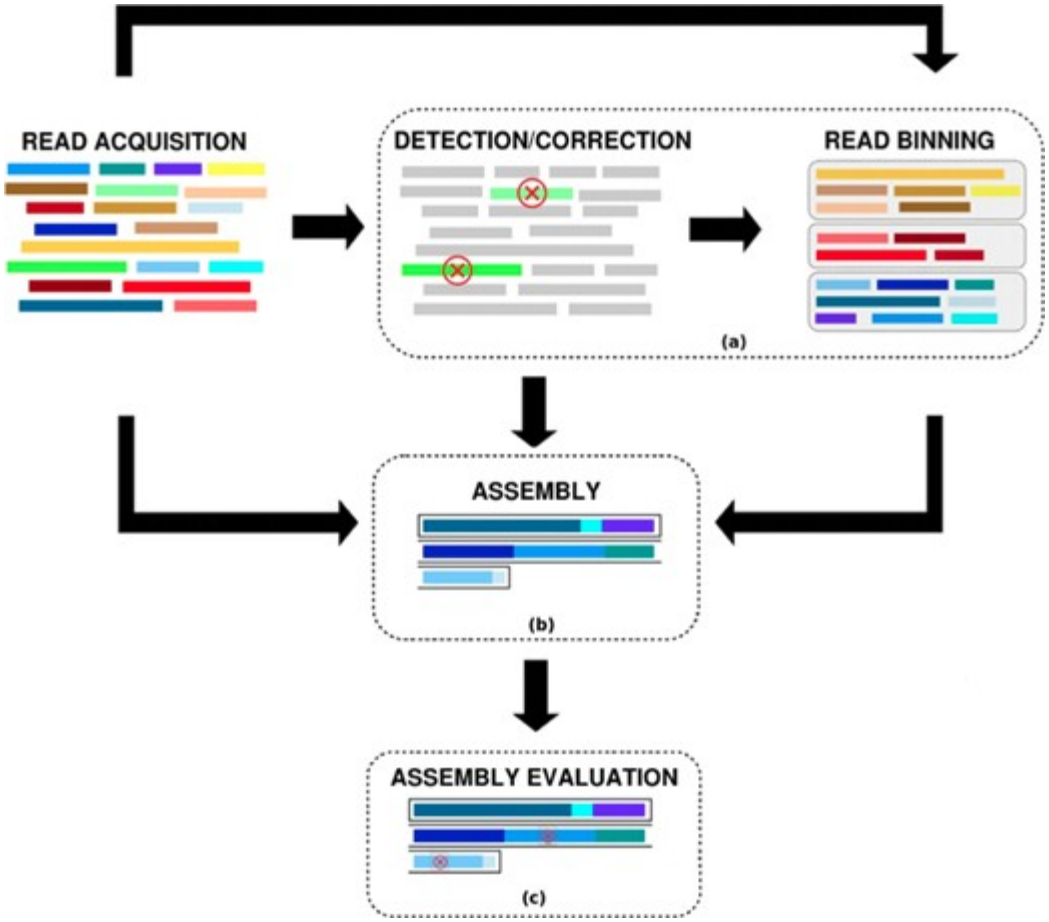
Sequencing generates 100 M reads with a length of 100 bp.



Genome assembly : A giant jigsaw puzzle



Genome assembly : A giant jigsaw puzzle



De Souza *et al.* (2018) Briefings in Bioinformatics

Step	Identification	Problem addressed	Main ML model	Data set	Simulated	Model features
				Real		
Pre-assembly	Angeleri (1999)	Binning of reads	Artificial neural networks	✓	✗	Nucleotide sequences
	Constantinescu (2015)	Binning of reads	Artificial neural networks		✗	Nucleotide sequences
	Krachunov (2017)	Detection of sequencing errors	Artificial neural networks	✗		K-mer frequencies
			Random forest			
			Hoeffding Trees			
			RIPPER			
Post-	Choi (2008)	Assembly evaluation	Decision tree	✓	✗	Error identification met-
assembly			Random forest			rics (e.g., RCN, RCX, CE,
			Random trees			etc.)
			Bayesian networks			
			Naïve bayes			
	Lanc (2013)	Assembly evaluation	K-Means	✓	✗	Adjusted AMOSvalidate features
	Bodily (2014)	Contig evaluation for	Binary classifiers		✗	Abundance information
		scaffolding step				(e.g. contig size and read coverage)
	Kuhring (2015)	Assembly evaluation	Random forest	✓		Contig size, read cover-age, read length, read quality, etc.
Auto-	Palmer (2010)	Read overlap evalu-	Random tree	✓		K-mer frequencies
assembly		ation	Random forest			
			Naïve Bayes			
	Bocicor (2011)	Read order definition	RL		✓	Overlapping score
	Zhu (2014)	Contig extension evaluation	SVMs	✓		Branch features (coverage level, path weight, etc.)
	Wang (2015)	Read assembly into contigs	Hidden Markov Mo-dels	✓		Target-gene features
	Afiahayati (2015)	Read assembly into contigs and scaffolds	SVMs	✓	✗	Number, length and coverage of sequences and K-mer frequenc
	Ji (2017)	Read assembly into contigs and scaffolds	Unsupervised learning SVMs	✓	✗	Codon frequency and TNF

Procedia Computer Science
Volume 108, 2017, Pages 1145-1154

Machine learning models in error and variant detection in high-variation high-throughput sequencing datasets

Milko Krachunov [1] ✉, Maria Nisheva [1], Dimitar Vassilev [1]

Show more

https://doi.org/10.1016/j.procs.2017.05.242

4.2 Learning input

The ML-based models would classify an input example for a given evaluated base r_k—preselected by (1) or otherwise—producing a Boolean prediction (correct or incorrect) as its output. Each example in our models would be a vector of frequency statistics about the base r_k. In particular, each number would be its frequency in a different subset of the reads.

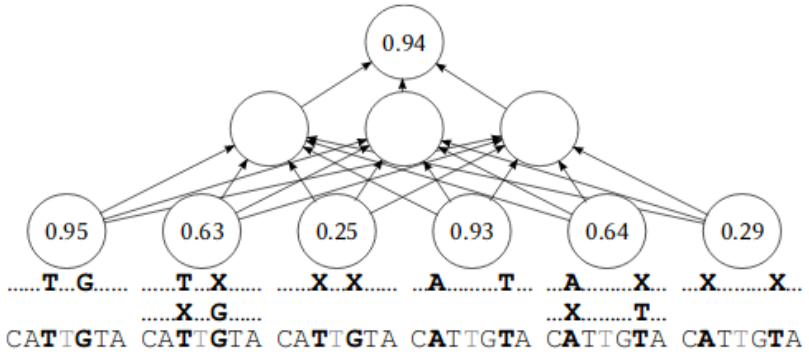
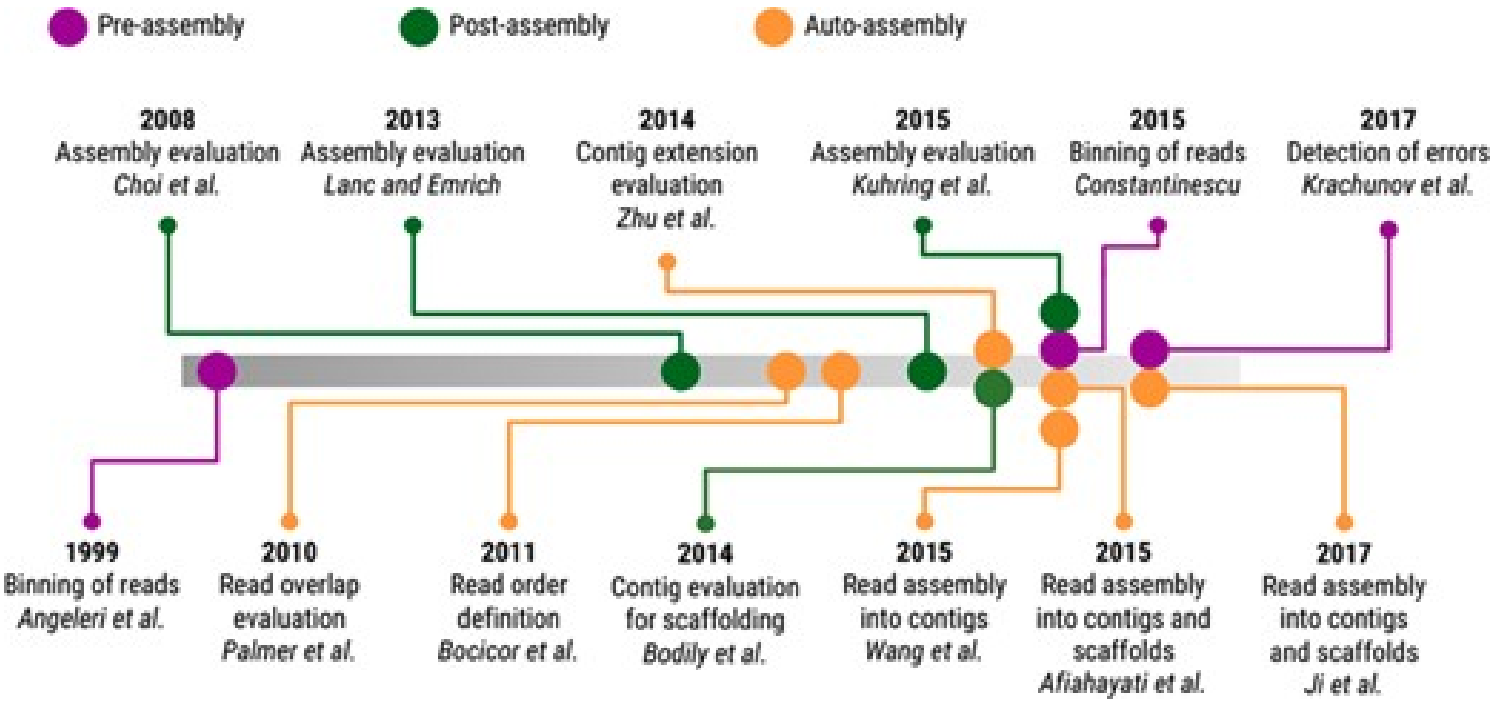


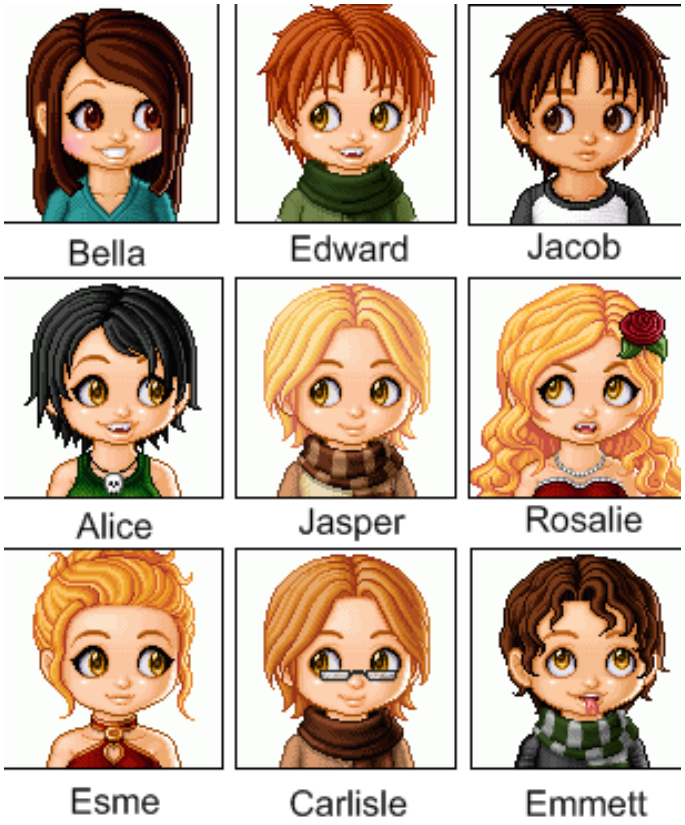
Figure 1: Example learning input

ML in genome assembly



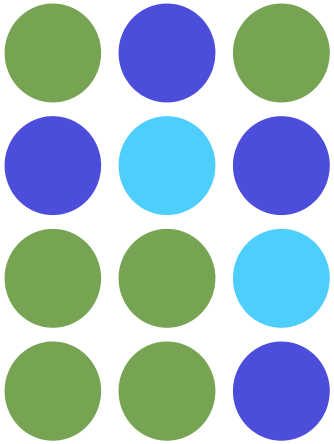
De Souza *et al.* (2018) Briefings in Bioinformatics

Genome-wide association studies (GWAS)



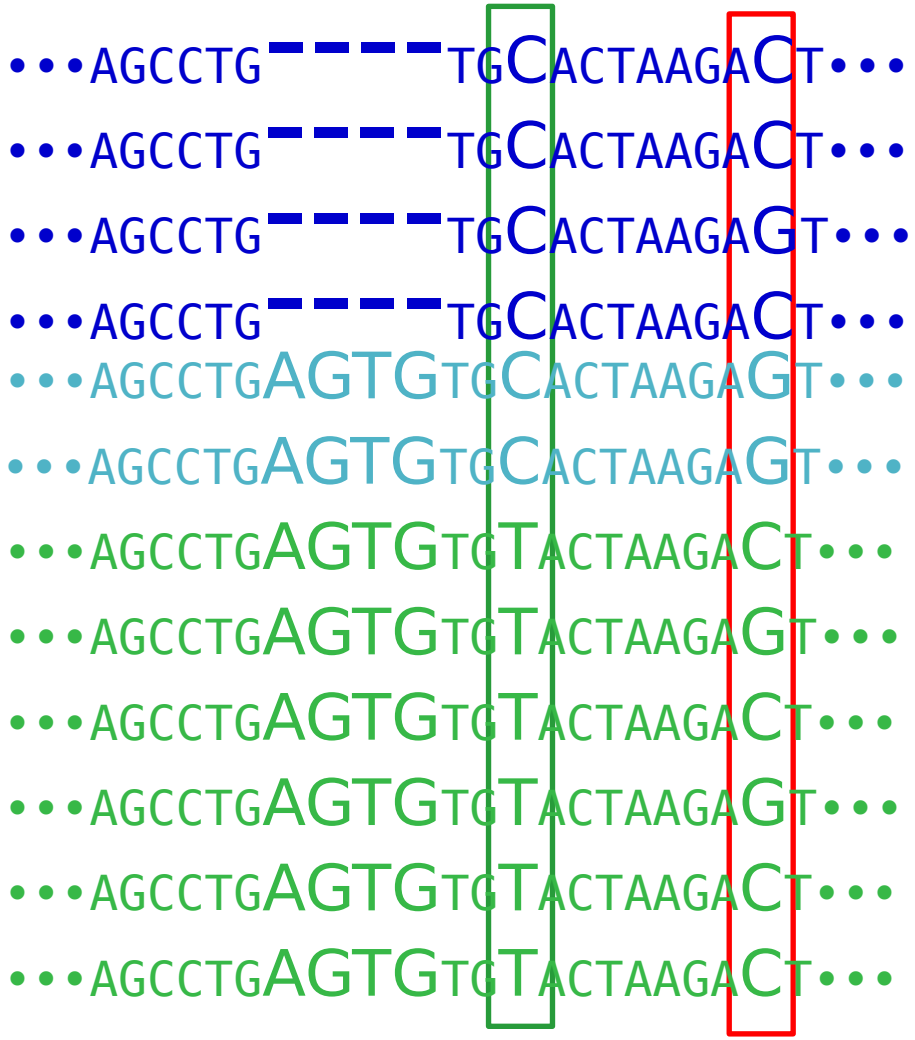
Associations between phenotype and genotype

Introduction: GWAS



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

Introduction: GWAS



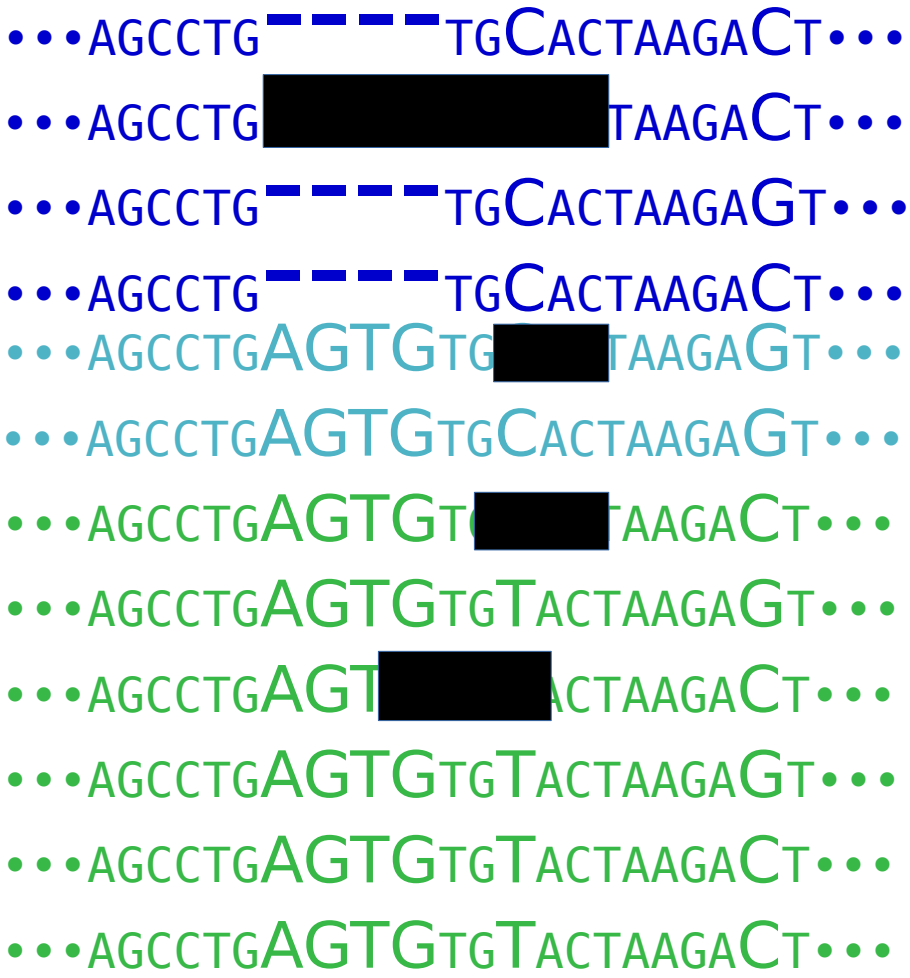
SNPs

Indels

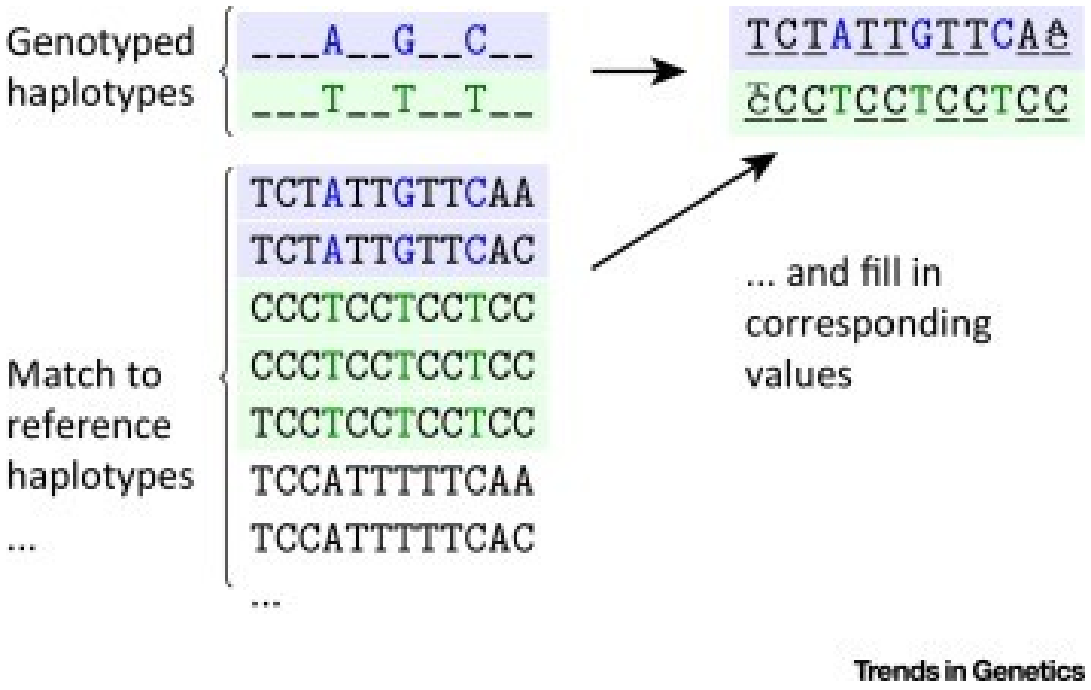
CNVs

Epigenetic markers

Real data are not perfect

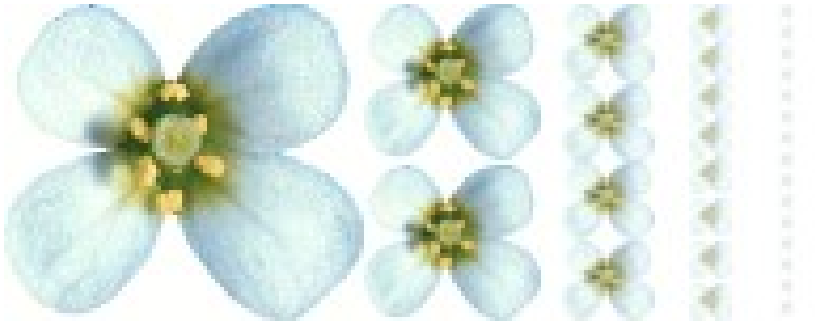


Imputation of missing data

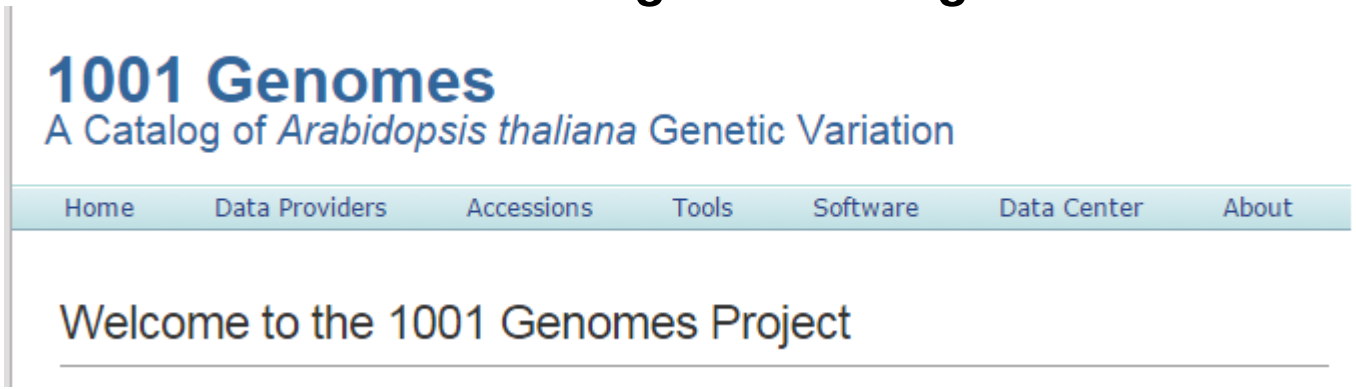


It is a simple classification problem

The 1001 Genomes Project



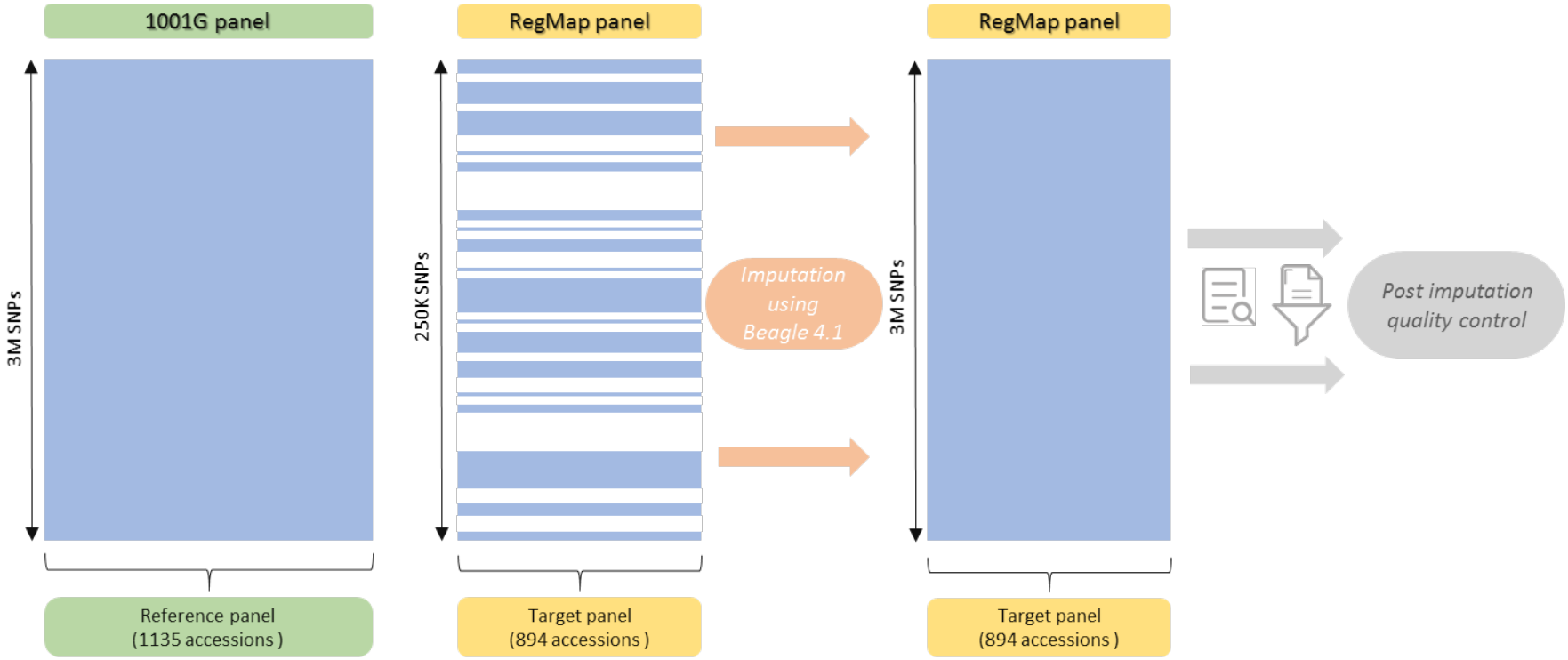
www.1001genomes.org



Data : 1,135 high quality genomes with more than 10 M SNP and 500k structural variants

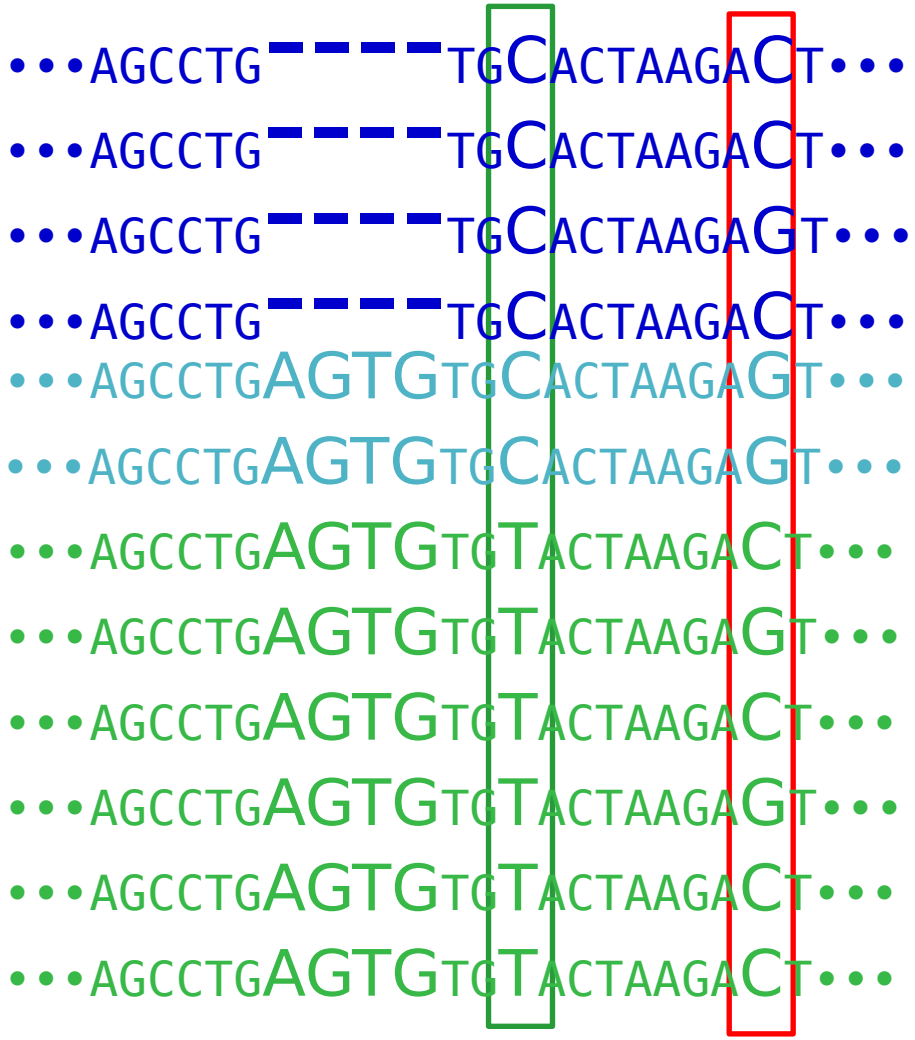
Previuos data : 1,307 SNP arrays with 250 k SNPs

Imputation of missing data



Bader *et al.* In prep.

Introduction: GWAS



SNPs

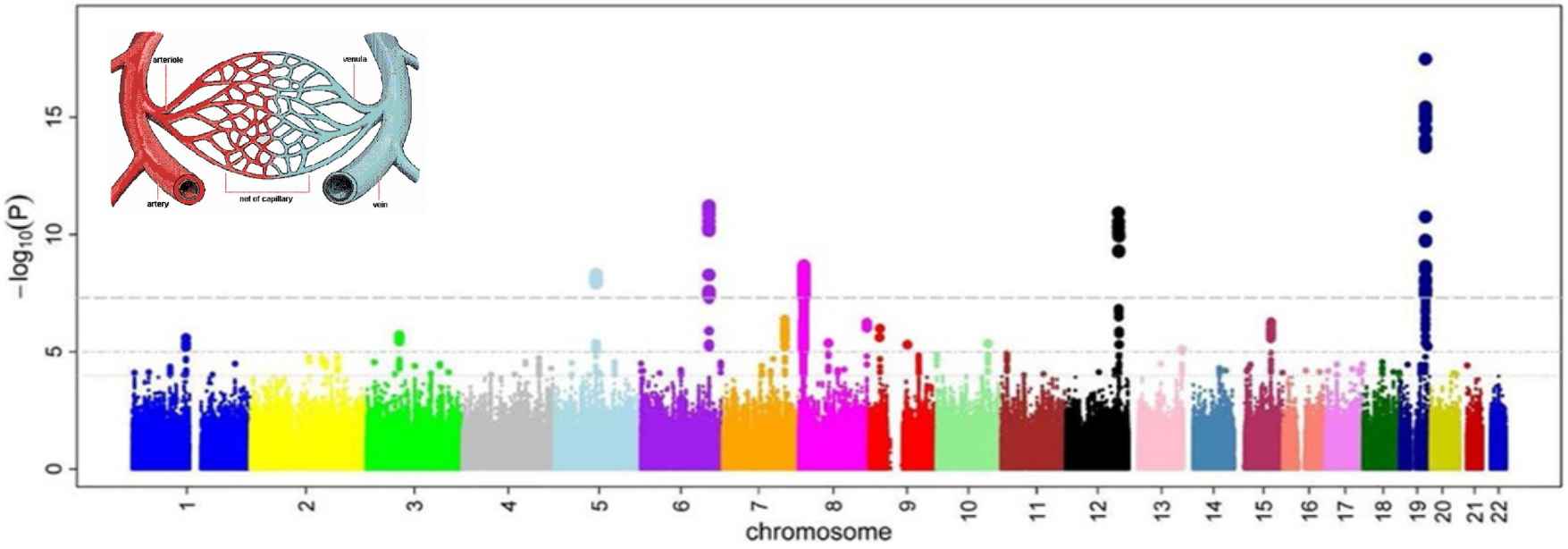
Indels

CNVs

Epigenetic markers

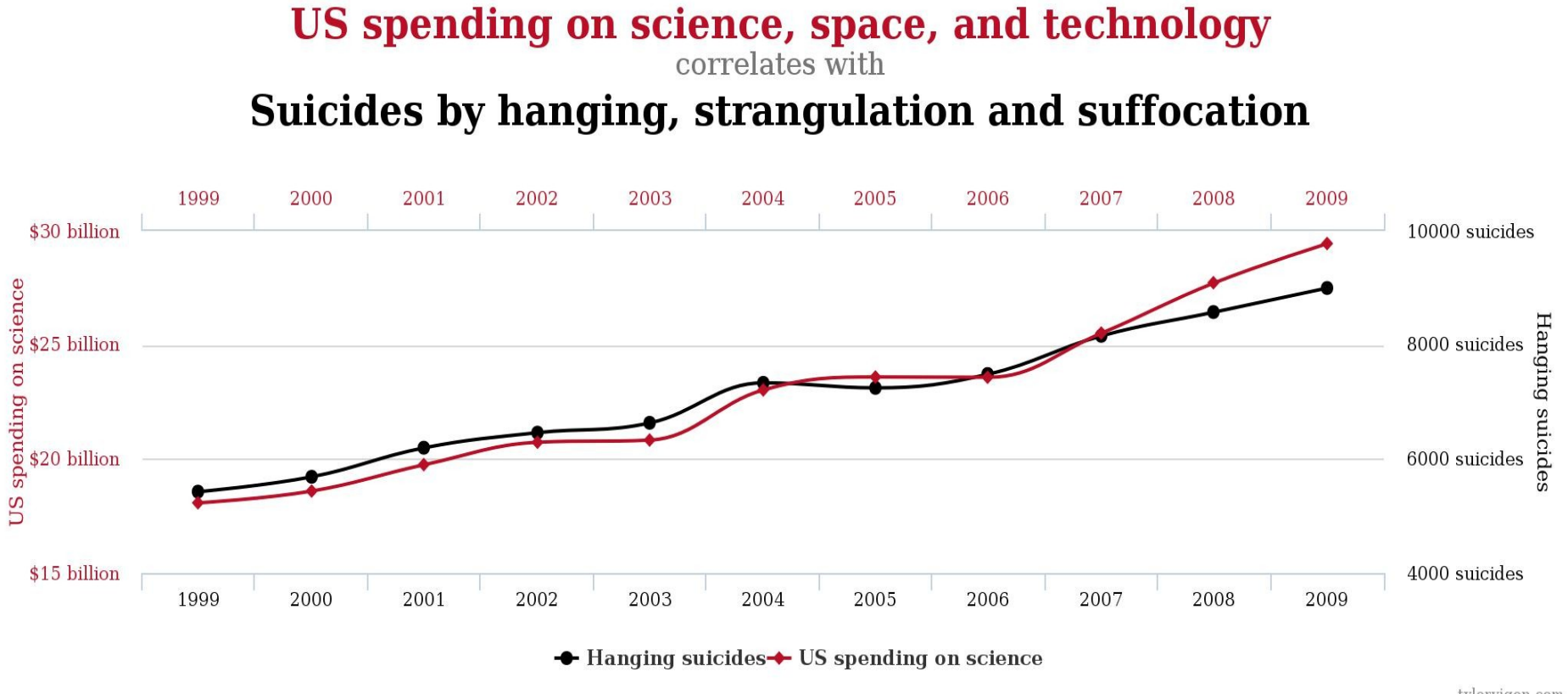
GWAS on blood microcirculation

Manhattan plot



Ikram *et al.* (2010) PLoS Genet.

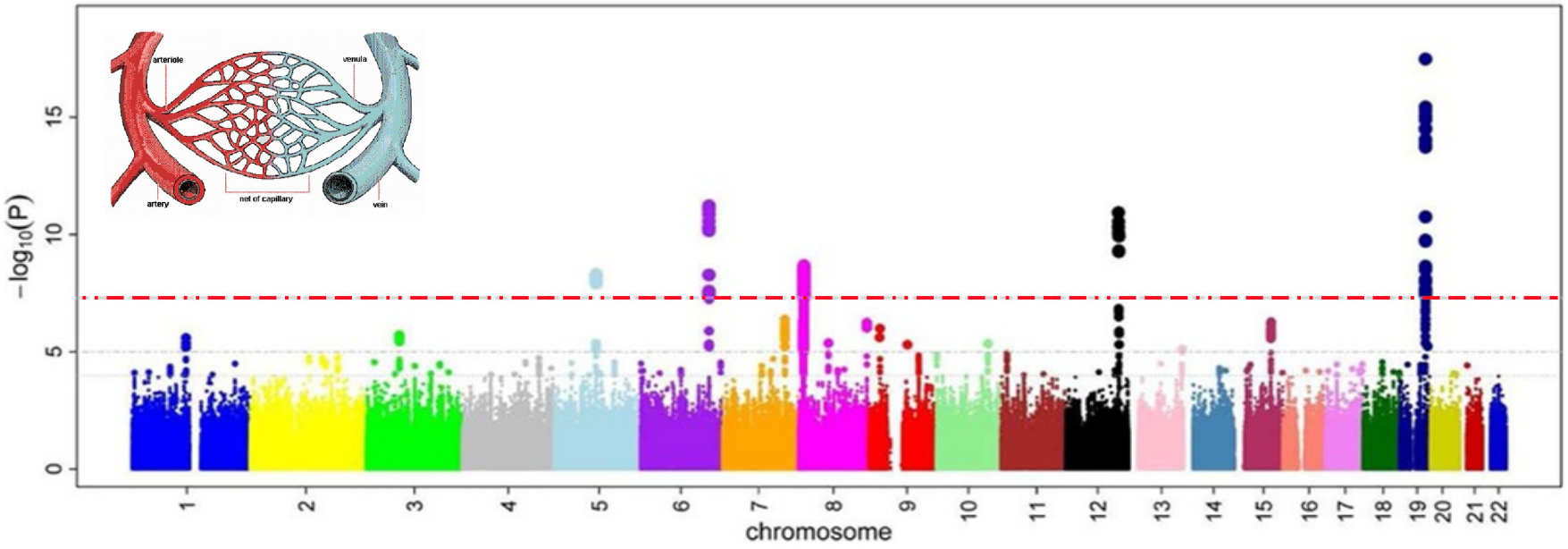
Correlation and Causality



http://www.tylervigen.com/spurious-correlations

GWAS on blood microcirculation

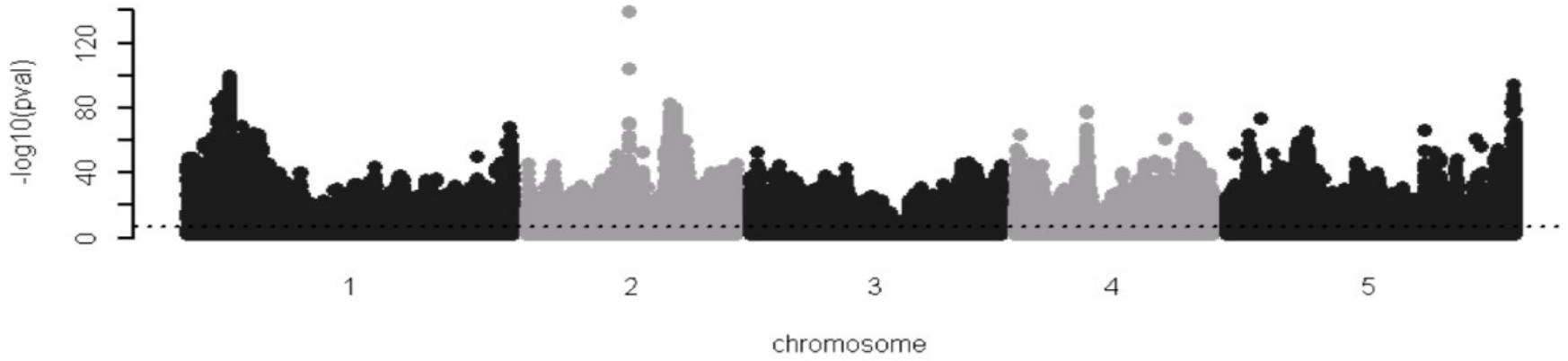
Manhattan plot



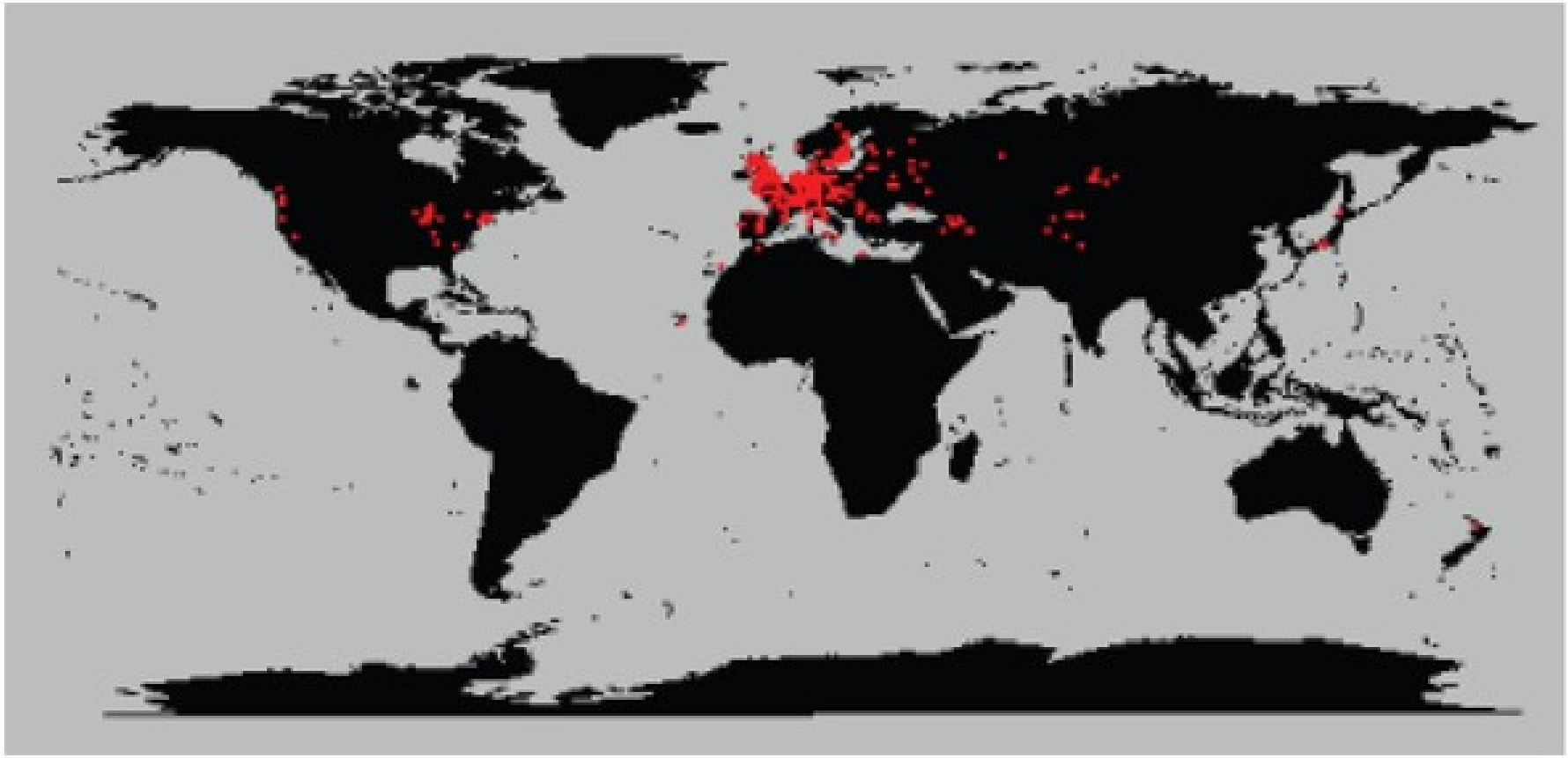
Ikram *et al.* (2010) PLoS Genet.

GWAS on flowering time in *A.thaliana*

Experiment with 925 different ecotypes in a controlled environment

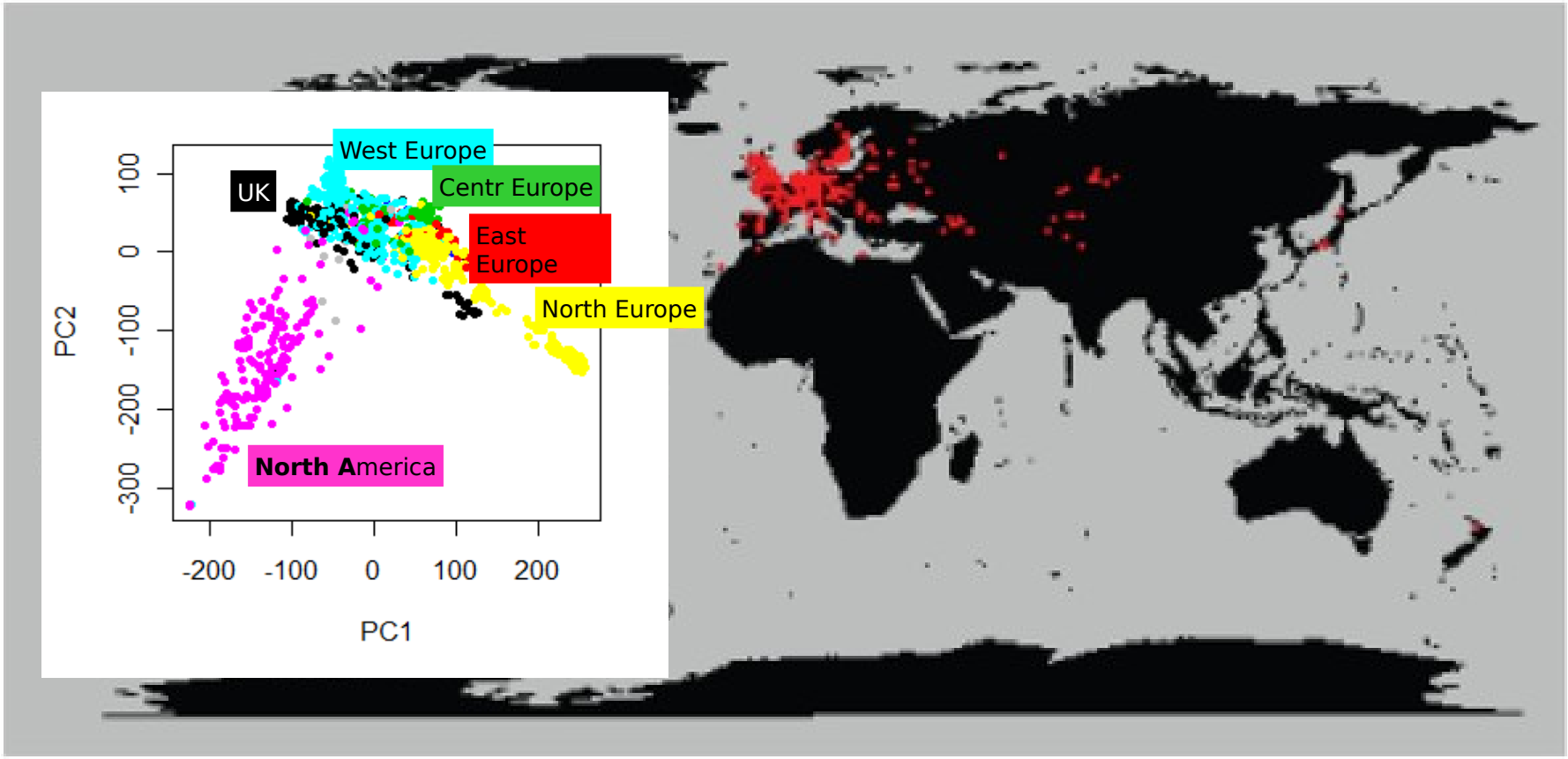


Population structure



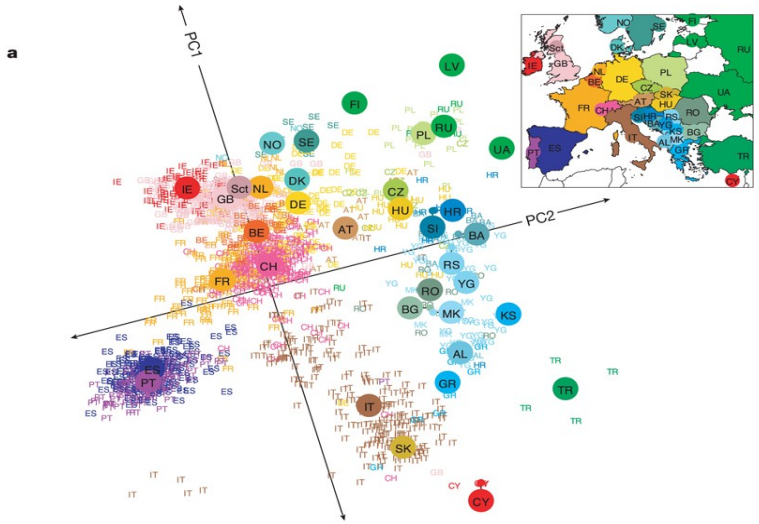
www.1001genomes.org

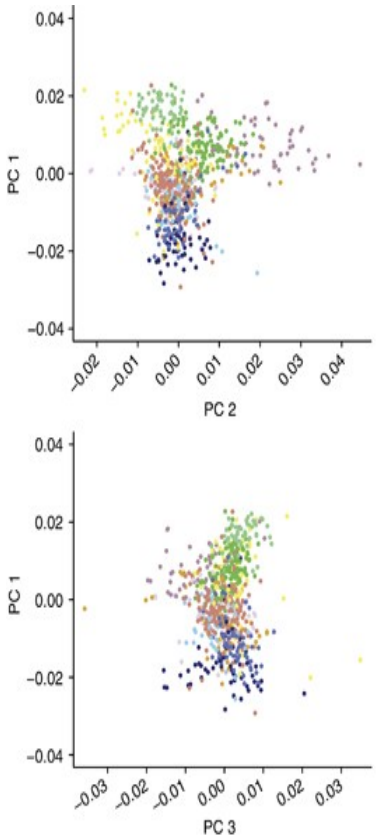
Population structure



www.1001genomes.org

Population structure in humans





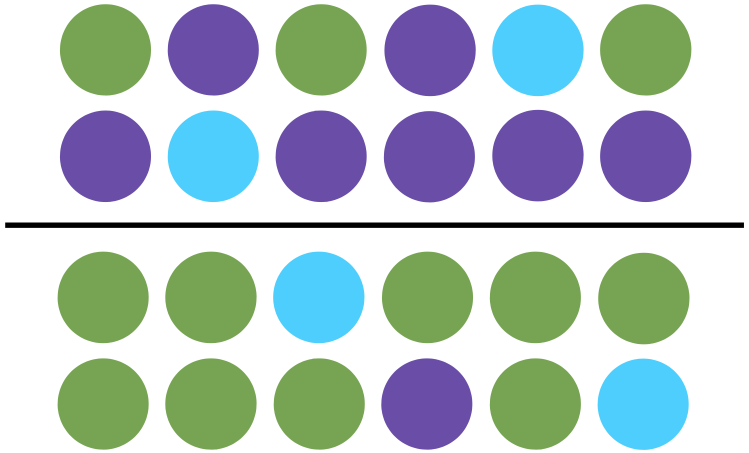
Novembre et al. (2008) Nature

Boomsma *et al.* (2013) EJHG

Population structure leads to confounding

Sub-population 1

Sub-population 2



Flowering time correlates with lattitude
Penetrance of many diseases differs in different ethic groups

How to correct for population structure

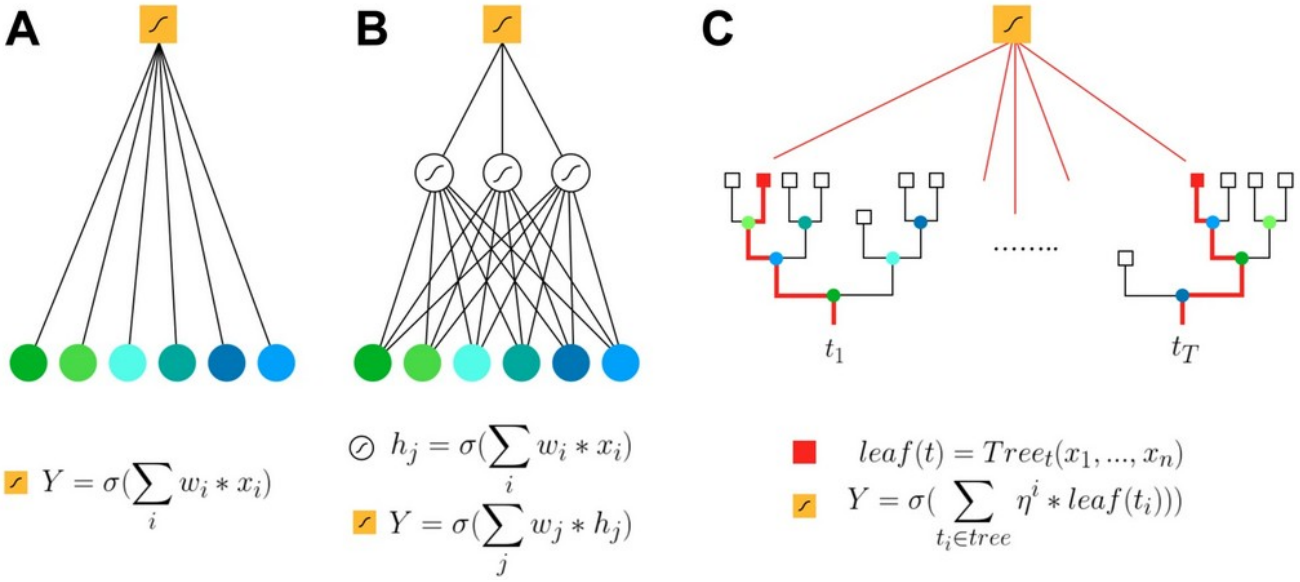
- Genomic control (Devlin & Roeder 1999, Biometrics)
- Structured association (Pritchard et al. 2000, Am.J.Hum.Genet.)
- Principal-components approach (Price et al. 2006, Nature Genet.)
- Mixed-model approach (Yu et al. 2006, Nat Genet.; Kang et al. 2008, Genetics)
- FaST-LMM -Select (Listgarten et al. 2013, Nature Genet.)
- ….
- ML (e.g. Rogmanoni et al. 2019, Scientific Reports)

SCIENTIFIC REPORTS

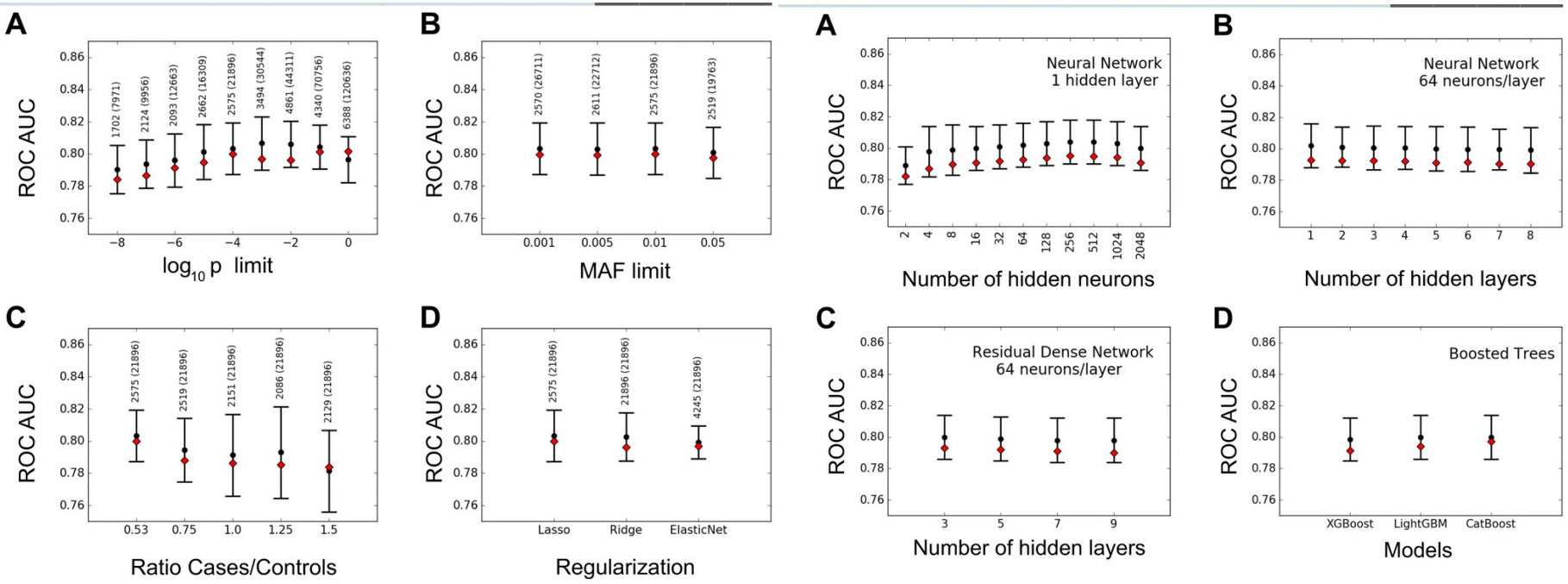
Article | Open Access | Published: 17 July 2019

Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data

Alberto Romagnoni, Simon Jégou, Kristel Van Steen, Gilles Wainrib, Jean-Pierre Hugot & International Inflammatory Bowel Disease Genetics Consortium (IIBDGC)

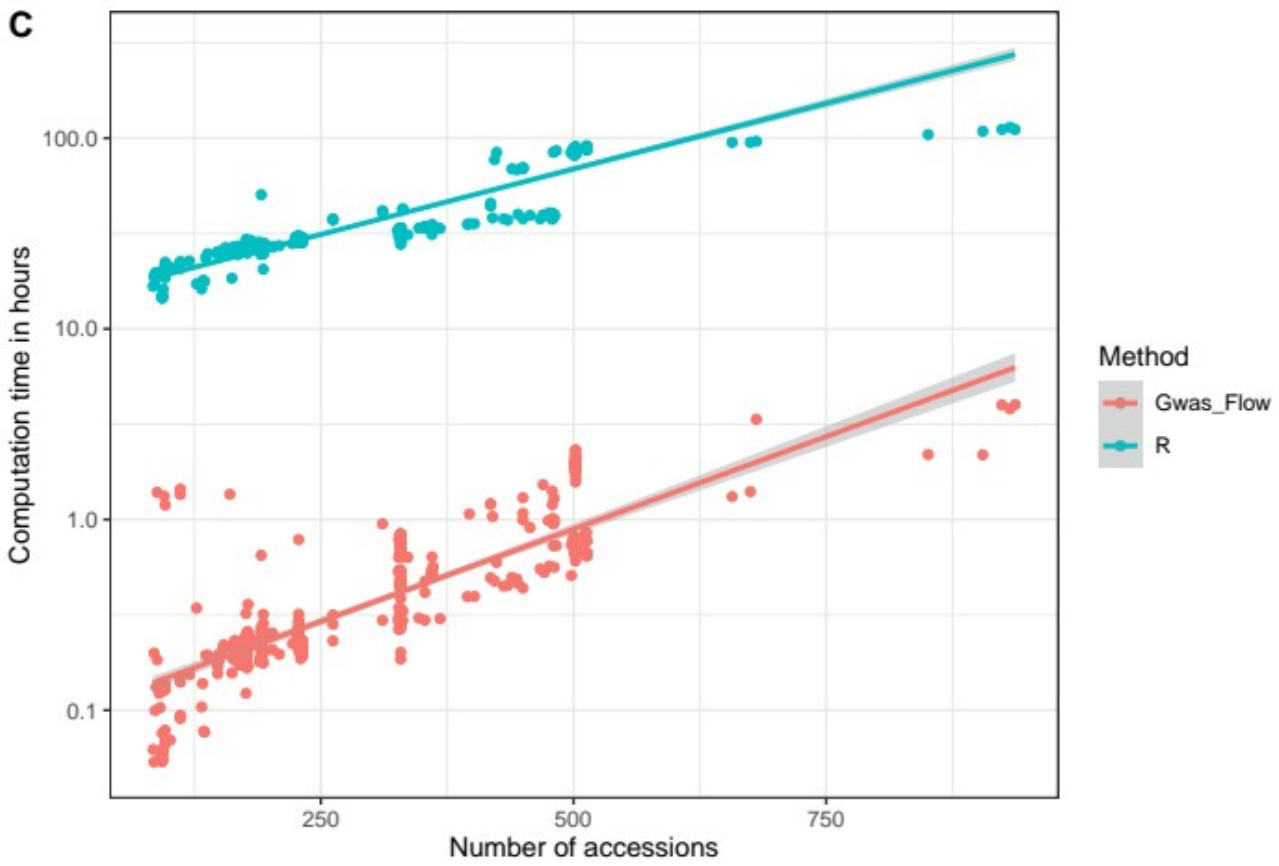


Network architecture and feature selection



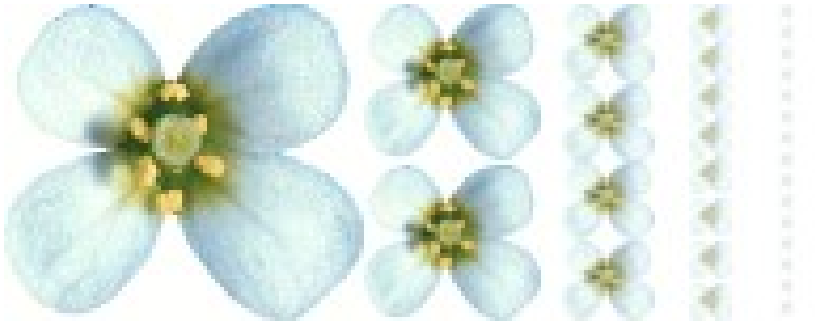
For GWAS ML (non-linear) and linear models perform on a compareable scale

Using TensorFlow for GWAS

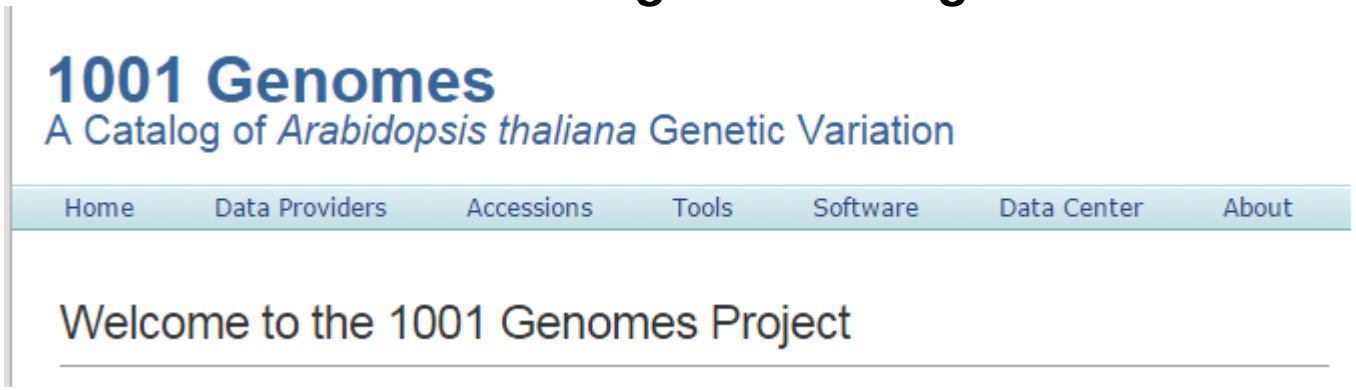


Freudenthal *et al.* In prep.

The 1001 Genomes Project



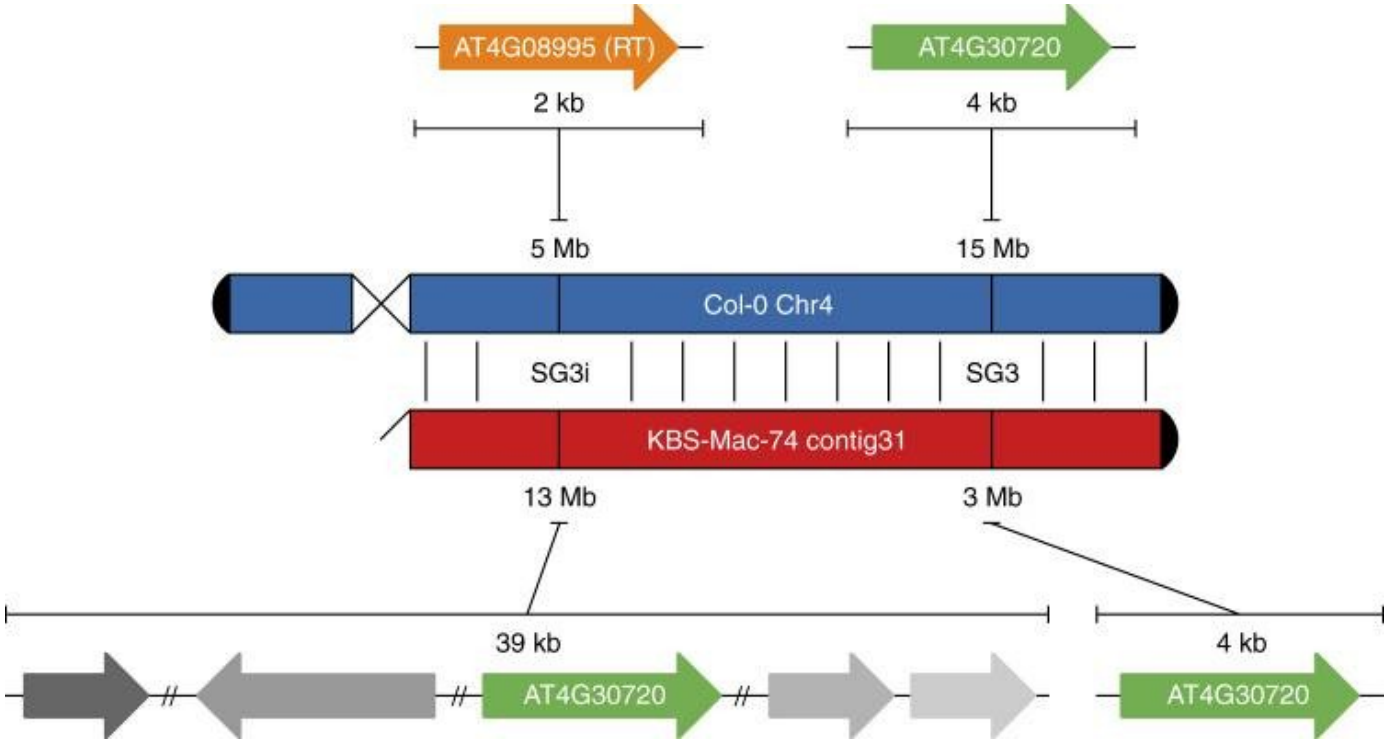
www.1001genomes.org



Data : 1,135 high quality genomes with more than 10 M SNP and 500k structural variants

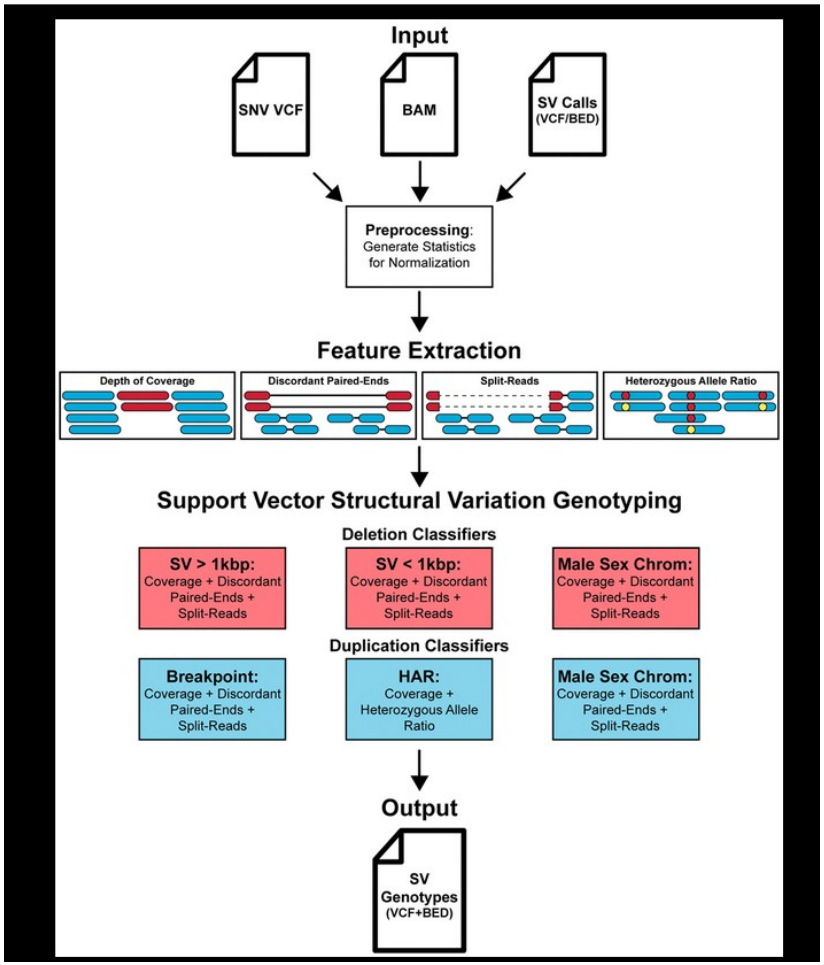
Previuos data : 1,307 SNP arrays with 250 k SNPs

Structural variation in the Arabidopsis genome



Michael *et al.* 2018 Nat Comms

ML for the detection of structural variants



Antaki *et al.* 2018 Bioinformatics

Summary GWAS

GWAS is a powerful tool to detect genotype-phenotype association which are valuable for functional follow-up studies or trait prediction,

but commonly

- **only marginal effects of markers a tested**
- **trait correlation is ignored**
- **model assumptions might be violated**
- **...**

Korte and Farlow *Plant Methods* 2013, **9**:29
http://www.plantmethods.com/content/9/1/29

PLANT METHODS

REVIEW Open Access

The advantages and limitations of trait analysis with GWAS: a review

Arthur Korte*† and Ashley Farlow†

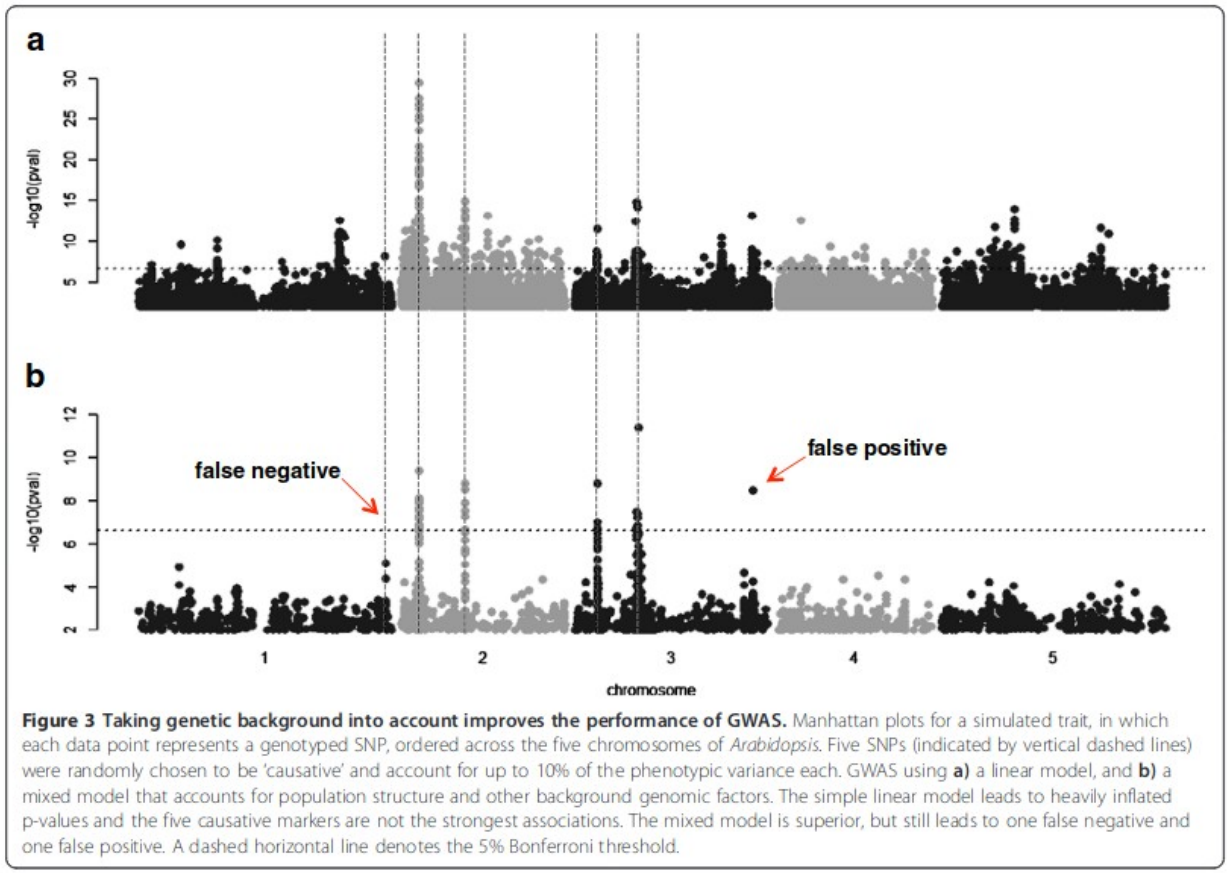
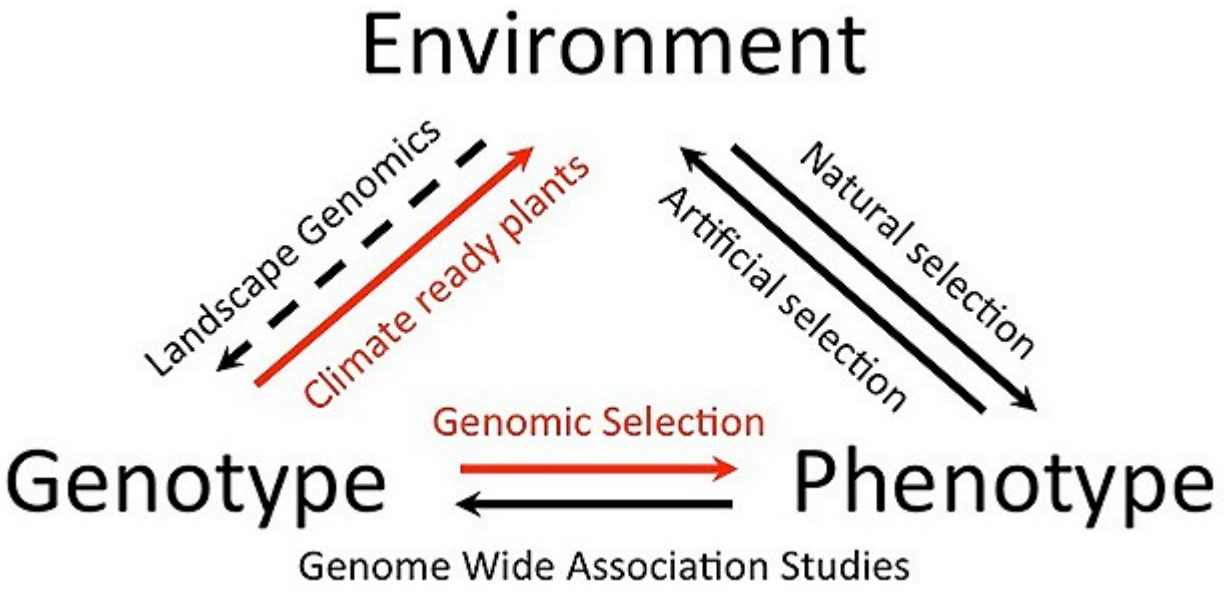
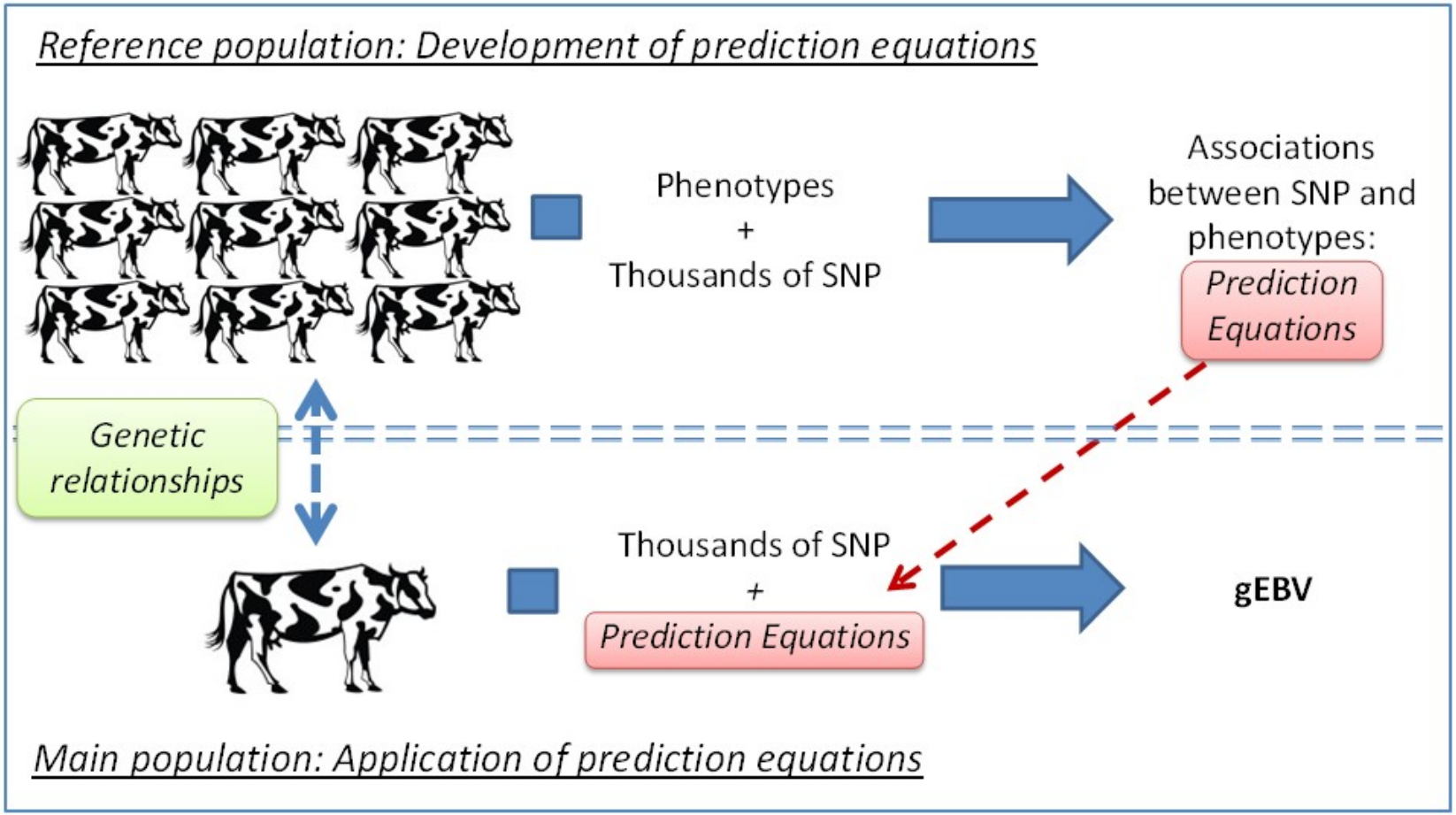


Figure 3 Taking genetic background into account improves the performance of GWAS. Manhattan plots for a simulated trait, in which each data point represents a genotyped SNP, ordered across the five chromosomes of *Arabidopsis*. Five SNPs (indicated by vertical dashed lines) were randomly chosen to be 'causative' and account for up to 10% of the phenotypic variance each. GWAS using **a)** a linear model, and **b)** a mixed model that accounts for population structure and other background genomic factors. The simple linear model leads to heavily inflated p-values and the five causative markers are not the strongest associations. The mixed model is superior, but still leads to one false negative and one false positive. A dashed horizontal line denotes the 5% Bonferroni threshold.

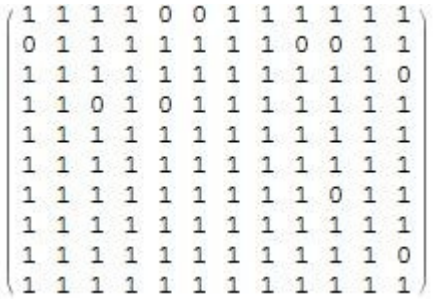
Evolutionary Genomics



Genomic prediction



Genomic Prediction



SNP Matrix

→ phenotype value (e.g. height, flowering time ...)

GWAS : $Y = \beta_0 + X\beta_1 + \upsilon + \varepsilon$, $\upsilon \sim N(0,\sigma_g K), \varepsilon \sim N(0,\sigma_e I)$

GP : $\beta_0 + X\beta_1 + \upsilon Z + \varepsilon = Y$

Might be possible to learn the betas instead of using a Gaussian distribution

Decomposition of the phenotypic variance

$$\sigma_P = \sigma_G + \sigma_E + \sigma_{GxE}$$

$$\sigma_G = \sigma_A + \sigma_D + \sigma_I$$

Decomposition of the phenotypic variance

GP : $\beta_0 + X\beta_1 + \upsilon Z + \varepsilon = Y$

$$\sigma_g = \sigma_A + \sigma_I$$

Decomposition of the phenotypic variance

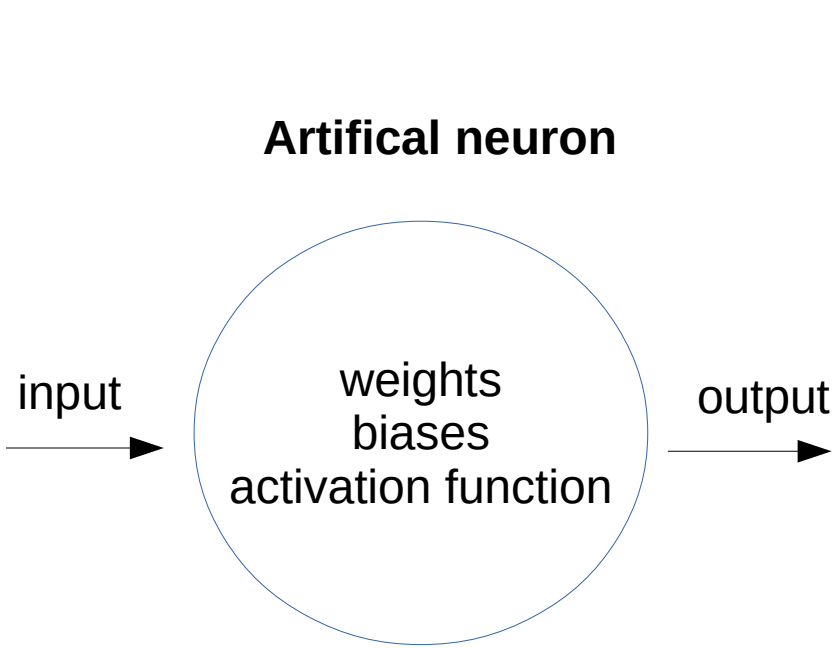
GP : $\beta_0 + X\beta_1 + \upsilon Z + \varepsilon = Y$

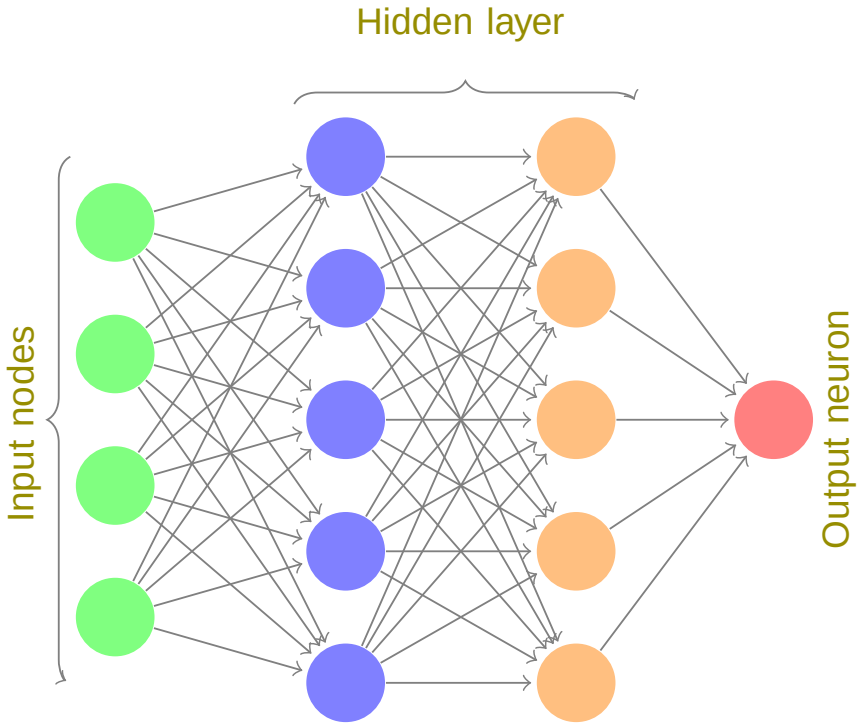
$$\sigma_g = \sigma_A + \sigma_I$$

perfect for linear models

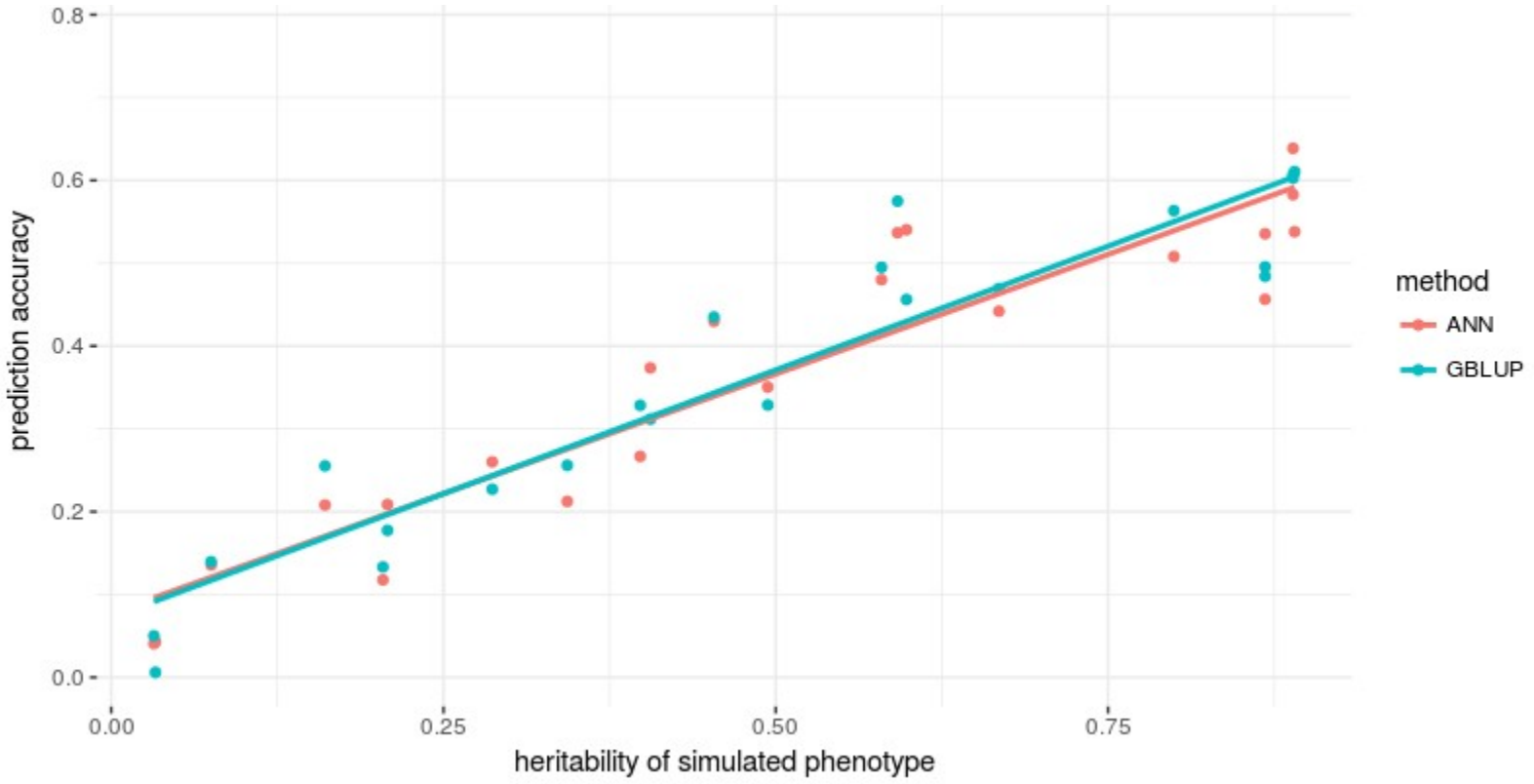
Non-linear interaction,
perfect for ML

ML for Genomic Prediction





ML for Genomic Prediction



Saiz *et al.* unpublished

ML does not outperform classical linear models

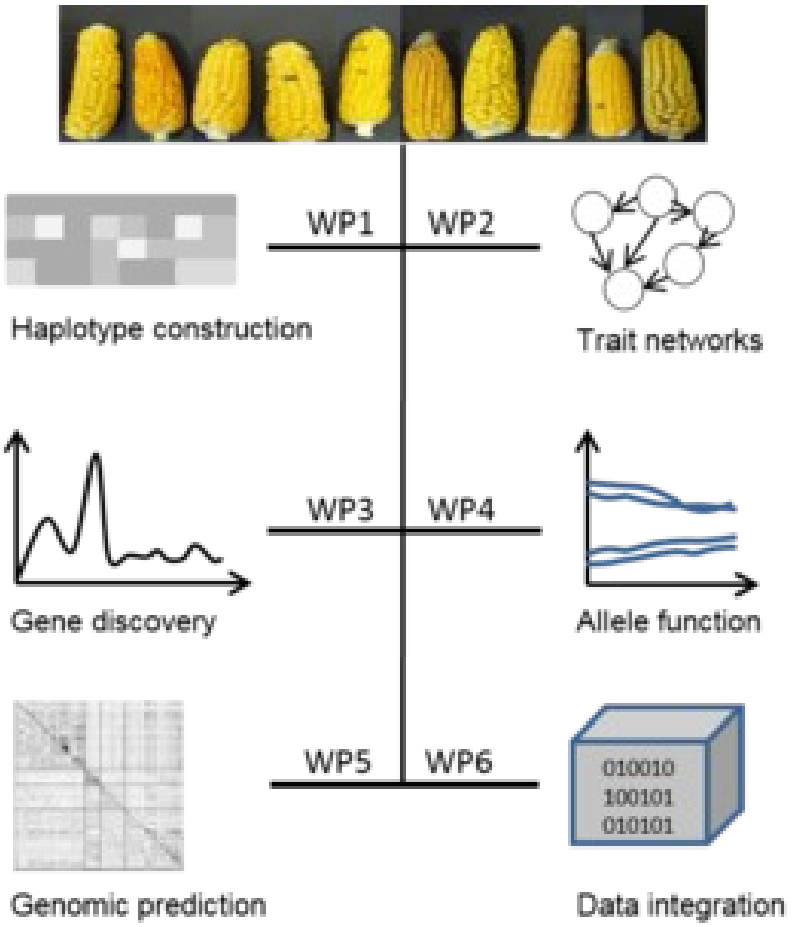
$$GP: \quad \beta_0 + X\beta_1 + \upsilon Z + \varepsilon = Y$$

$$\sigma_g = \sigma_A + \sigma_I$$

perfect for linear models

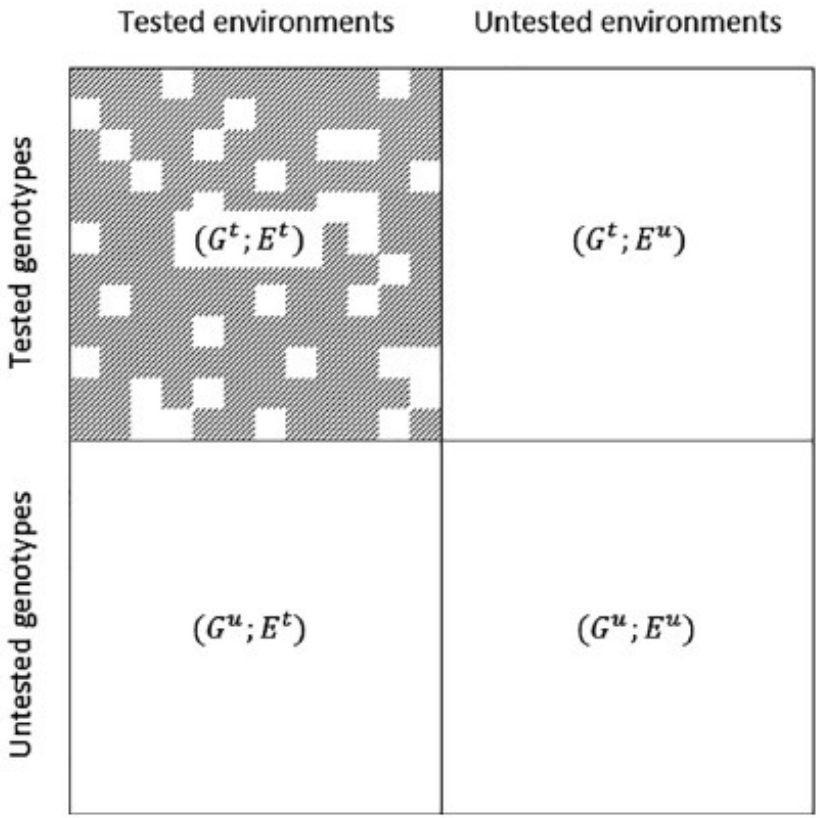
Non-linear interaction, perfect for ML

MAZE - Assessing the genomic and functional diversity of maize to improve quantitative traits



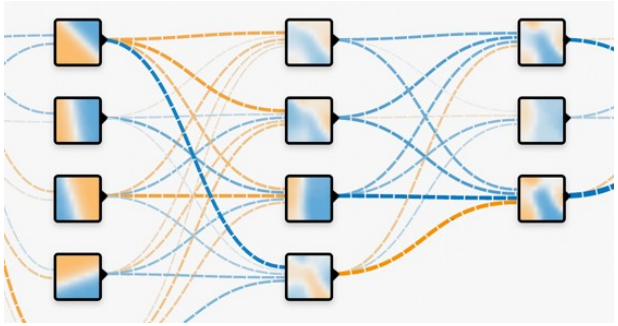
SPONSORED BY THE
Federal Ministry of Education and Research

Genomic prediction

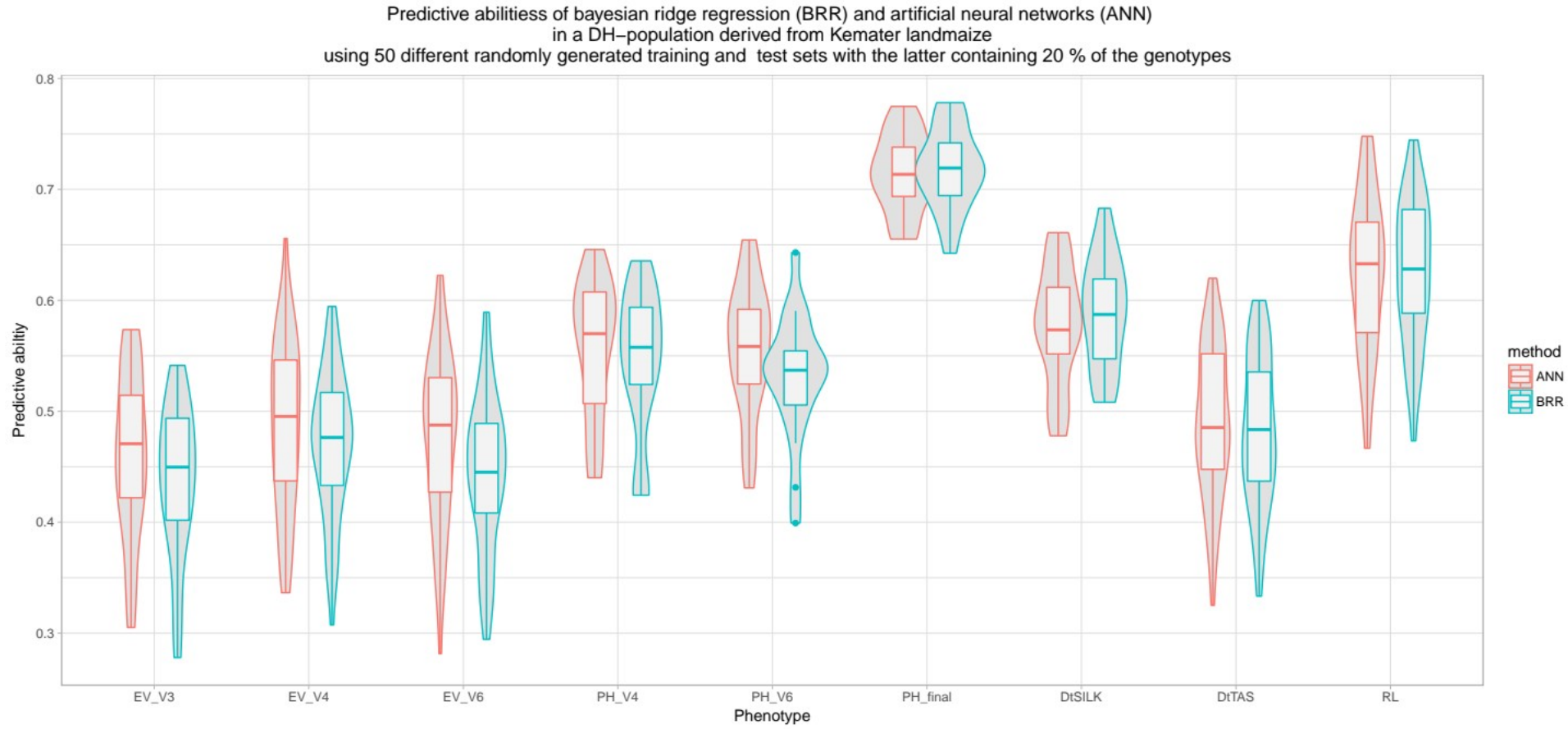


Malosetti et al. (2015) Crop Science

Machine learning



ML for Genomic Prediction



Freudenthal *et al.* unpublished

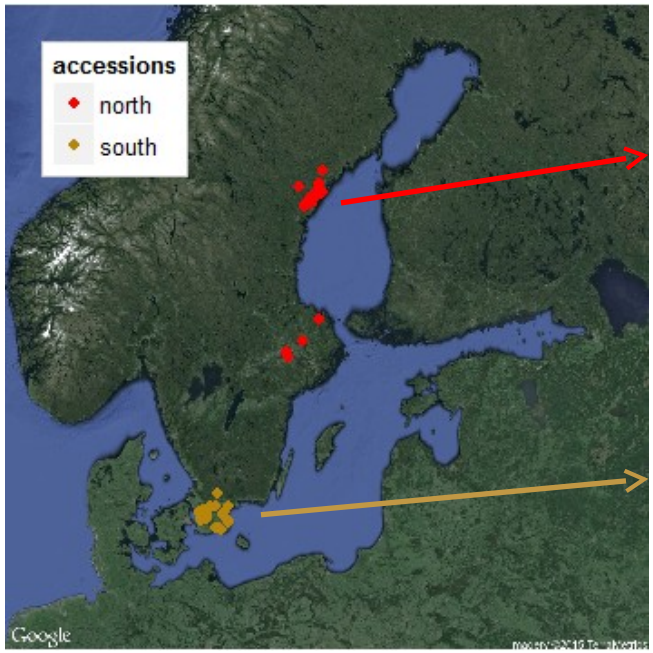
Summary GP

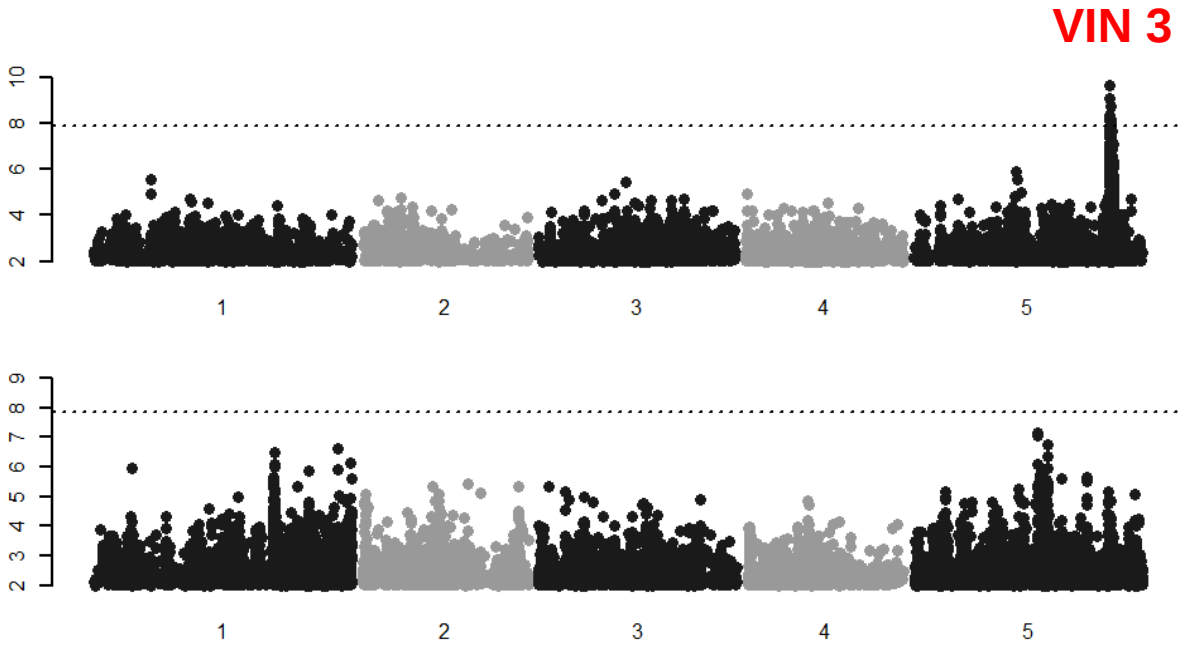
ML offers some potential for Genomic Prediction

but

current models do not outperform existing additive models

Back to GWAS

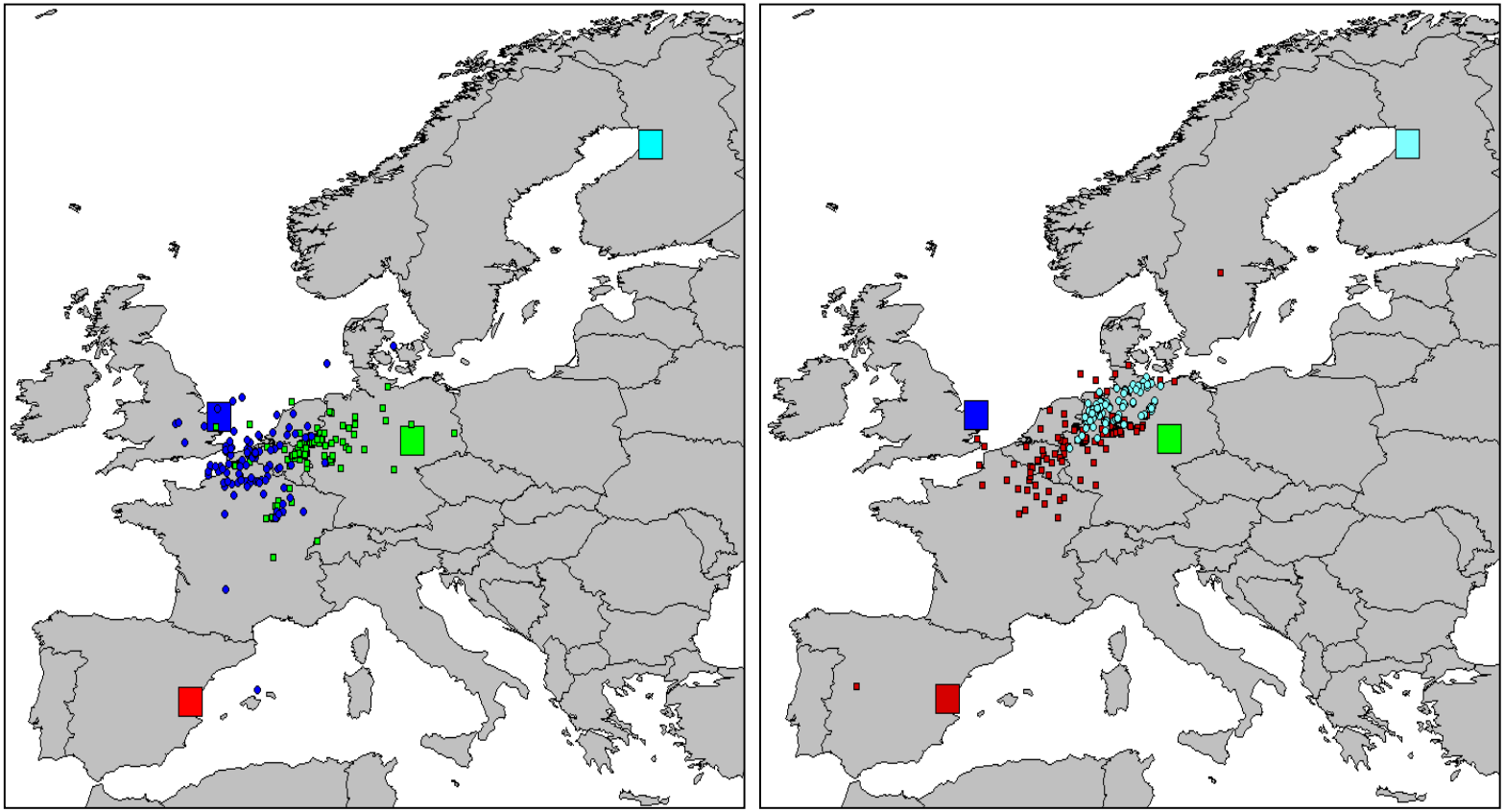




Different results in different subsets

Lopez *et al.* in prep

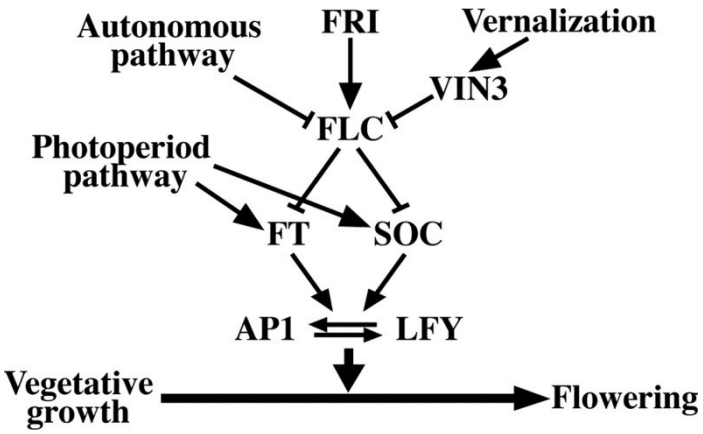
Local Adaptation



Genes providing fitness under natural conditions are mostly local

Fournier-Level *et al.* (2011) Science

GWAS in local subsets



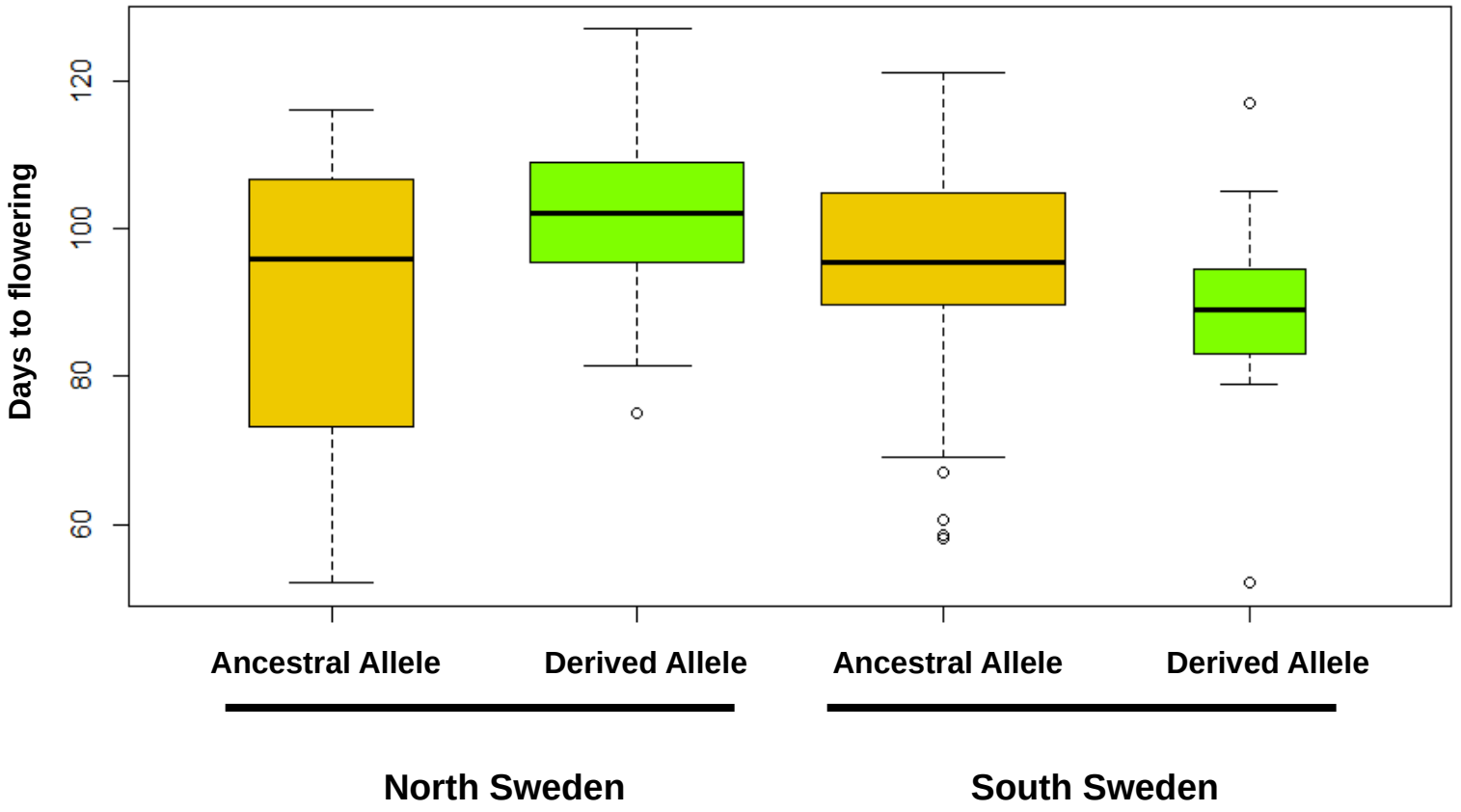
Caicedo *et al. (2004)* PNAS

- **Epistasis**

$$GWAS : Y = \beta_0 + X\beta_1 + \upsilon + \varepsilon$$

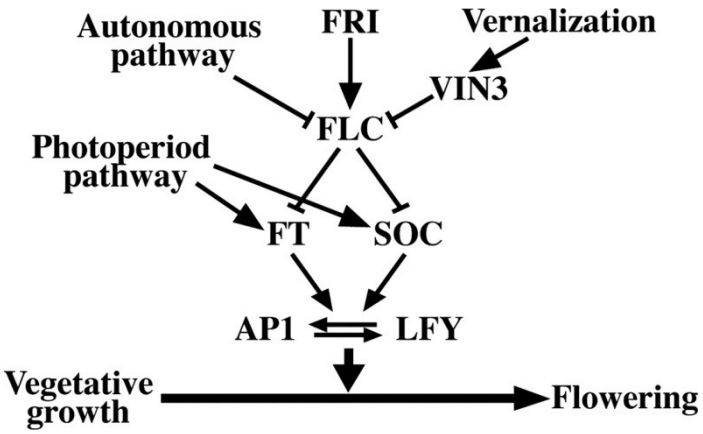
$$GWAS : Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_1X_2\beta_{12} + \upsilon + \varepsilon$$

Effect of the *VIN3* Allele on Flowering time



The effect of the respective allele depends on the genetic background

GWAS in local subsets



Caicedo *et al. (2004)* PNAS

- **Epistasis**

$$GWAS : Y = \beta_0 + X\beta_1 + \upsilon + \varepsilon$$

$$GWAS : Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_1X_2\beta_{12} + \upsilon + \varepsilon$$

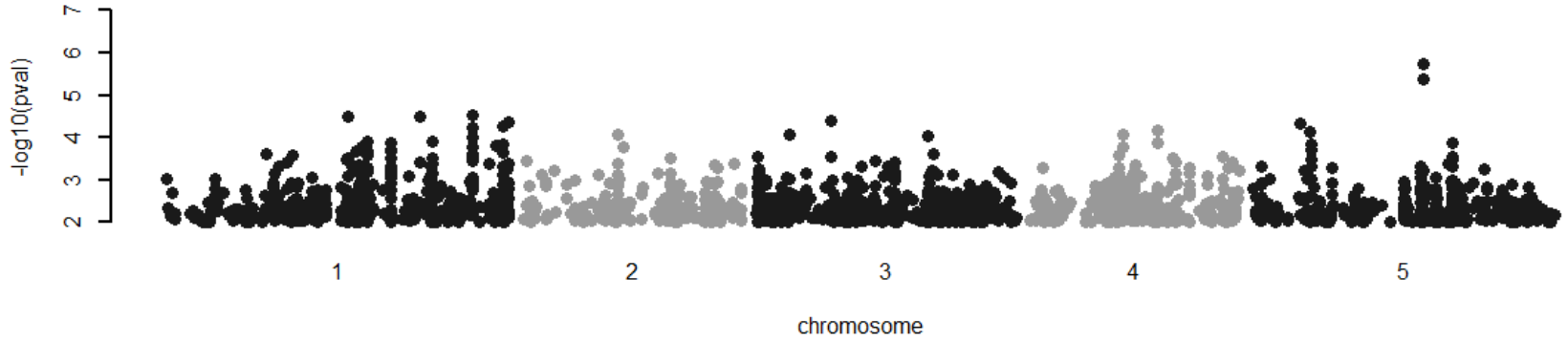
GWAS models to test for epistasis

SNP by SNP epistasis

Genome-wide epistasis with the lead SNP in *VIN3*

3 different models with two homozygous SNPs

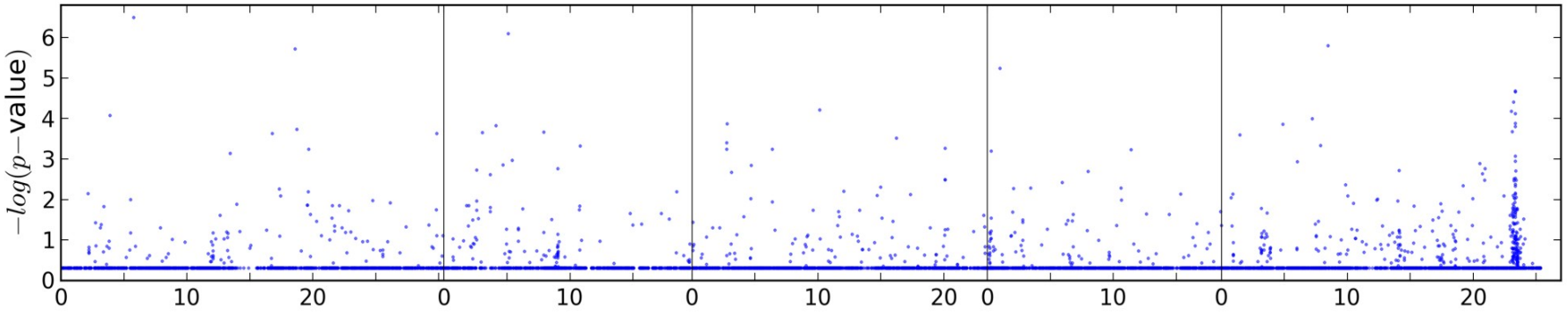
AND, OR and XOR



GWAS models to test for epistasis

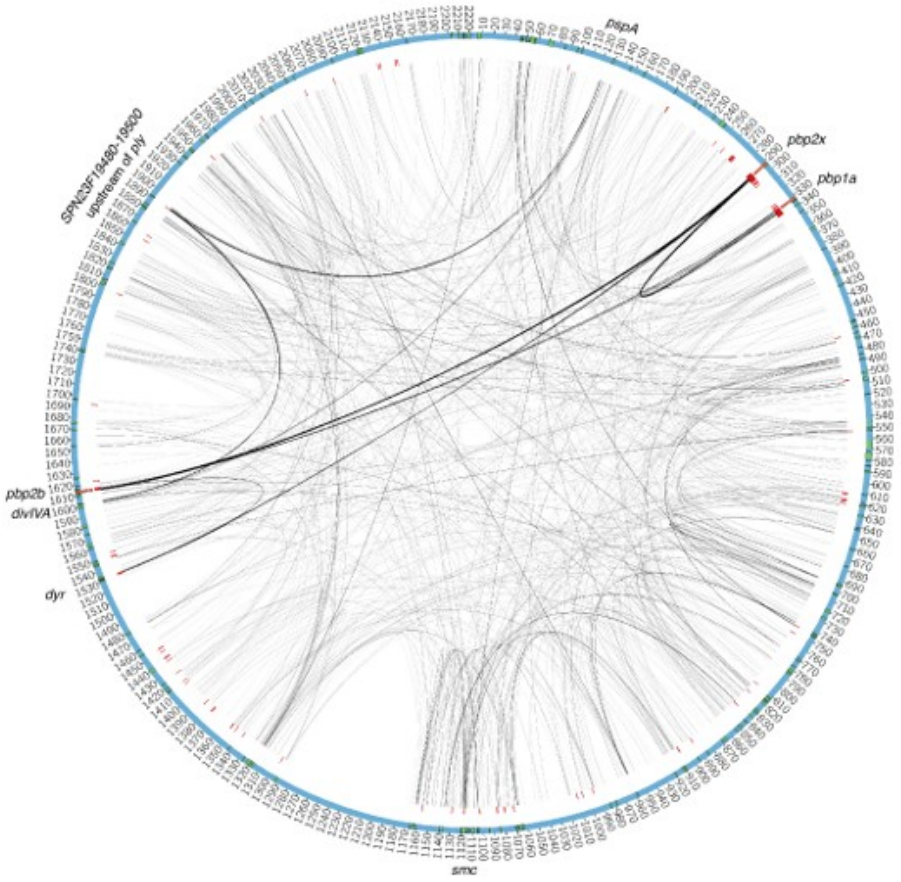
Epistasis with the genomic background in a mixed model setting

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



Different SNPs in *VIN3* seems to interact with the genomic background

Epistatic interactions in a bacterial genome



Skwark *et al.* (2017) Plos Genetics

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

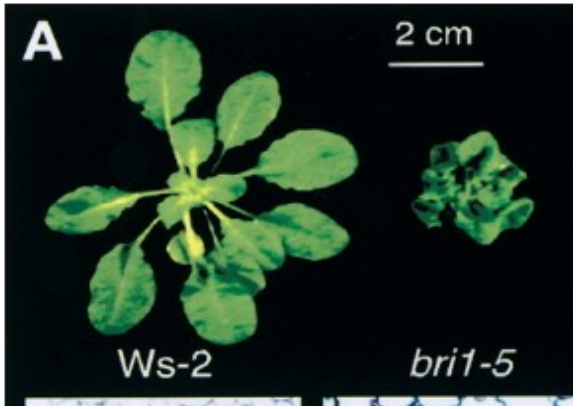
28,148 SNPs lead to a premature STOP codon

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

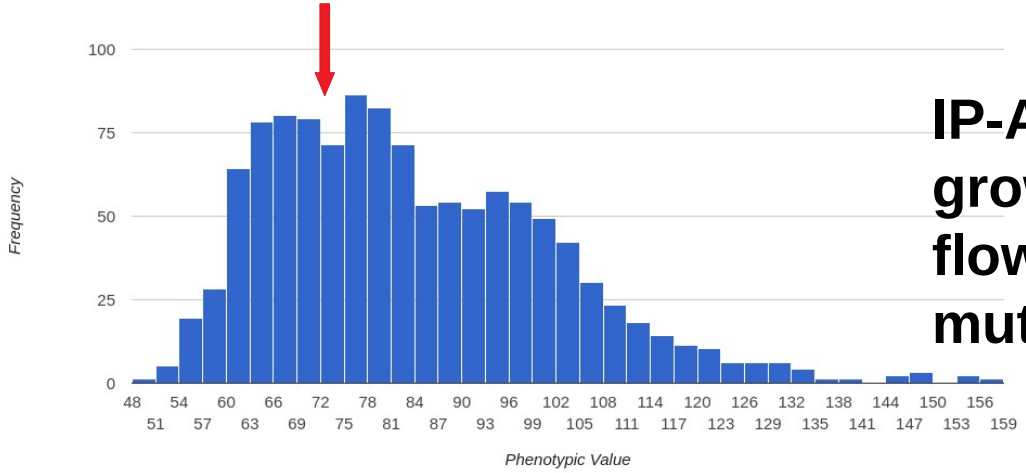
Natural BRI1 knock out in an accession from Portugal (IP-Alo-0)

Natural BRI1 knock out in an accession from Portugal (IP-Alo-0)

Natural BRI1 knock out in an accession from Portugal (IP-Alo-0)

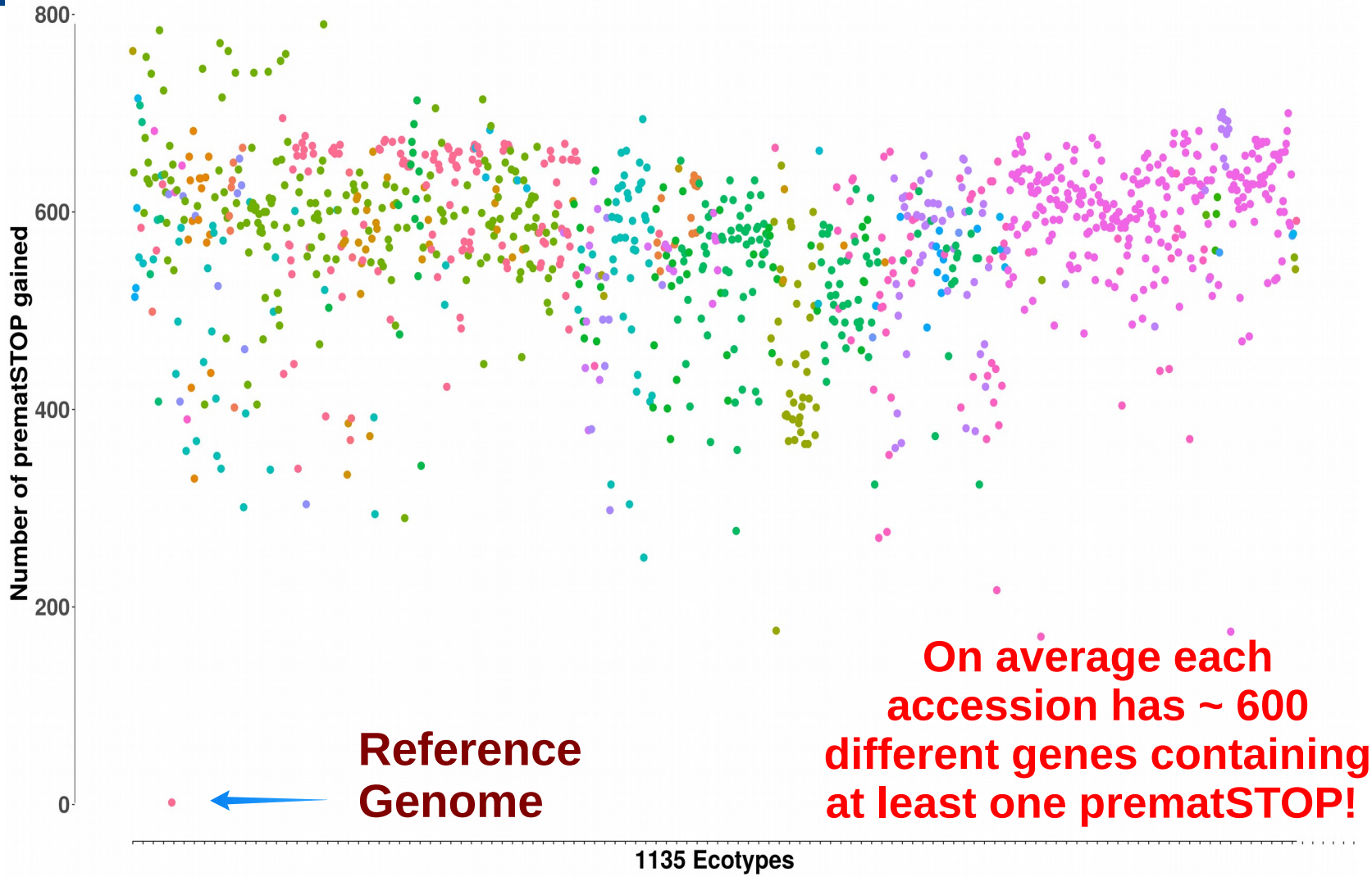


Santiago Mora-García *et al.* 2004, Genes Dev.



IP-Alo-0 does neither show the growth defect nor the delay in flowering time observed in *bri1* mutants

Frequency of prematSTOP codons by accessions



Co-occurence of premature Stop codons

Filter for functional prematSTOP by using available RNAseq data (Kawakatsu *et al.* 2016)

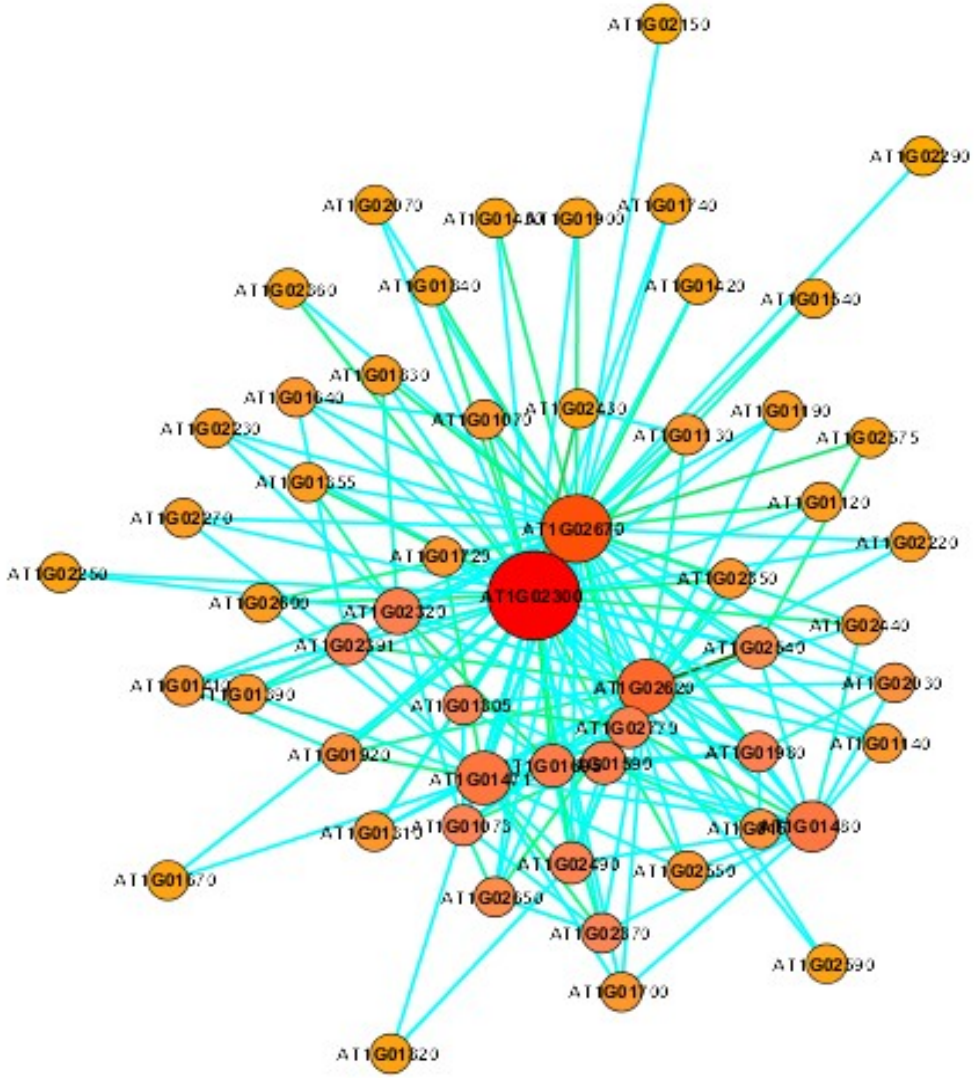
~ 6.000 genes / 16.000 prematSTOP

Gene 1	Gene 2	gene1_count	gene2_count	co-occurrence	P-value over	P-value under
AT2G25850	AT3G45910	27	72	2	0.52	0.75
AT3G55780	AT5G51000	450	435	237	$<10^{-20}$	1

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

If we use synonymous SNPs as a control we don't observe this co-occurence

Network of epistatic interactions in *A.thaliana*



Use this network architecture as prior knowledge for GWAS

Anwar *et al*. in prep

MTMM (Multi-trait mixed model)

Combine correlated traits
(same trait in two environments | longitudinal data)

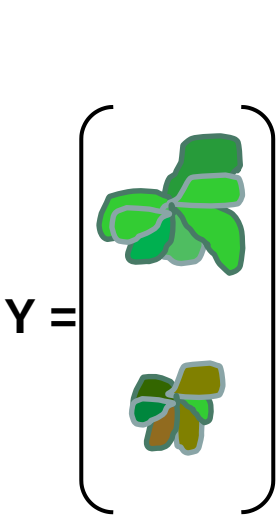
Flowering time and Yield

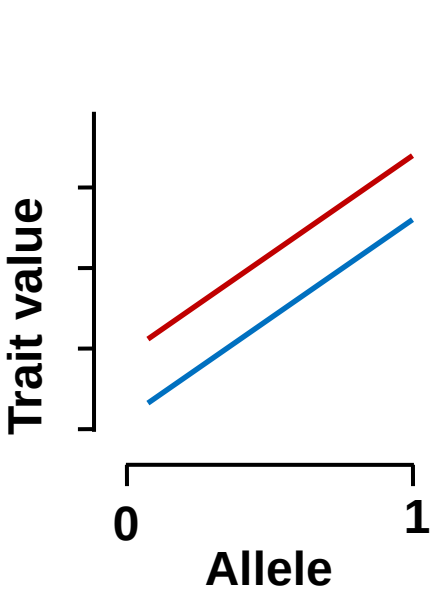
Leaf growth in normal and water limiting conditions

Plant size on day x as a function of plant size on day x-1

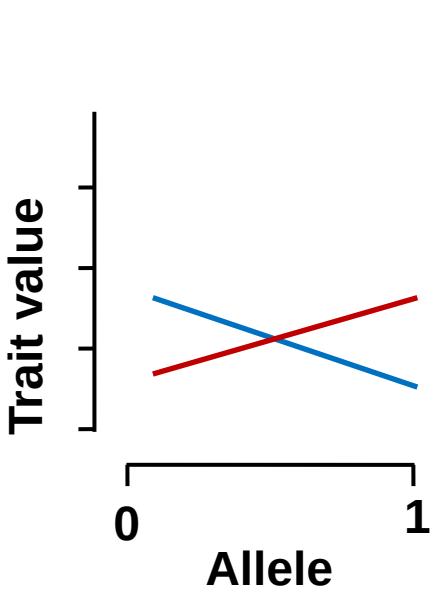
MTMM (Multi-trait mixed model)

Combine correlated traits
(same trait in two environments | longitudinal data)

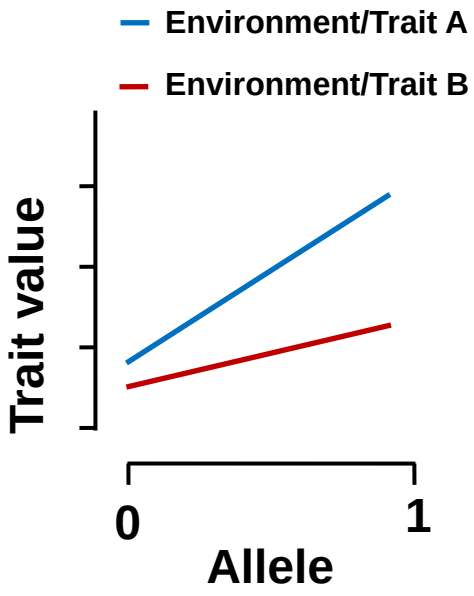




Trait common



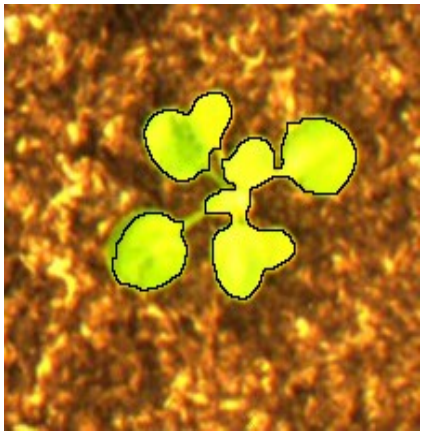
Trait speciic

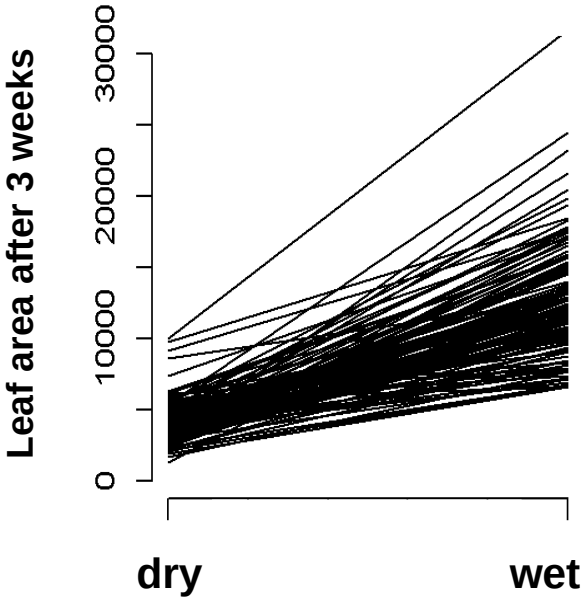


Full test

Korte *et al.* (2012) Nat.Genetics

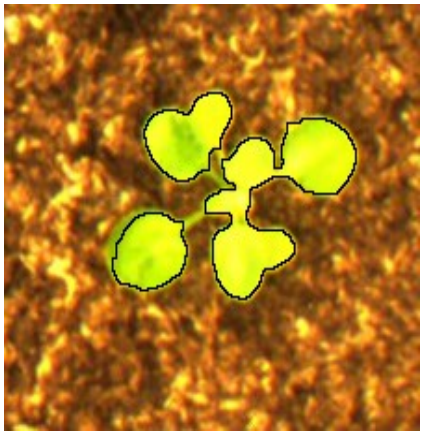
Drought experiment

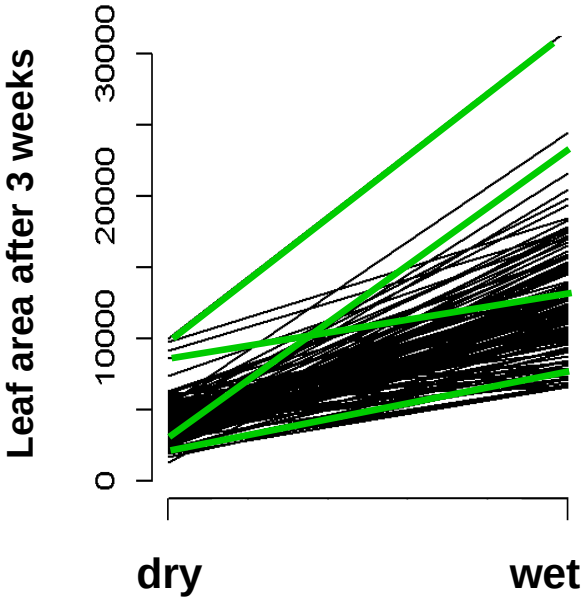




Phenotypic differences in how drought affect seedling growth

Drought experiment

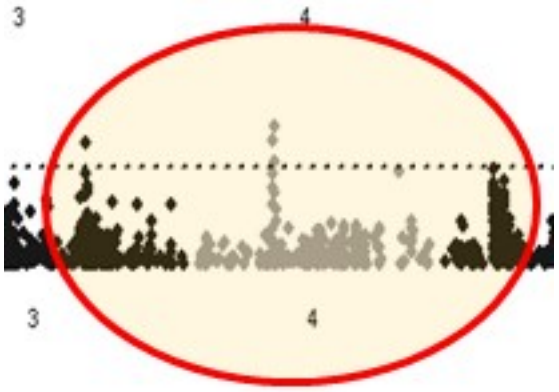




Phenotypic differences in how drought affect seedling growth

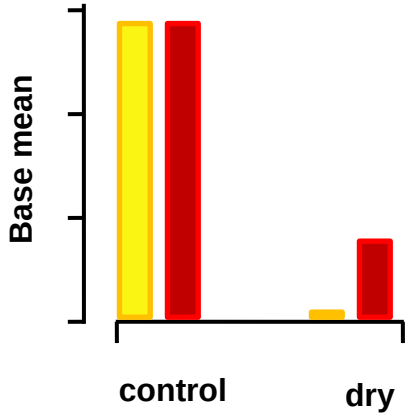
Example : MTMM on Leaf area day 21

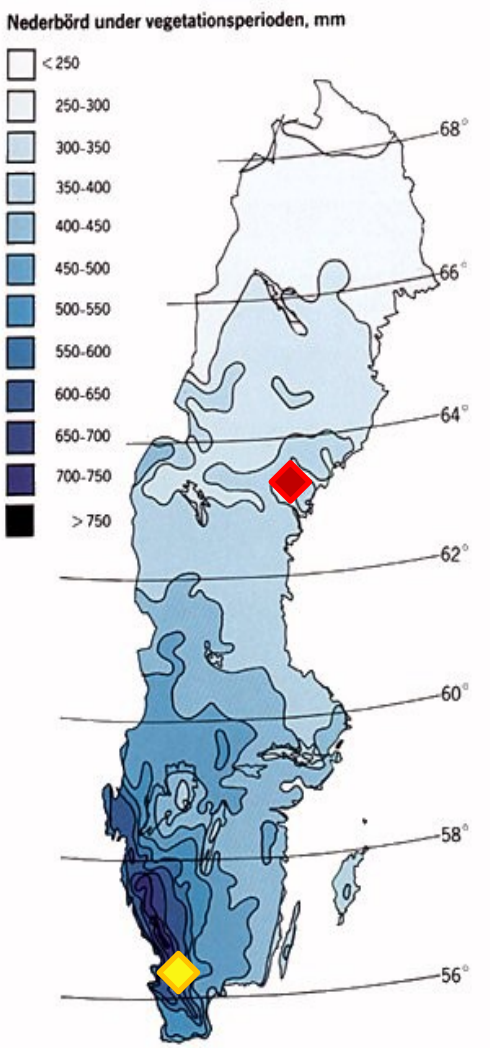
MTMM
Trait specific test



AT4G08950 EXO
(response to brassinosteroids)

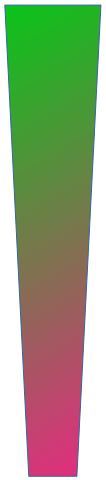
RNA expression of AT4G08950



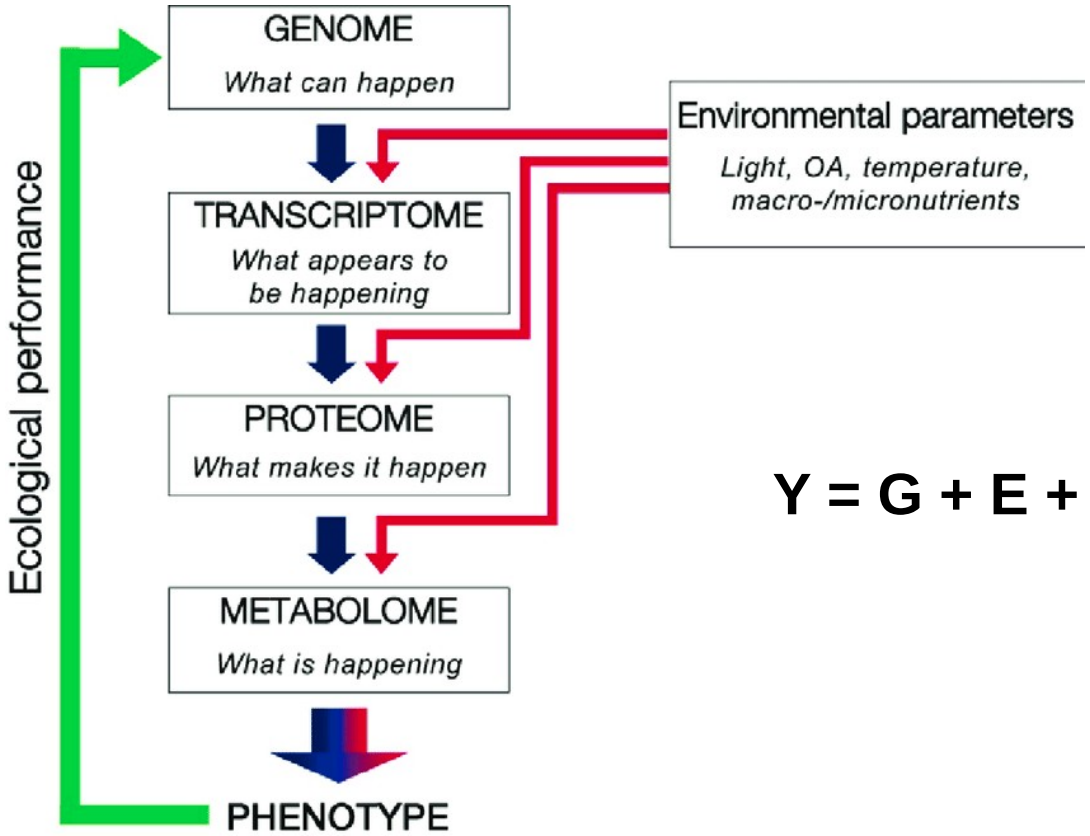


The Omnics cascade

stable



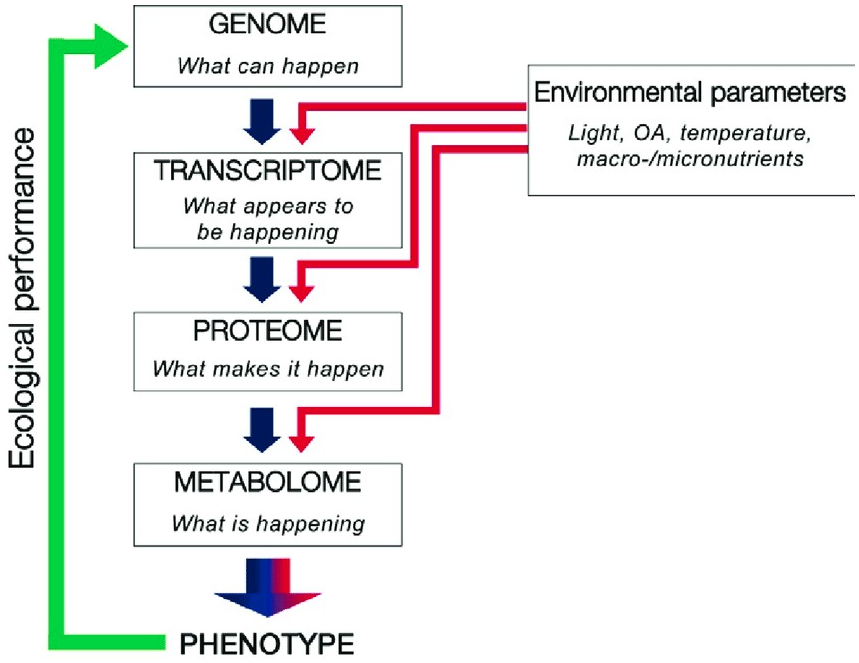
instable

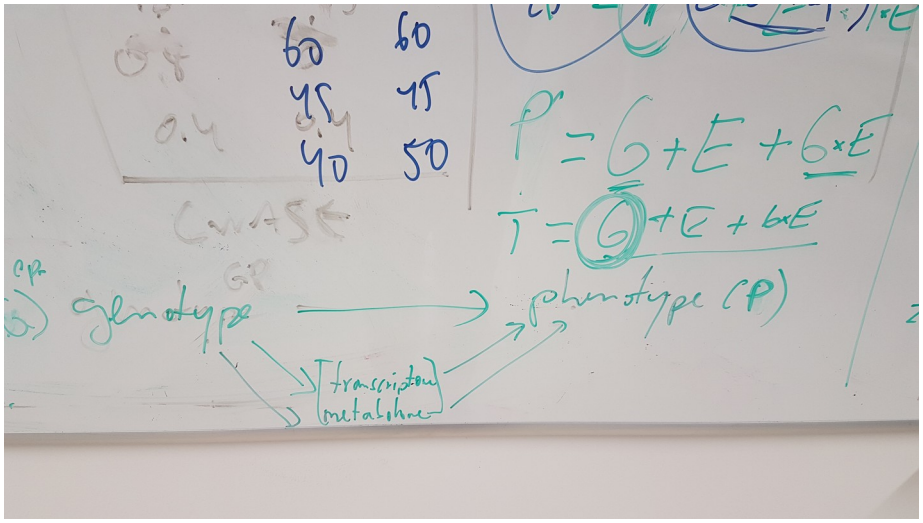


$$Y = G + E + GxE$$

Dettmer et al. (2007) Mass Spect Rev

GWAS / TWAS / eGWAS



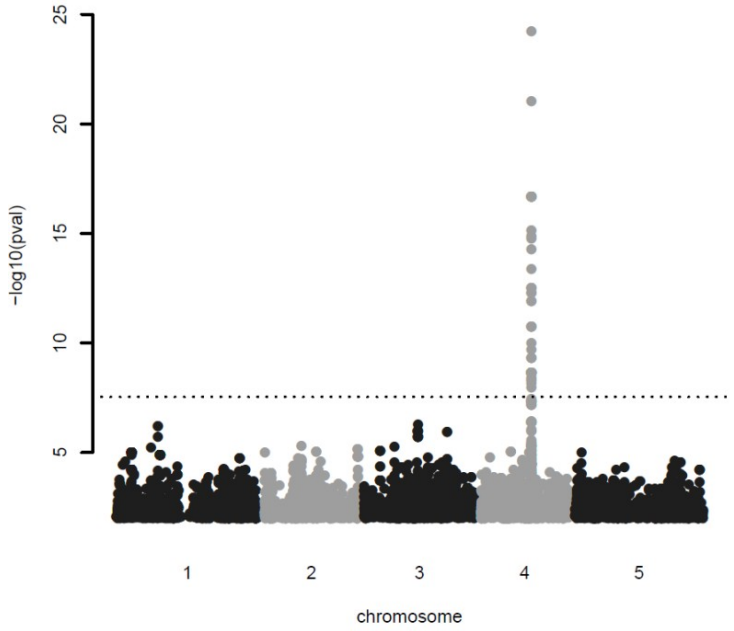


GWAS : Genome - Phenotype

TWAS : Transcriptome - Phenotype

eGWAS : Genome - Transcriptome

eGWAS



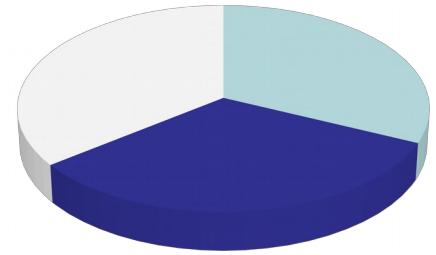
A SNP on Chromosome 4 affects the expression of the gene AT4G19480
How does this affect the phenotype ?

Integration of different data layers into joint models
(Structural Equation Models, ML)

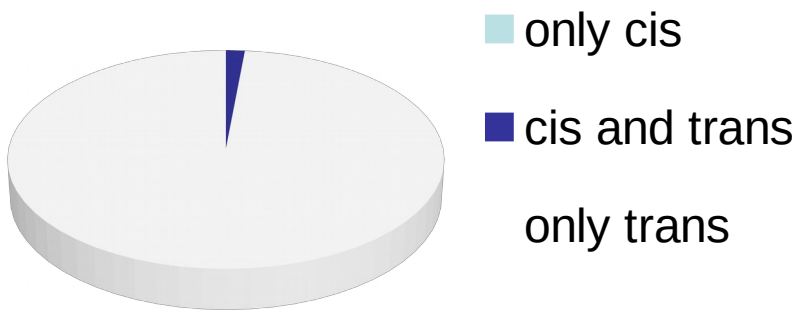
Clauw *et al.* (2016) Plant Cell

Summary of MTMM on RNAseq data

Environment independend



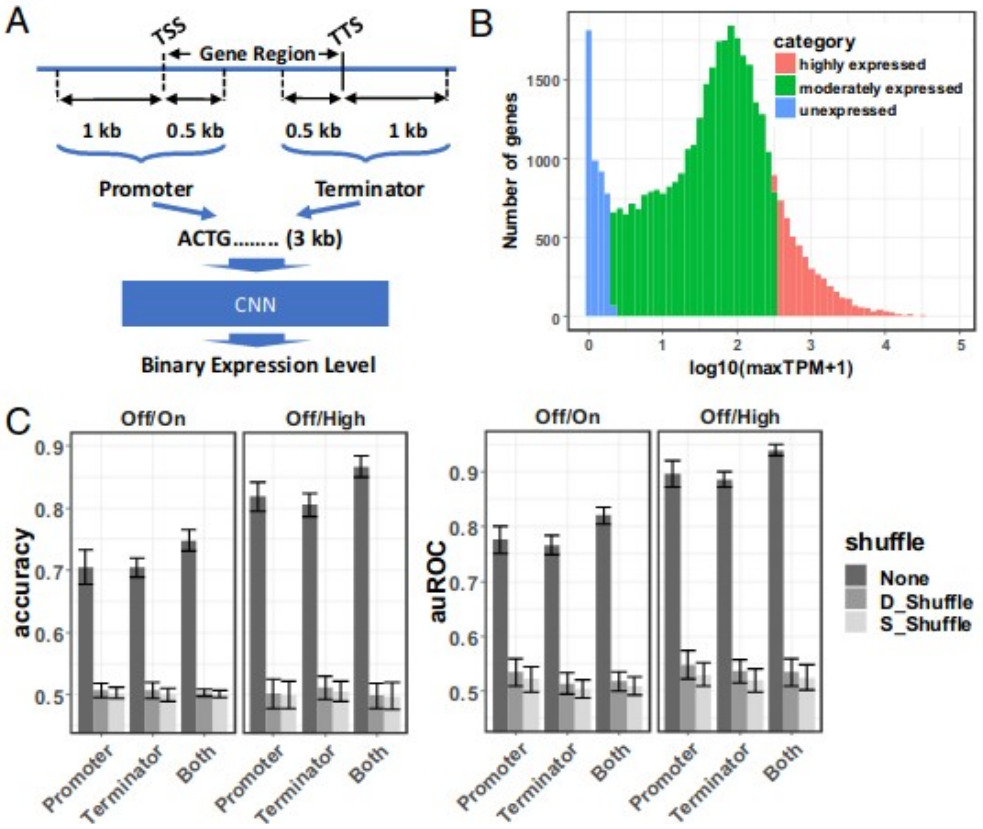
G x E



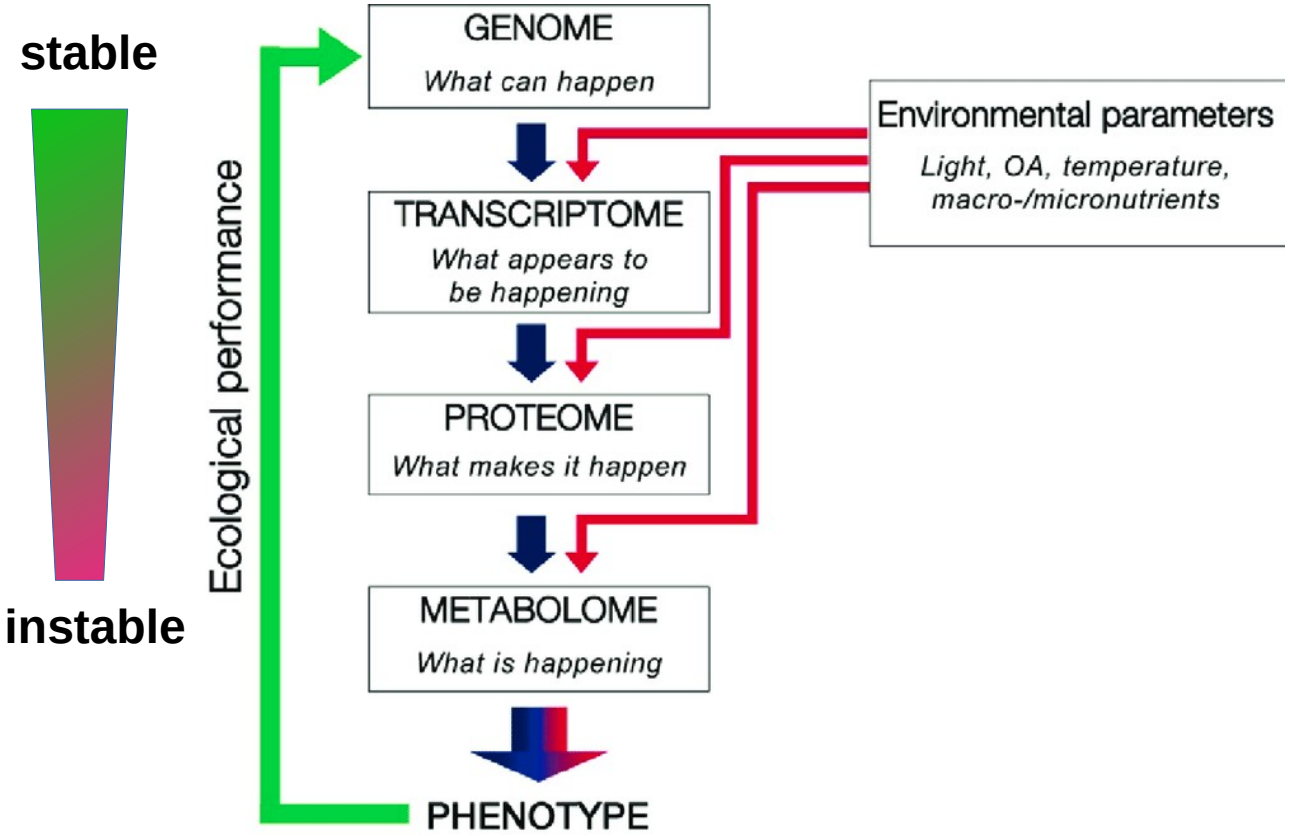
Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence

Jacob D. Washburn[a,1], Maria Katherine Mejia-Guerra[a], Guillaume Ramstein[a], Karl A. Kremling[a], Ravi Valluru[a], Edward S. Buckler[a,b,2], and Hai Wang[c,a,1,2]

[a]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853; [b]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY 14850; and [c]Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, 100081 Beijing, China

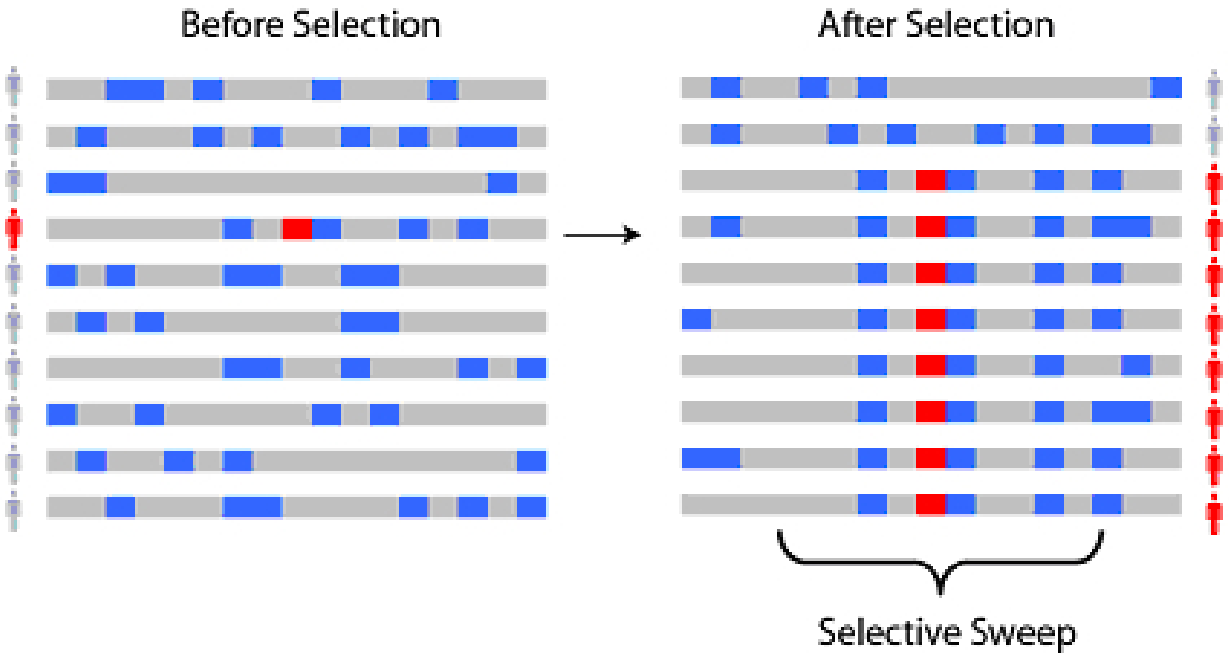


Selection shapes the genome



Dettmer et al. (2007) Mass Spect Rev

Selective Sweeps



Beneficial variants increase their frequency
The same holds true for neutral variants in their proximity

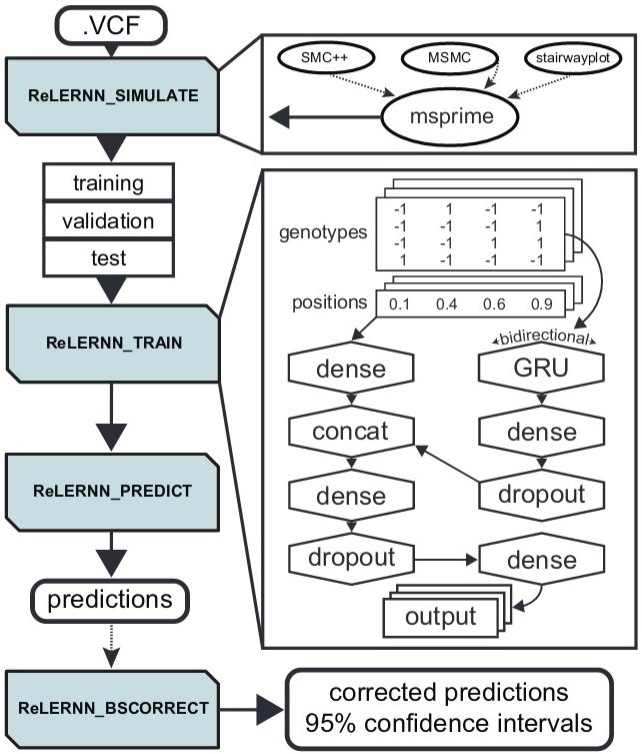
Schaffner and Sabatei (2008) Nature Education

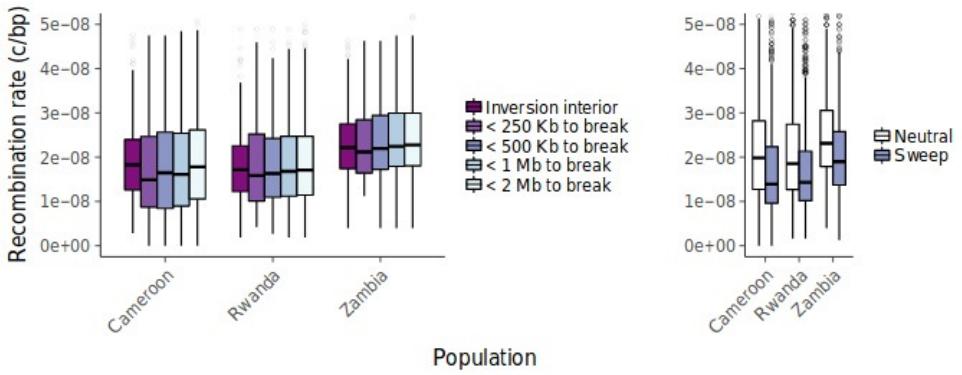
Inferring the landscape of recombination using recurrent neural networks

Jeffrey R. Adrion[1,†], **Jared G. Galloway**[1,†], **Andrew D. Kern**[1]

[1]Institute of Ecology and Evolution, University of Oregon

***For correspondence:**
jadrion@uoregon.edu



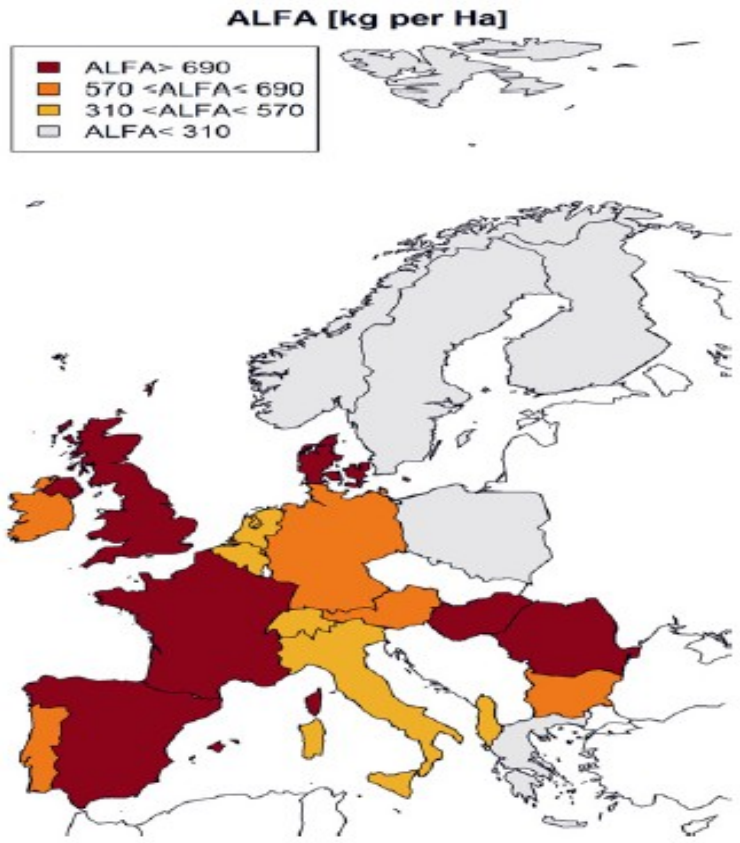


Summary ML in Genomics

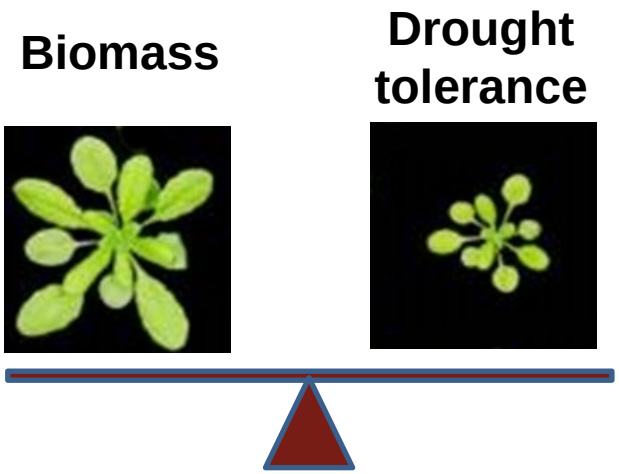
It still feels a bit like the Wild West
but we are trying...

- **Genome assembly**
- **Imputation**
- **Detection of structural variants**
- **GWAS**
- **Genomic Prediction**
- **Predict gene expression**
- **Predict recombination rate and selective sweeps**
- **joint models including different data layers**

Application to the real world

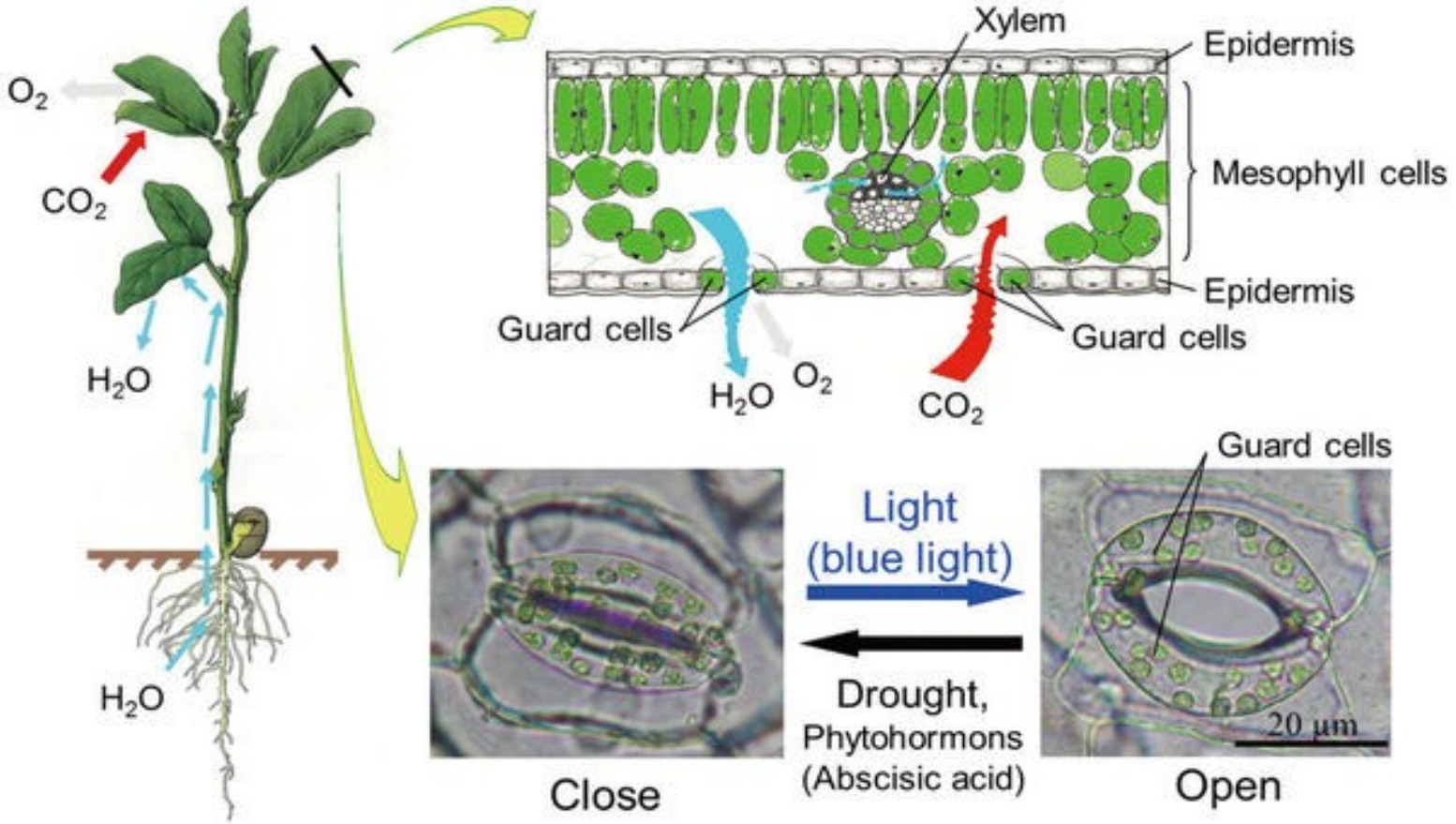


Naumann *et al.* Env. Research Letters **2015**



Drought = less biomass production

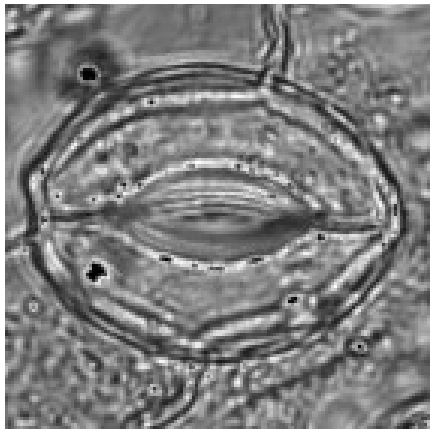
Guard Cells regulate plant water status

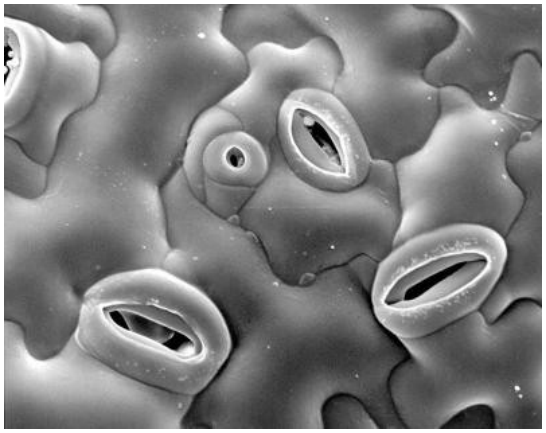


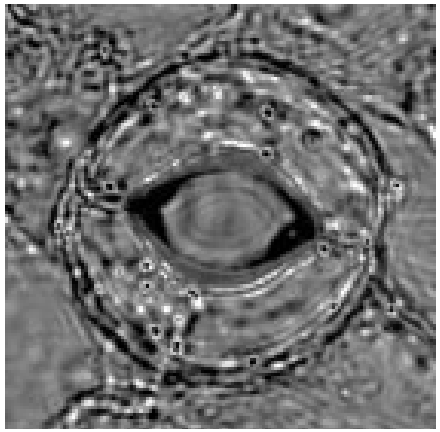
Natural variation in guard cell signalling

Low humidity
High CO_2
Darkness

Heat
Low CO_2
Light





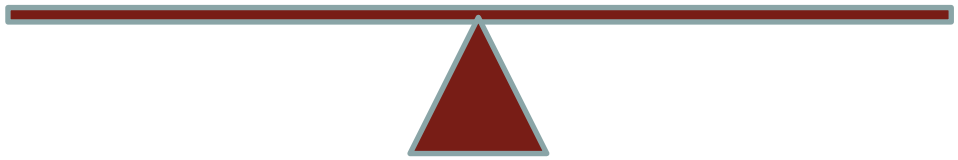


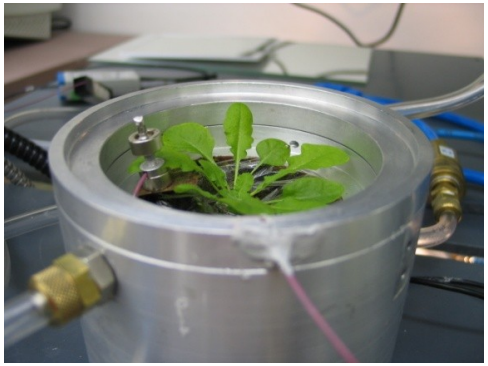
Guard cells directly sense environmental cues and balance water loss and CO_2 uptake

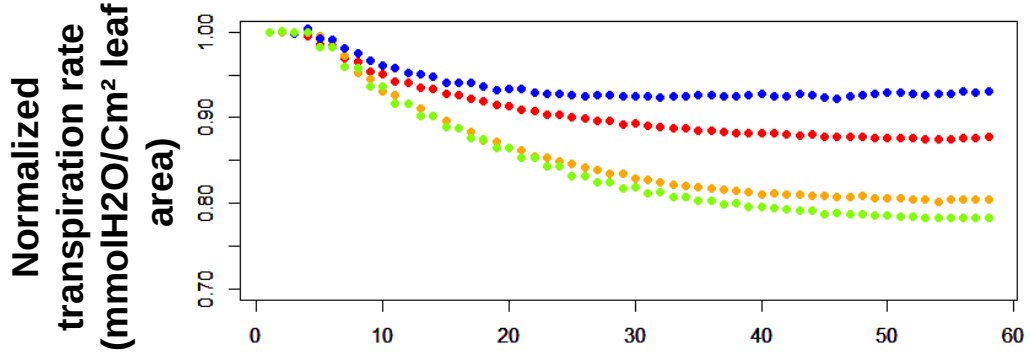
Natural variation in guard cell signalling

Drought tolerance

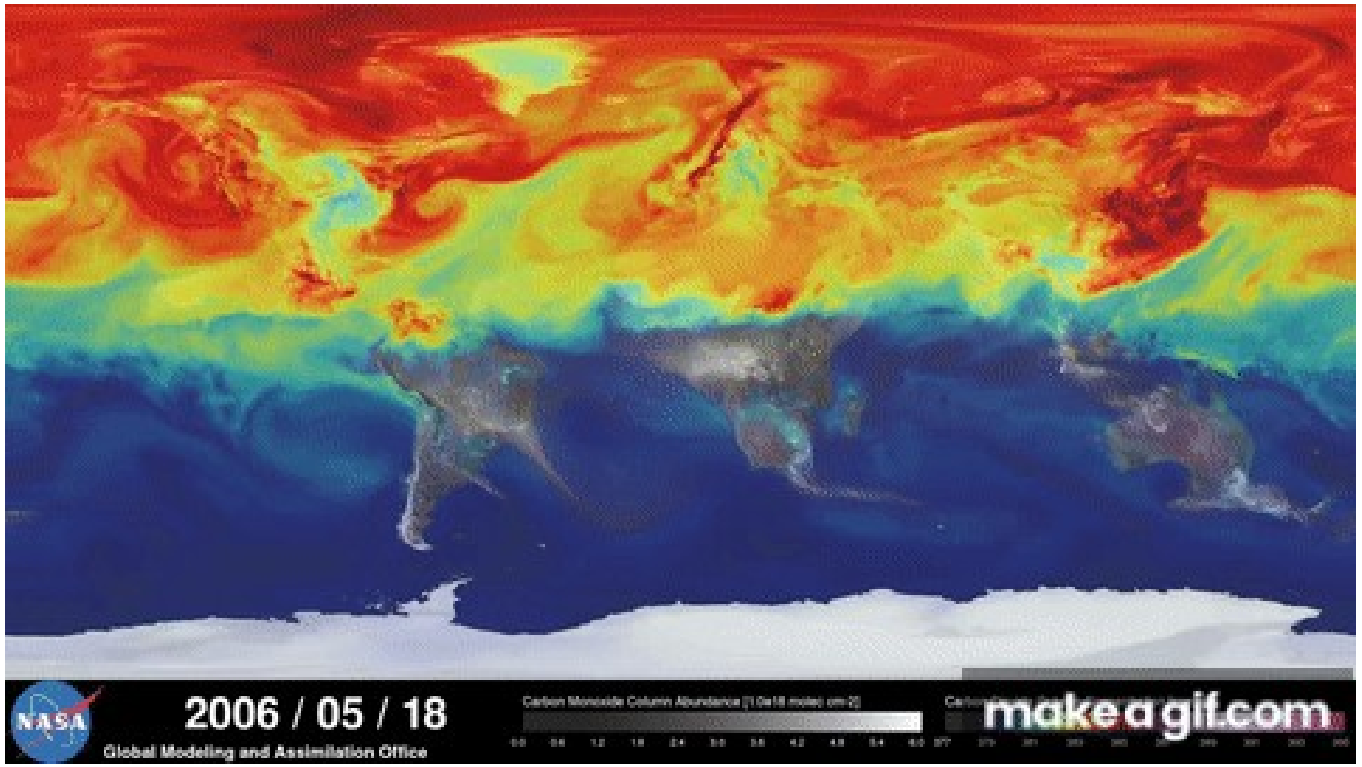
Biomass







CO_2 distribution in the atmosphere

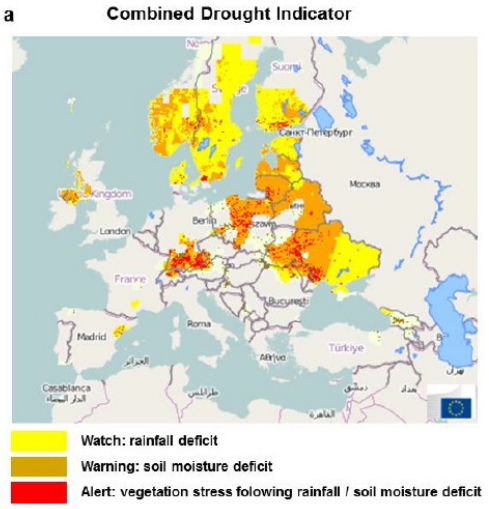


Understanding genotype – phenotype relationships and how the environment influences these

Statistical Genetics

$$Y = \beta_0 + \beta_1 X + u + \varepsilon$$

Drought stress



Integration of different data layers into joint models

Acknowledgments

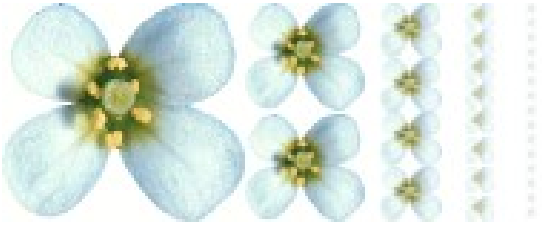
Uni Würzburg
Jan Freudenthal
Ammarah Anwar
Willia Lopez

GMI Vienna
Magnus Nordborg
Ümit Seren

VIB Ghent
Dirk Inze
Frederik Coppens
Pieter Clauw

Uni Köln
Juliette DeMeaux
Hannes Dittbrenner

Center for Computational and Theoretical Biology



The 1001 Genomes

MAZE

SPONSORED BY THE
Federal Ministry of Education and Research