

T4.2 Genomic Selection and its potential in organic breeding

MODULE 4 -Development and application of molecular methods in organic breeding

Training courses in Organic Breeding

04.03.2025

Michael Schneider, FiBL

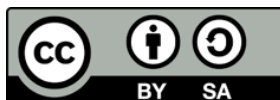
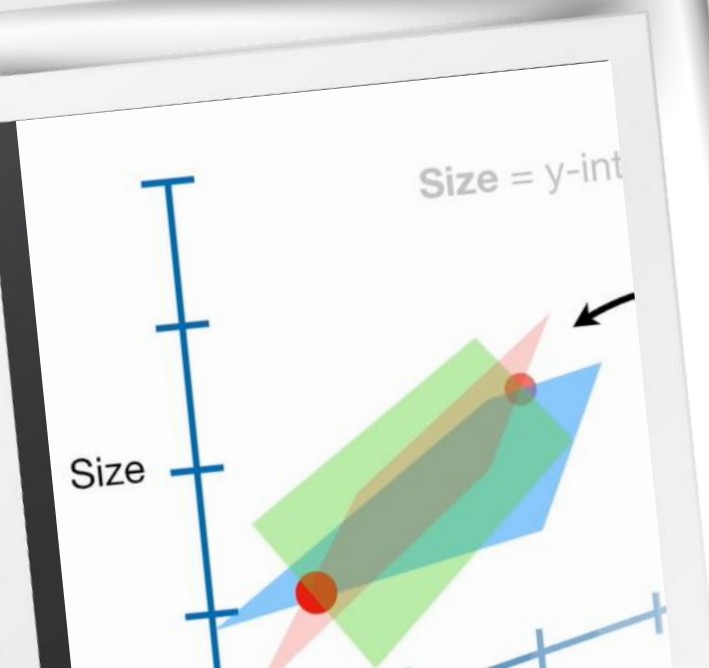
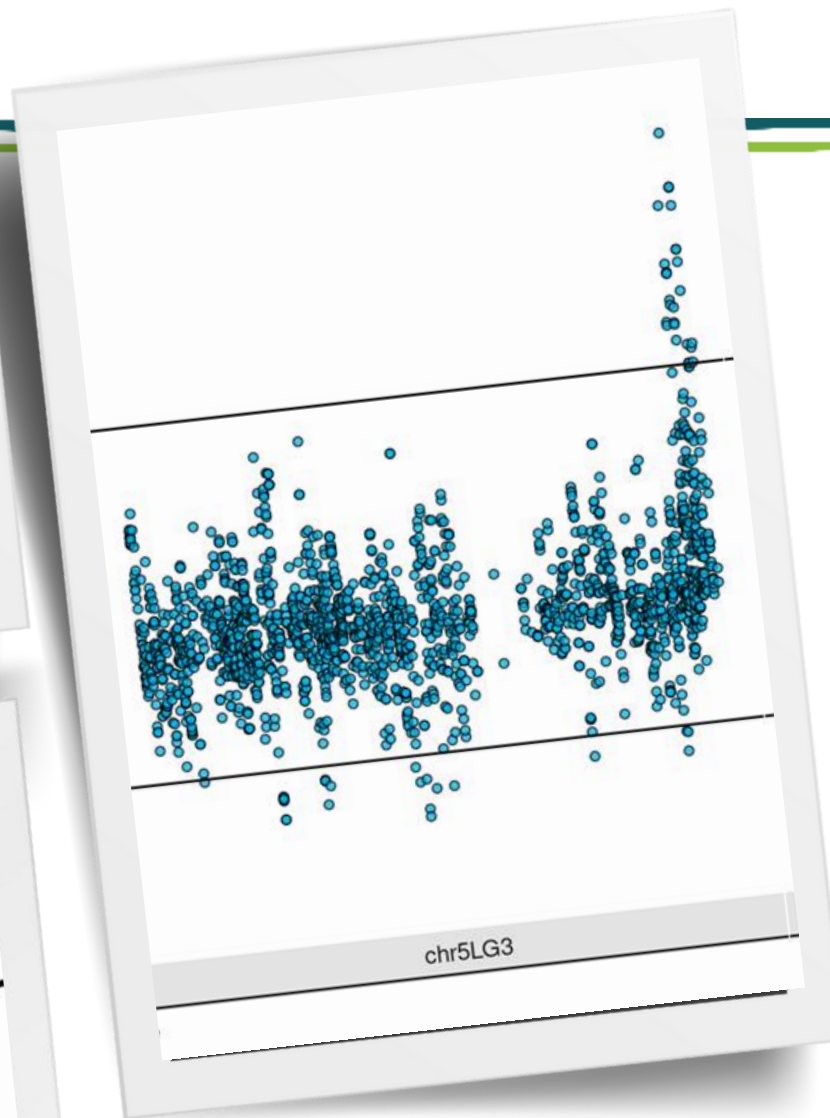


Training in organic breeding

**Module 4: Development and application of molecular
methods in organic breeding**

**Unit 4.2: Genomic Selection and its potential in
organic breeding**

Michael Schneider, FiBL



Funded by the European Union, the Swiss State Secretariat for Education, Research and Innovation (SERI) and UK Research and Innovation (UKRI).



UK Research
and Innovation

Training in organic breeding organized in 5 Modules

1. **Module 1 - Plant Genetic Resources (PGRs): collection, conservation and exchange to support the increase of agrobiodiversity in farming systems**
2. **Module 2 - Phenomics: approaches and tools for genetic resources and breeding material characterisation - FEBRUARY 3rd 2025, 9:00 to 17:30 CET**
3. **Module 3 - Breeding methods fundamentals - FEBRUARY 13th 2025, 9:00 to 18:00 CET**
4. **Module 4 - Development and application of molecular methods in organic breeding - MARCH 4th 2025, 9:00 to 18:00 CET**
5. **Module 5 - Organic heterogeneous material (OHM) design and development - MARCH 7th 2025, 9:00 to 18:00 CET**

Planned for today

- **Genomic prediction – what for?**
 - Introduction – what is the use?
 - Mathematical background
 - Model fit and validations
- **Get into the code – applied programming**
 - Checking out different R packages
- **Recap**
- **Exercise – completion important for the CERTIFICATE**

Aim of breeders and scientists

Making predictions of a phenotype or trait, prior trait measurement

- **Examples**

- Milk quantity of cattle
- Plant health status in a new environment
- Seed yield
- .. (these are called complex traits, because multiple genes contribute to the phenotype)

How can you gain this information (without knowledge of the genotype)?

- **By calculating «BLUP's» (Best linear unbiased prediction)**

- Using additional information of relatness of genotypes (Pedigree)
- Testing (many) related genotypes in multiple environments / replications
- This is summarized in the EBV (estimated breeding value)
- = value can be defined as its genetic merit for each trait

Including genomic data can improve those predictions

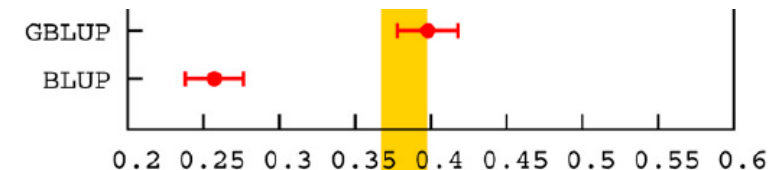
Limitations of EBV

Using only pedigree information..

- is not necessarily correct for all loci (recombination)
- High variants of parents lead to high prediction error
- Predicted individual must be related to known founders
- Requires a lot of phenotyping
- Adding genomic data => GEBV / MBV



Predict phenotype from DNA

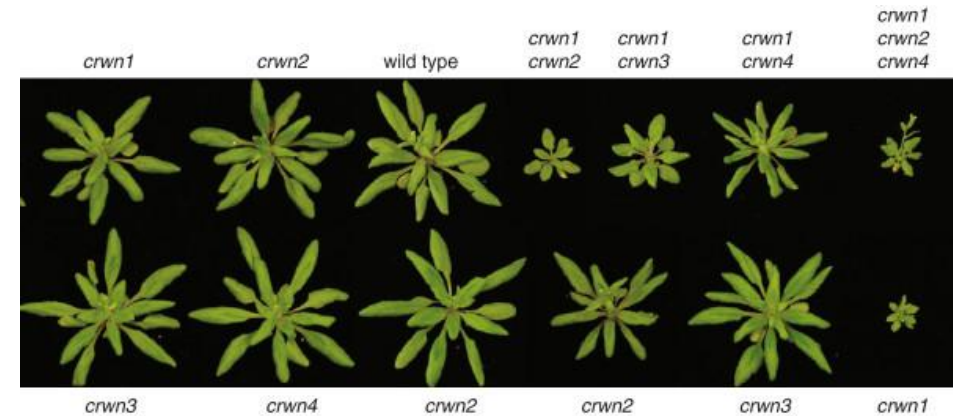


Complex traits – the genetic background

Phenotype = Genotype + Environment

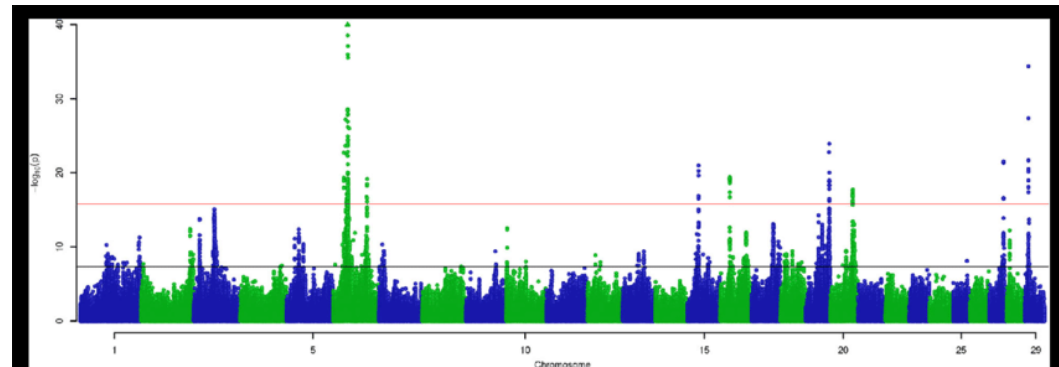
variance of phenotype traits are determined by *multiple genetic loci*, described by different effect sizes

1. **Few genes have big effects**
2. **Many genes have no effect**
3. **Many genes have very small effects**
 - can be infinitely small



[10.1186/1471-2229-13-200](https://doi.org/10.1186/1471-2229-13-200)

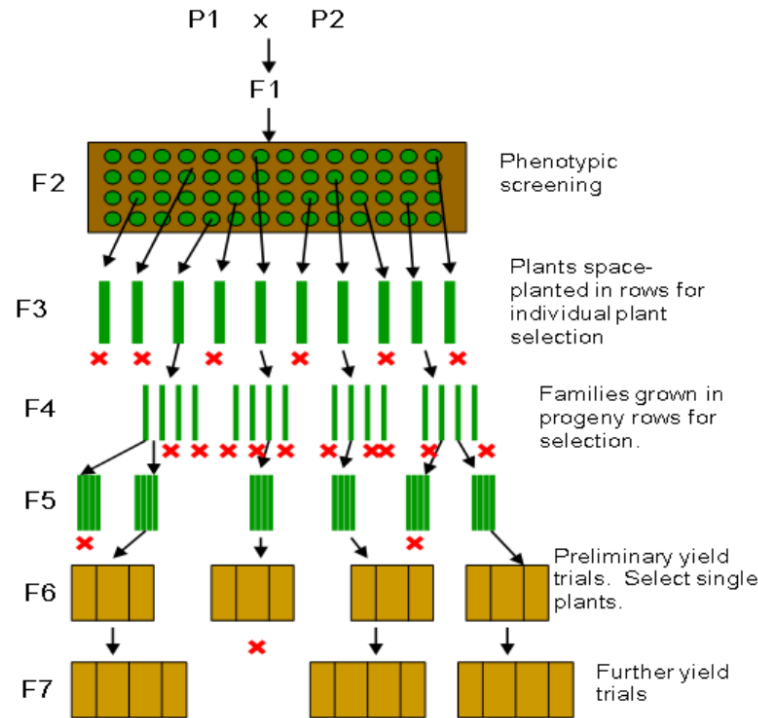
GWAS – genome wide association study
find few loci with big effects



[10.1186/s12864-017-4320-3](https://doi.org/10.1186/s12864-017-4320-3)

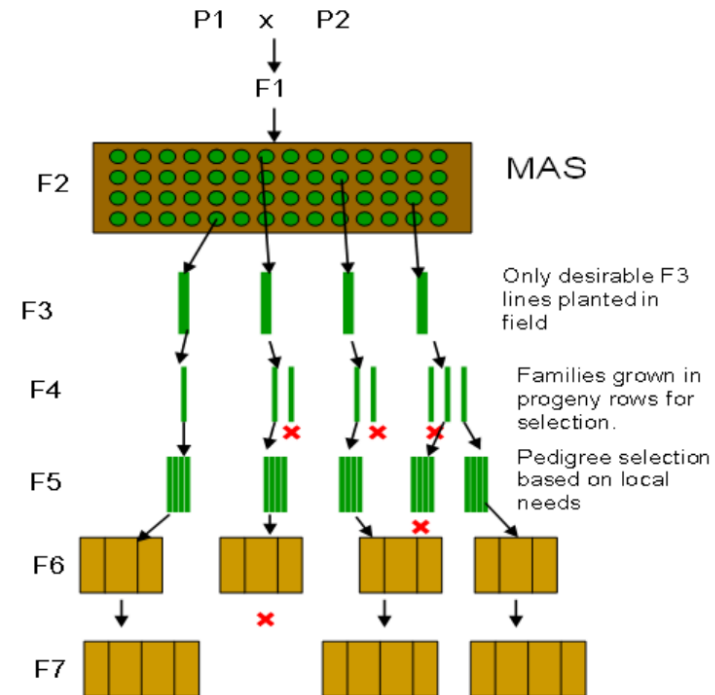
Using these QTLs for marker assisted selection (MAS)

PEDIGREE METHOD



F8 – F12 Multi-location testing, licensing, seed increase and cultivar release

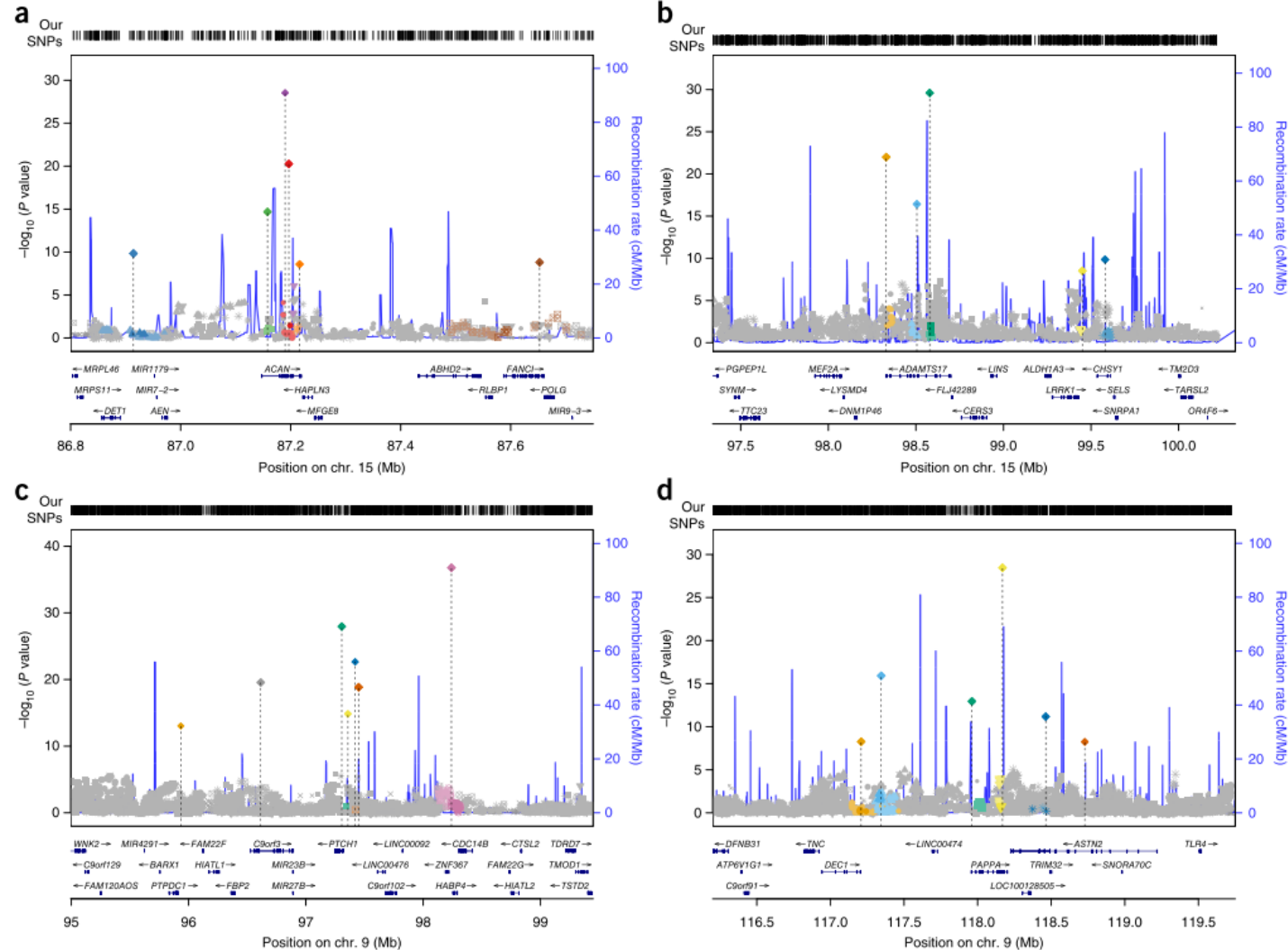
EARLY GENERATION SELECTION MARKER ASSISTED SELECTION



F8 – F12 Multi-location testing, licensing, seed increase and cultivar release

<http://dx.doi.org/10.19080/ARTOAJ.2018.14.555914>

GWAS – find few loci with big effects



What is the take?

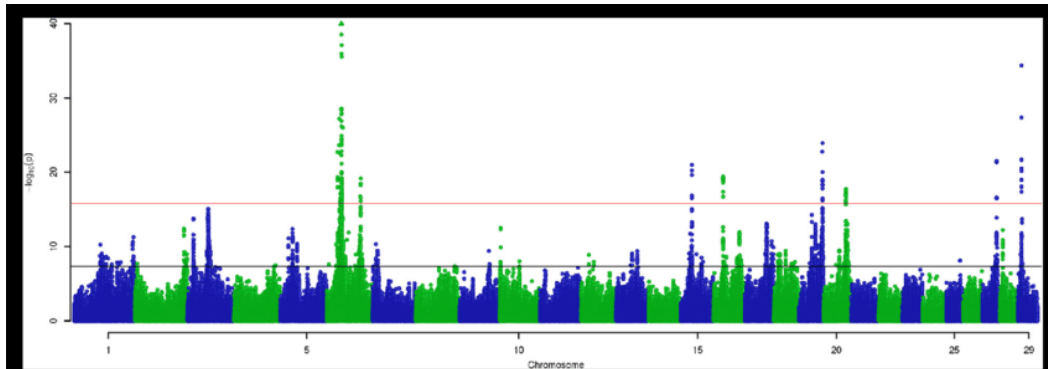
- Although there were few applications in (cattle) breeding, MAS based on a few markers was not contributing appreciably to livestock improvement simply because most of the traits of interest are quantitative and complex, meaning *phenotypes are determined by thousands of genes with small effects* and influenced by environmental factors.
- only a few genes that contribute more than 1% of the genetic variation for any given polygenic trait

Approaches to link genomic information to a trait

GWAS / QTL mapping

- Find few loci with big effects

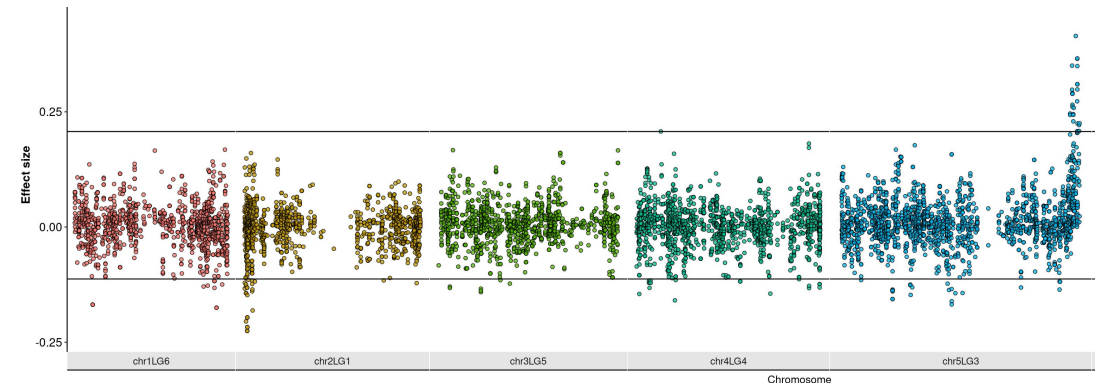
- most of the variation is not considered
- Needs few markers to make predictions



▪ Genomic prediction

- All loci have an effect on the phenotype

- all the variation is covered when using all markers
- Many more markers are needed



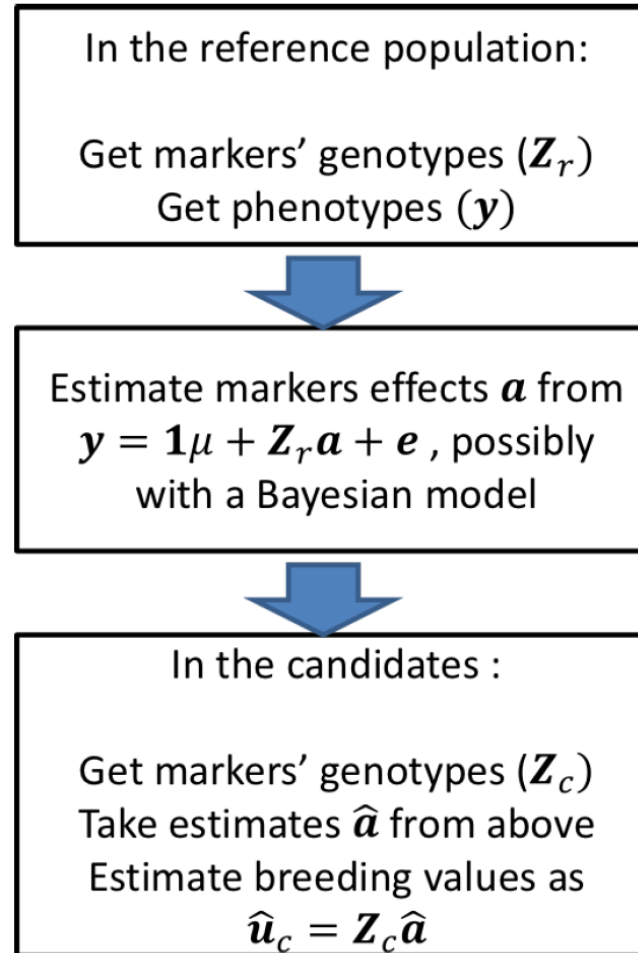
Key message



Genomic prediction can be an (alternative) approach to boost marker-assisted selection and with that, improve the breeding success

Biological background: Need and use:

So, how does the genomic prediction works?



Backbone of genomic prediction – the maths

$$y = Xb + Za + e$$

Breeding value = $u = Za$

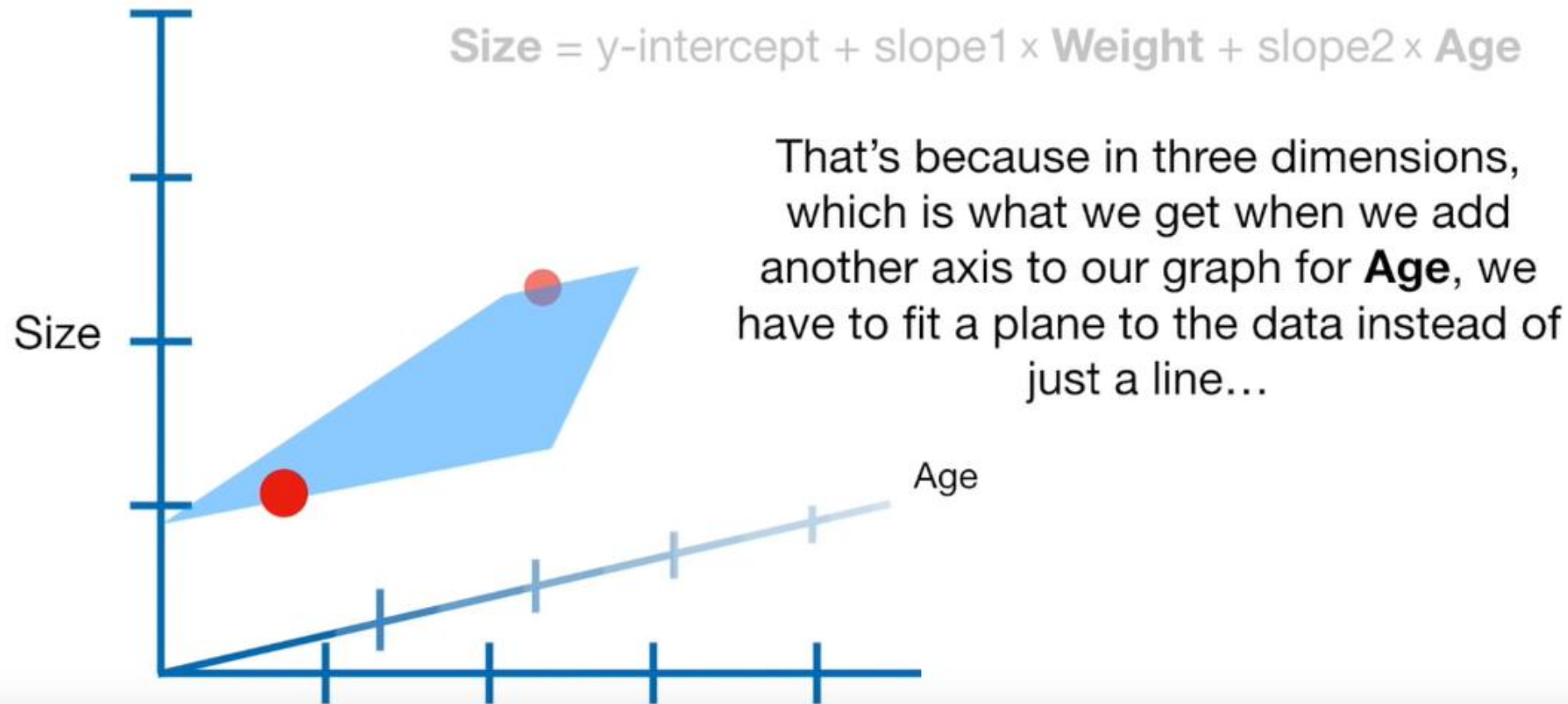
$$Za = \begin{pmatrix} 0 & 1 & 1 & 0 & \vdots & 2 & 0 \\ 2 & 0 & 0 & 0 & \vdots & 1 & 1 \\ 0 & 1 & 0 & 1 & \vdots & 0 & 2 \end{pmatrix} \begin{pmatrix} a_A \\ a_B \\ a_C \\ a_D \\ \dots \\ a_E \\ a_F \end{pmatrix}$$



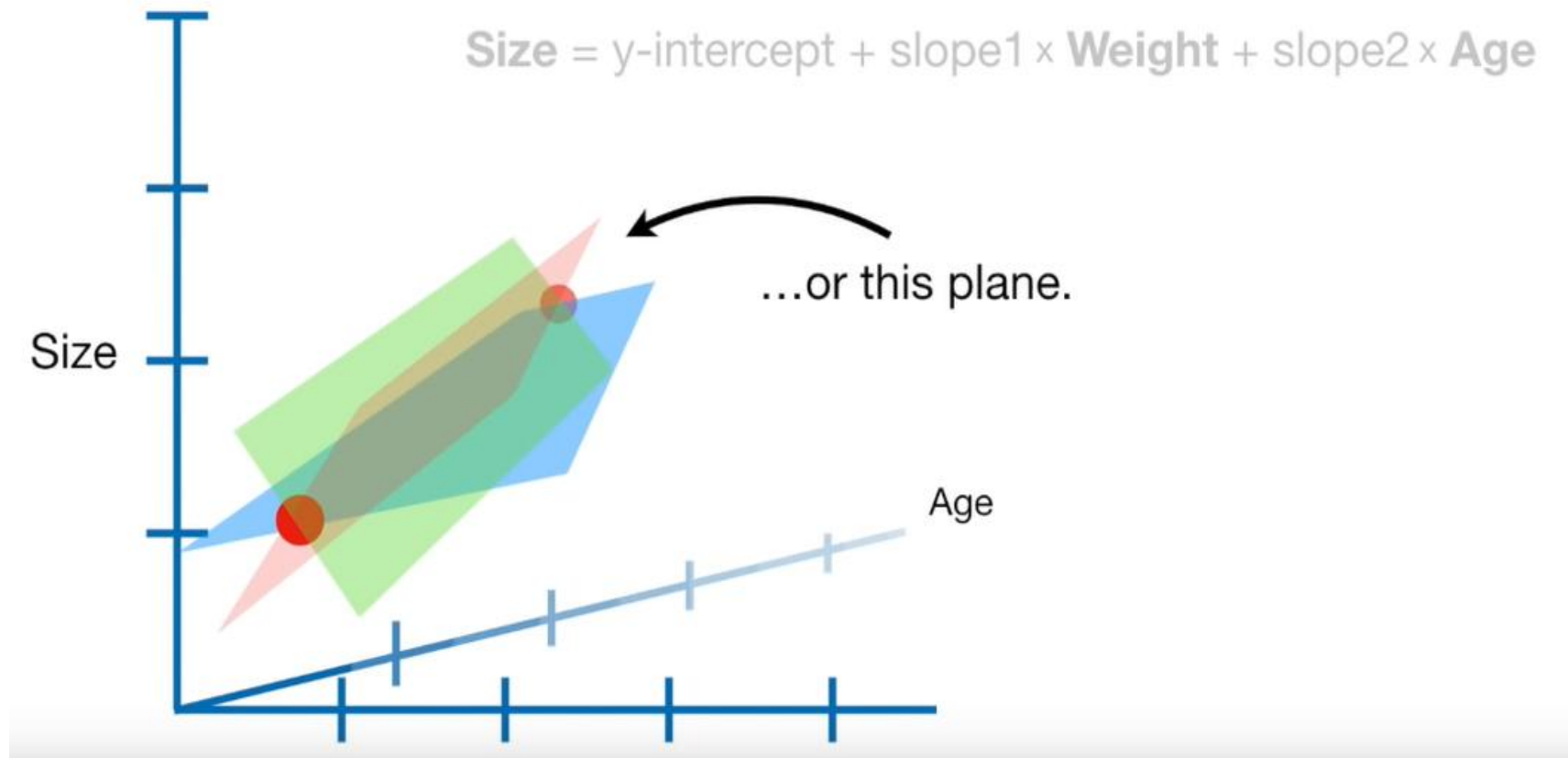
Marker effect

Excuse to the maths part of the problem..

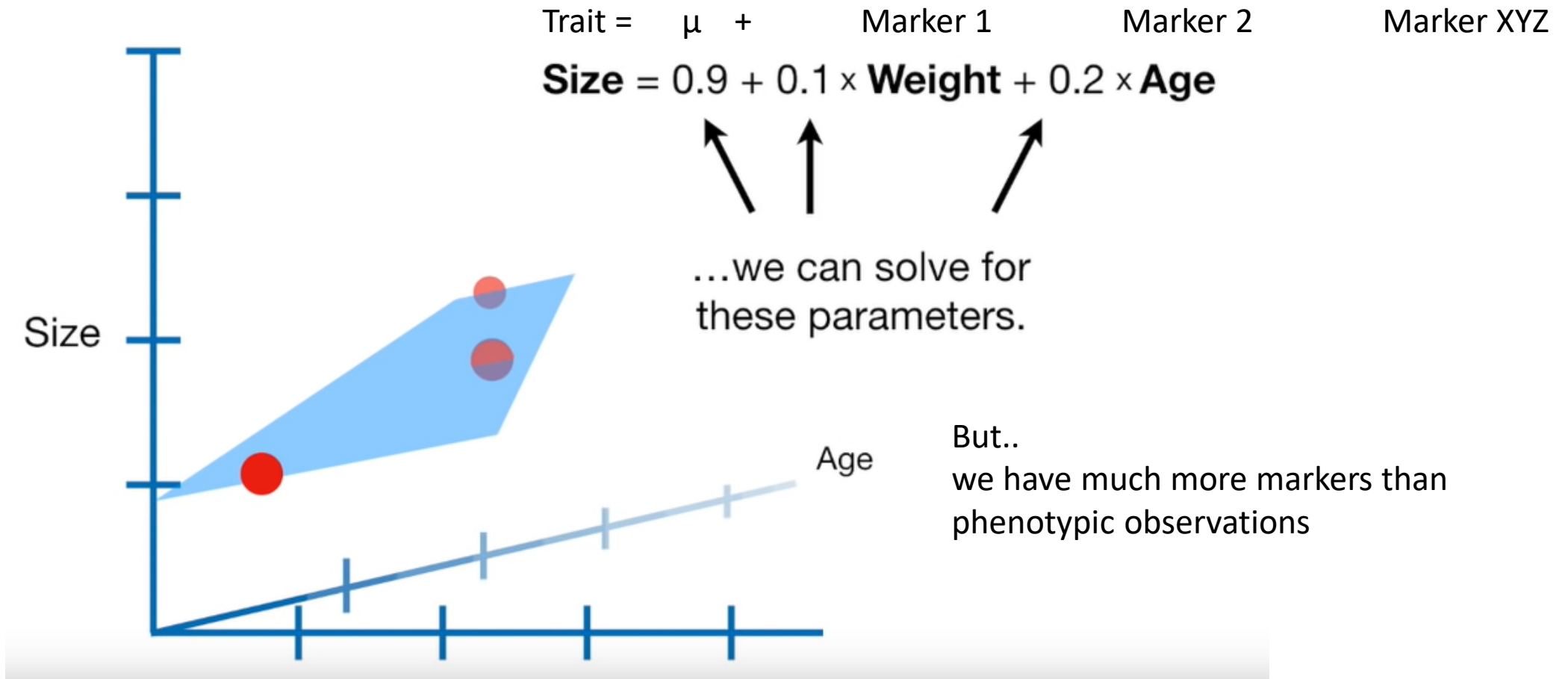
We have many SNPs (predictors) and few phenotypes (measurements), which causes a problem



With simple linear regression, we cannot solve the missing terms in the equation



With more y values than x (& z axis), we could solve the equation



But we do not have.

- **More predictors than target values available**

- **Problem in solving the equations**

- a simple linear model
 - with fixed effects does not work

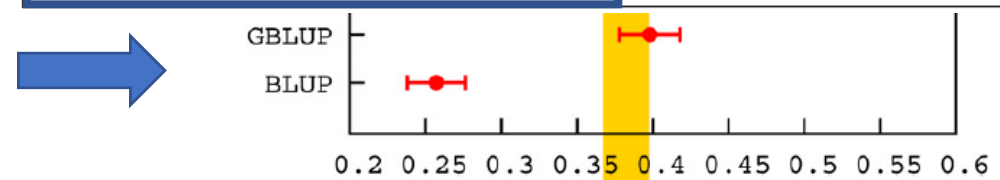
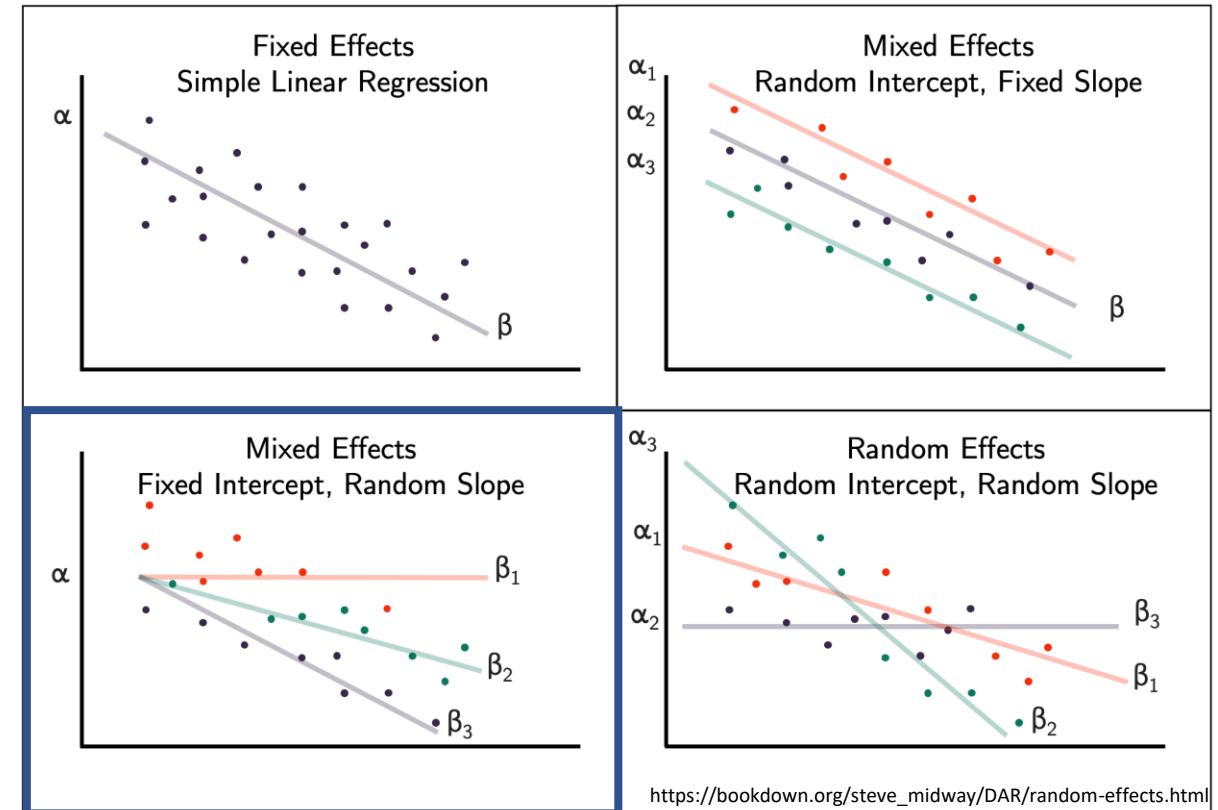
- **We can solve this problem by**

- **Considering all SNPs as *random effects***

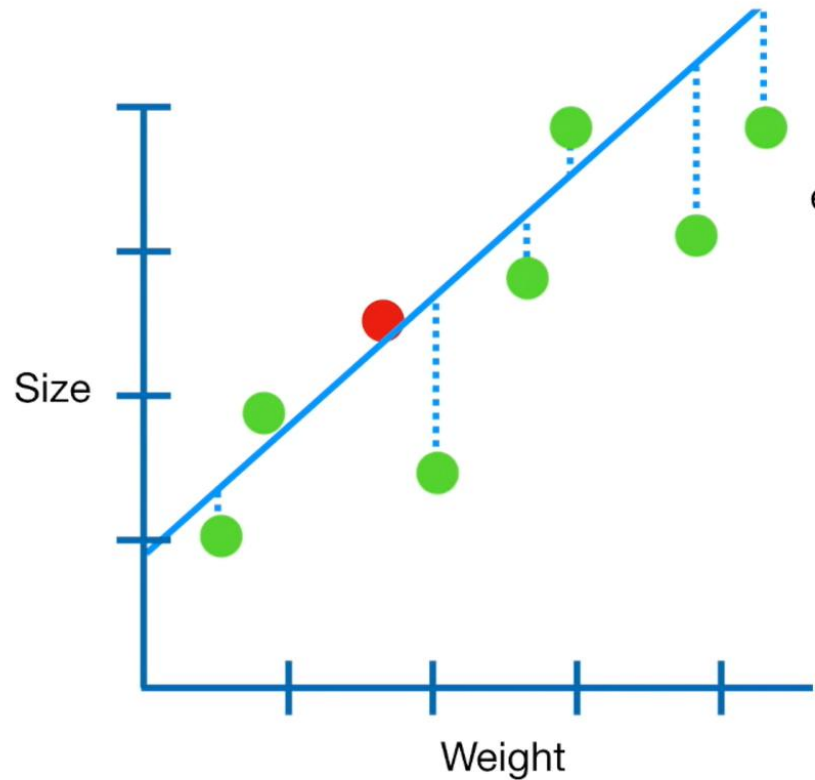
- this way, all effects can be jointly estimated

- **An alternative idea is to calculate the genetic distance between genotypes from markers**

- Use this information as predictor for the phenotypes (called GBLUP)



Solving it with priors (bayesean) or shrinkage (ridge, lasso, elastic net)



Lastly, even when there isn't enough data to find the **Least Squares** parameter estimates, **Ridge Regression** can still find a solution with **Cross Validation** and the **Ridge Regression Penalty**.



the sum of the squared residuals

+

$$\lambda \times \text{Slope}^2$$

3 approaches exist to shrink the regression – shrinkage = reduction of errors

- ***Ridge regression***

- Sum of squared residuals + $\lambda * \text{slope}^2$
 - $\lambda * \text{slope}^2 = \text{penalty}$
 - will lead to the regression becoming less steep as λ becomes bigger
 - How to determine the best λ ? By cross validation
 - The slope attribute contains all parameters despite the intercept (so, all markers)
 - Not all parameters are shrunk equally
 - Ridge Regression assumes that effects are a priori normally distributed

- ***Lasso (least absolute shrinkage and selection operator)***

- Sum of squared residuals + $\lambda * |\text{slope}|$
- Many similar attributes to ridge regression
- Differences:
 - Lasso can shrink the slope all way to 0
 - => Meaningless variables can be eliminated as terms in the equations.
- Lasso can exclude useless variables from the equation
 - Ridge can only minimize their effect
- Lasso assumes that (marker) effects are a priori distributed following a Laplace (double exponential) distribution

- ***Elastic net***

- a combination of both lasso and ridge regression



▪ **LiveSeeding** λ & slope are both estimated using prior information, so that the marker information is utilized as a conditional probability, which depends on priory gained or assumed knowledge, e.g. a marker should not have an effect of, one phenotypic standard deviation of the trait

Bayes way – using a prior information

1. Bayes's Theorem

- $p(A\&B|B) = \frac{p(A\&B|A) p(A)}{p(B)}$
- Where A is unknown (can be a parameter) and B is known (can be a trait phenotype). Therefore, we want to infer values A by knowing B .
 - $p(A\&B|B)$: posterior probability of unknown A given B is known.
 - $p(B\&A|A)$: likelihood function, determined by both A and B .
 - $p(A)$: prior probability of unknown A .
 - $p(B)$: probability to observe B without having any knowledge of A .

2. Bayesian methods are non-linear and likely to be affected by shrinkage, meaning small effects became even smaller and big effects even bigger.

3. Bayesian regressions are affected by the prior distribution that we assign to marker effects

- each marker has a priori a different variance

$$p(a_i | \sigma_{ai}^2) = N(0, \sigma_{ai}^2)$$

Back to the practical application..

Which approaches exist?

1. **Genomic relationship-based method (GBLUP)**
2. **SNP effect-based method (SNP-BLUP)**
 - RR-BLUP – random/ridge regression-BLUP
 - BayesA
 - BayesB
 - BayesC
 - Bayes Lasso
 - ...

GBLUP method

- **Based on genomic relationships - identical by state (IBS)**
- **quantifying the number of alleles shared between two individuals**
- **genomic relation ship can be calculated for**
 - additive ..
 - dominance ..
 - and epistasis effects
- **the genomic relationship is given in a $n \times n$ matrix, where n denotes the genotypes objected to study – this matrix is also known as additive relationship matrix**
 - multiple ways exist to calculate this matrix
- **instead of SNP effects (u), genomic breeding values (Za) are estimated**
- **GBLUP is a BLUP where the pedigree relationship matrix is replaced by the genomic relationship matrix G.**
 - G contains only information from genotypes individuals.

Ridge regression BLUP

•RR-BLUP or SNP-BLUP provides SNP effects, but genomic estimated breeding values (u) can be derived as linear combinations of the SNP effects:

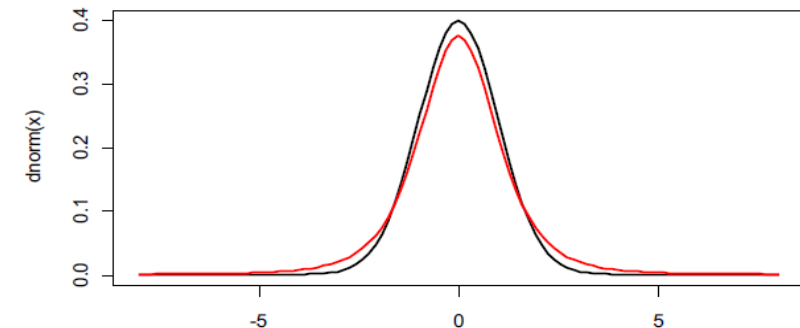
- $u = Za$ (Z marker matrix, a allele effect)
- marker effect follows a priori a normal distribution with a variance σ_{a0}^2 (variance of marker effects)
- markers are independent one from each other
 - the prior assumption of normality precludes few markers of having very large effect

All markers with the same variance

$$p(a_i) = N(0, \sigma_{a0}^2)$$

Bayes variance prior

$$p(a_i | \sigma_{ai}^2) = N(0, \sigma_{ai}^2)$$



Bayesian approaches

- **BayesA**

- All SNPs have an effect on the trait
 - Few have a large effect, most have a small effect
 - => different variances are assumed

- **BayesB**

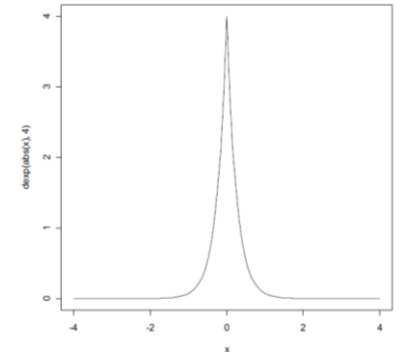
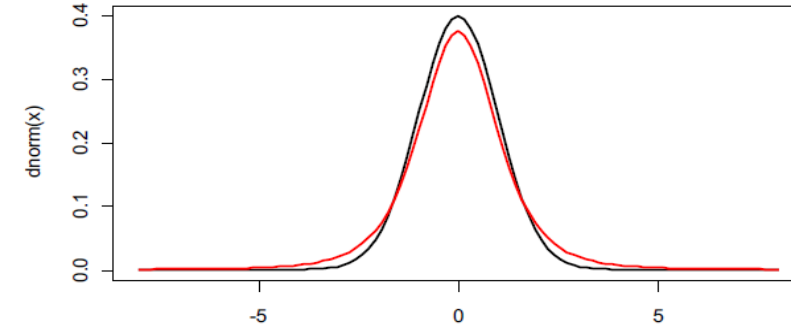
- Not many QTL were effecting the traits => many loci have zero variance
- π = proportion of SNP have no effect
- $1 - \pi$ = have a non-zero effect
 - When $\pi = 0$, BayesB becomes BayesA

- **BayesC**

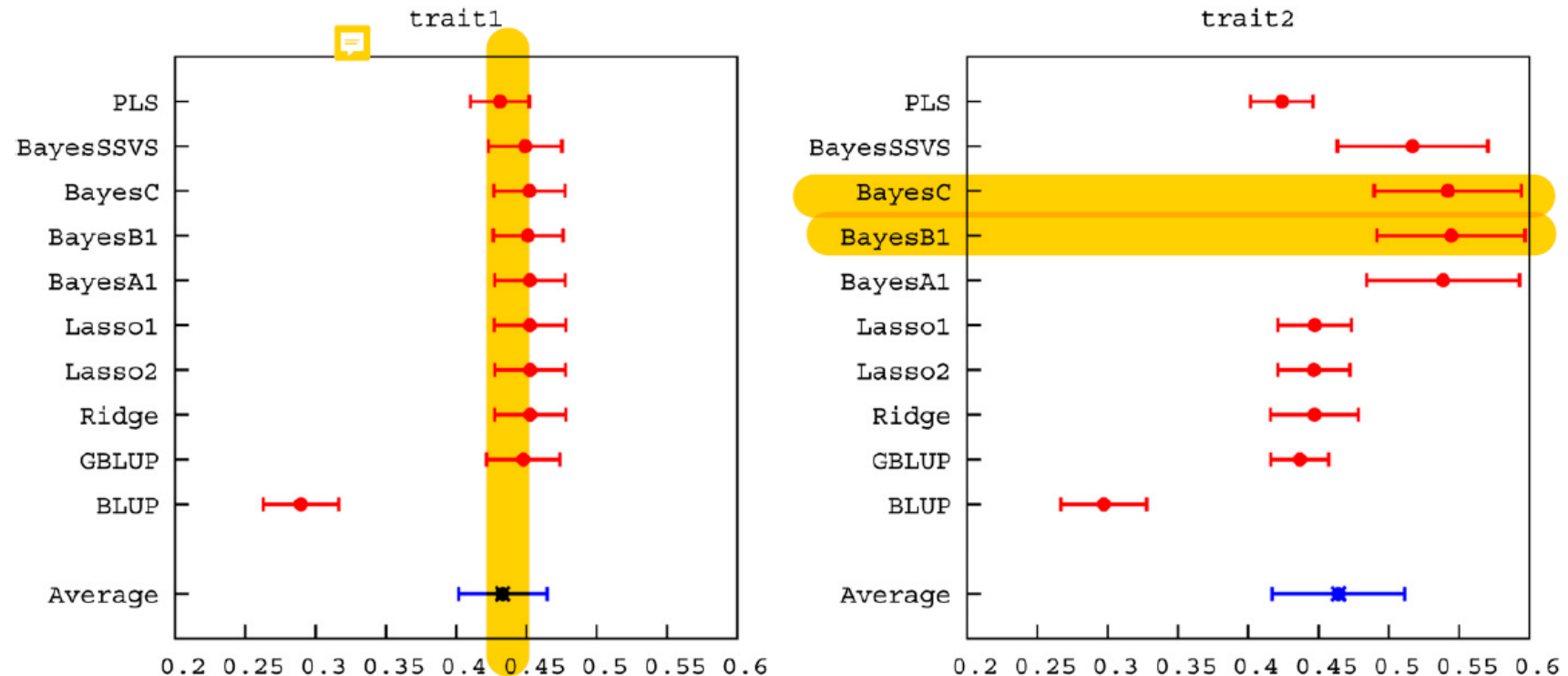
- A combination of SNP-BLUP and BayesB
- Combines a distribution with constant variance (SNP-BLUP) and assumes some fraction π of SNP have no effect (BayesB)
- If $\pi = 0$, BayesC = SNP-BLUP

- **BayseanLasso**

- Sets marker values a prior to small values, instead setting some to 0
- This is very similar to BayesA, in that a prior distribution is postulated for marker variances. The difference is the nature of this prior distribution (exponential in Bayesian Lasso and inverted chi-squared in BayesA),



How do these differentiate and perform anyway?



Why not simply using machine learning to predict phenotypes?

GBM was best, while in simpler cases lasso was superior. In the wheat and rice studies the best two methods were SVM and BLUP. The most robust method in the presence of noise, missing data, etc. was random forests. The classical statistical genetics method of genomic BLUP was found to perform well on problems where there was population structure. This suggests that standard machine learning methods need to be refined to include population structure information when this is present. We conclude that the application of machine learning methods to phenotype prediction problems holds great promise, but that determining which methods is likely to perform well on any given problem is elusive and non-trivial.

Table 2 cvR^2 for the five ML methods and for BLUP across 10 resampling runs applied to the wheat dataset. The best performance for each trait is in boldface. The average ranks for computation of the Friedman test are on the bottom line

Trait/method	Lasso	Ridge	BLUP	GBM	RF	SVM
Yield (drought)	0.023	0.060	0.217	0.051	0.172	0.219
Yield (irrigated)	0.084	0.162	0.253	0.132	0.184	0.258
TKW	0.172	0.240	0.277	0.218	0.242	0.304
DTH	0.292	0.325	0.381	0.325	0.358	0.394
Average rank	6.00	4.12	2.00	4.88	3.00	1.00



How is the performance of the prediction model assessed?

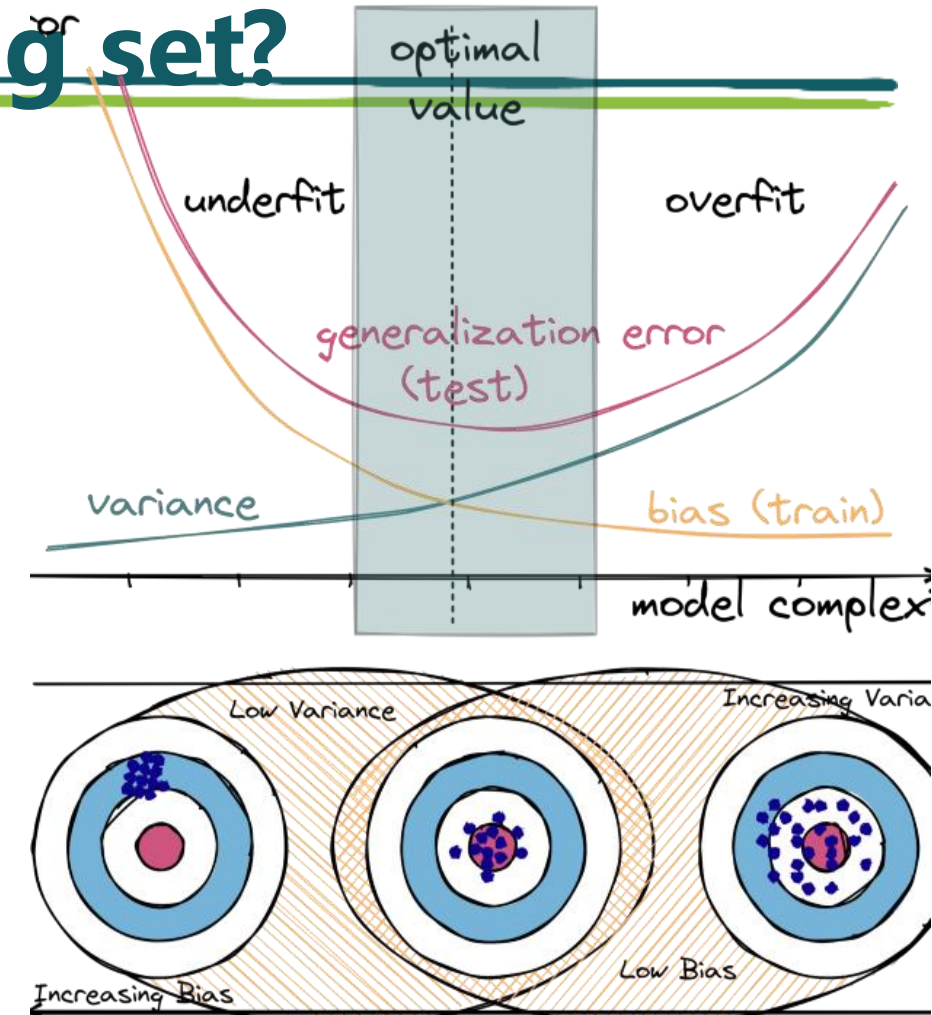
Validation works like..

	Complete data				
P1	Val	Train	Train	Train	Train
P2	Train	Val	Train	Train	Train
P3	Train	Train	Val	Train	Train
P4	Train	Train	Train	Val	Train
P5	Train	Train	Train	Train	Val

- We split the population of individuals in two groups – training and testing
- **K-fold cross-validation**
 - In this method, the genotyped population **is randomly divided into k subsets**, and phenotypes are removed from one subset a time
 - **Predictivity** for each fold is calculated by a **Pearson correlation**
 - The final prediction ability across all k-folds is calculated by summing up the correlation scores and dividing them by the number of k-folds.
 - *partitioning of training and testing populations will affect the accuracy attained*
 - large testing sets will reduce the reference population size and reduce accuracy
- when the testing set is too small, assessing differences in accuracy between methods for a particular data set may not be possible

Why do we need to split in a training and testing set?

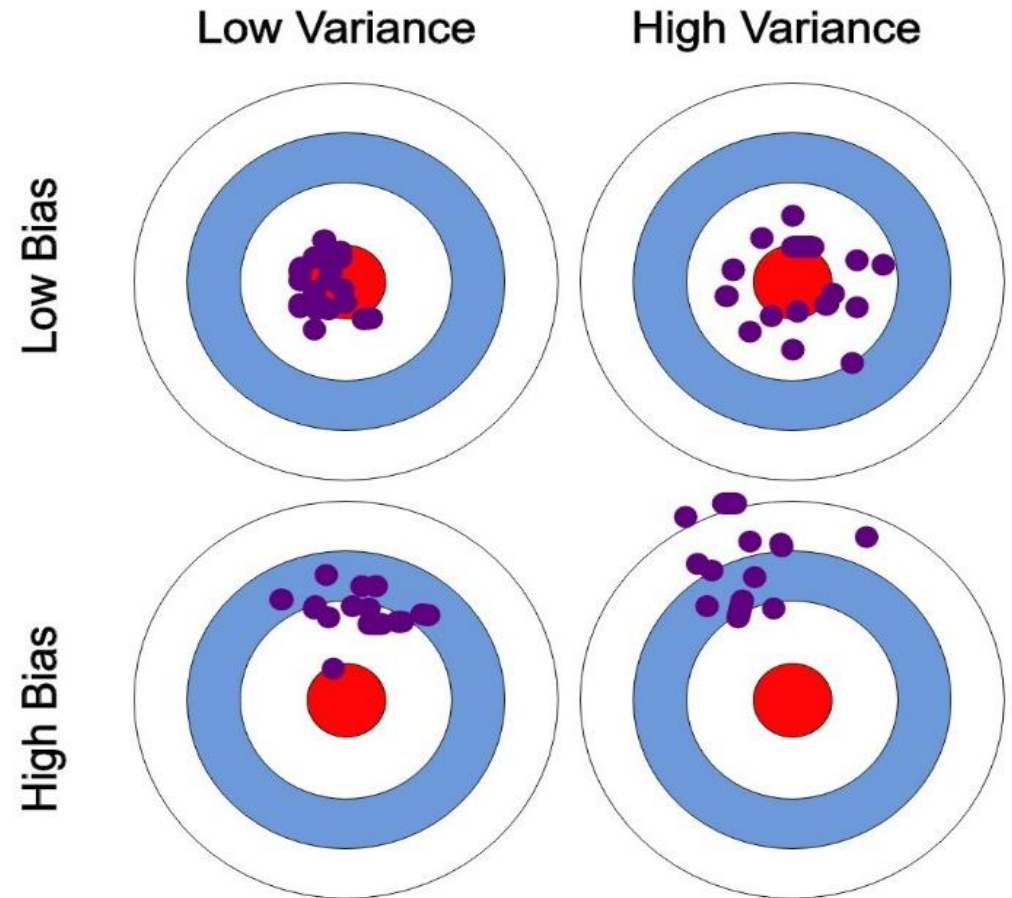
- *Bias-Variance Trade-off*
- "The two variables to measure the effectiveness of your model are bias and variance."
- **Bias** is the error or difference between points given and points plotted on the line in your training set.
 - (Sums of squares of the points to the regression lines in the training data set)
- **Variance** is the error that occurs due to sensitivity to small changes in the training set
 - (Sums of squares of the trained regression to the test-set points)



<https://nvsyashwanth.github.io/machinelearningmaster/bias-variance/>

What means what?

- High variance high bias – chose a different model to predict
- High bias low variance – the overall variation in the training set must have been much smaller than in the testing set, run cross-folds
- Low bias high variance – the variance in the testing set might be higher than in the training set – change composition of test/train sets.
- Low bias low variance – this is what we ideally want



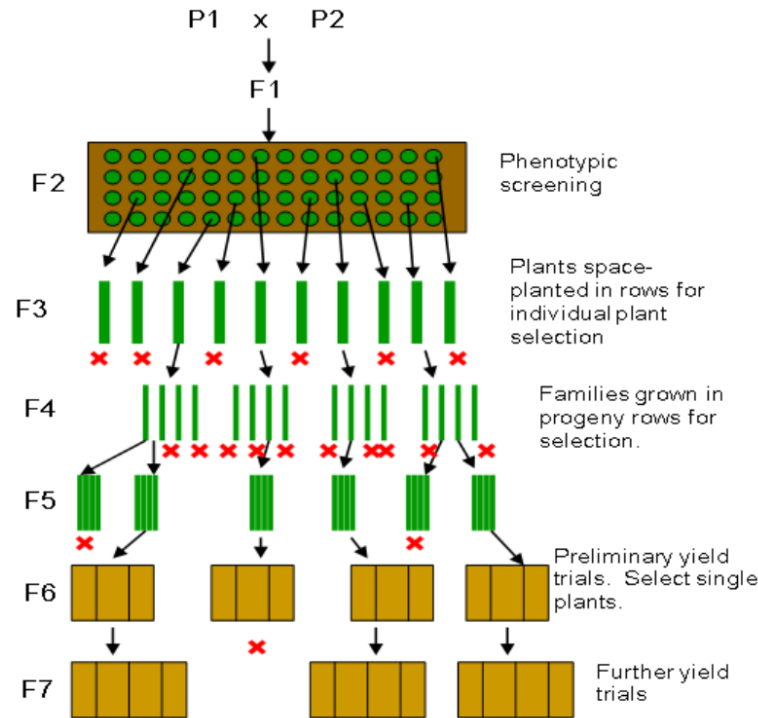
Summary

Steps running a genomic prediction

1. Measure phenotypes in multiple environments / years / experiments
2. Separating the genotypes in 10 or more groups for cross-fold validations
3. Perform the genomic prediction with any of the presented models to estimate the marker effect (a) or the breeding value (Za)
4. Using a with marker alleles (Z) or Za to predict genotype's pheno scores

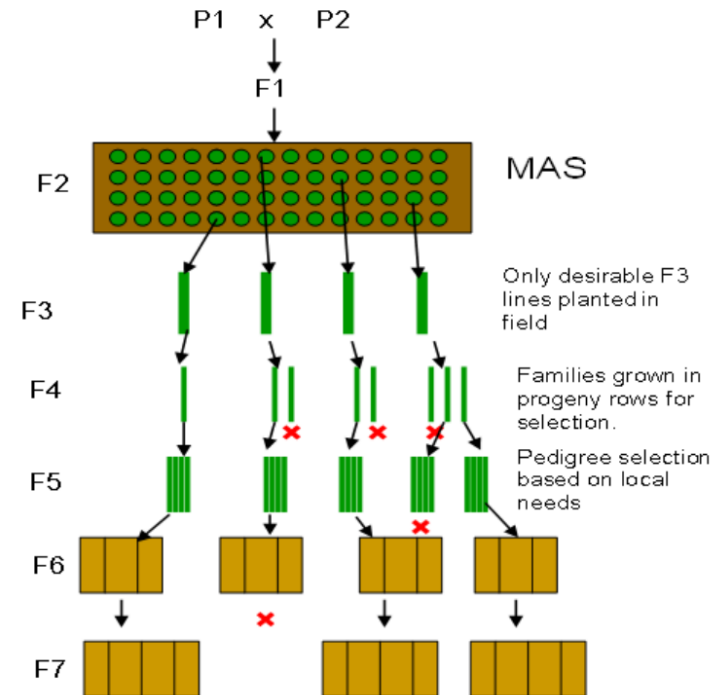
Using GP for marker assisted selection (MAS)

PEDIGREE METHOD



F8 – F12 Multi-location testing, licensing, seed increase and cultivar release

EARLY GENERATION SELECTION MARKER ASSISTED SELECTION



F8 – F12 Multi-location testing, licensing, seed increase and cultivar release

<http://dx.doi.org/10.19080/ARTOAJ.2018.14.555914>

Additonal – extracting the most useful markers for GP

Many markers explain very little to none of the phenotypes variation

- Reducing the complexitiy and costs by «selecting» markers

Different approaches exist

- **Top down**

- Using all markers as start set and reduce to a number of markers with an equal prediction accuracy
 - Advantage: no prior QTL knowledge requiered; User defined marker count & models
 - Disadvantage: untargeted approach with a lot of noise at the beginning

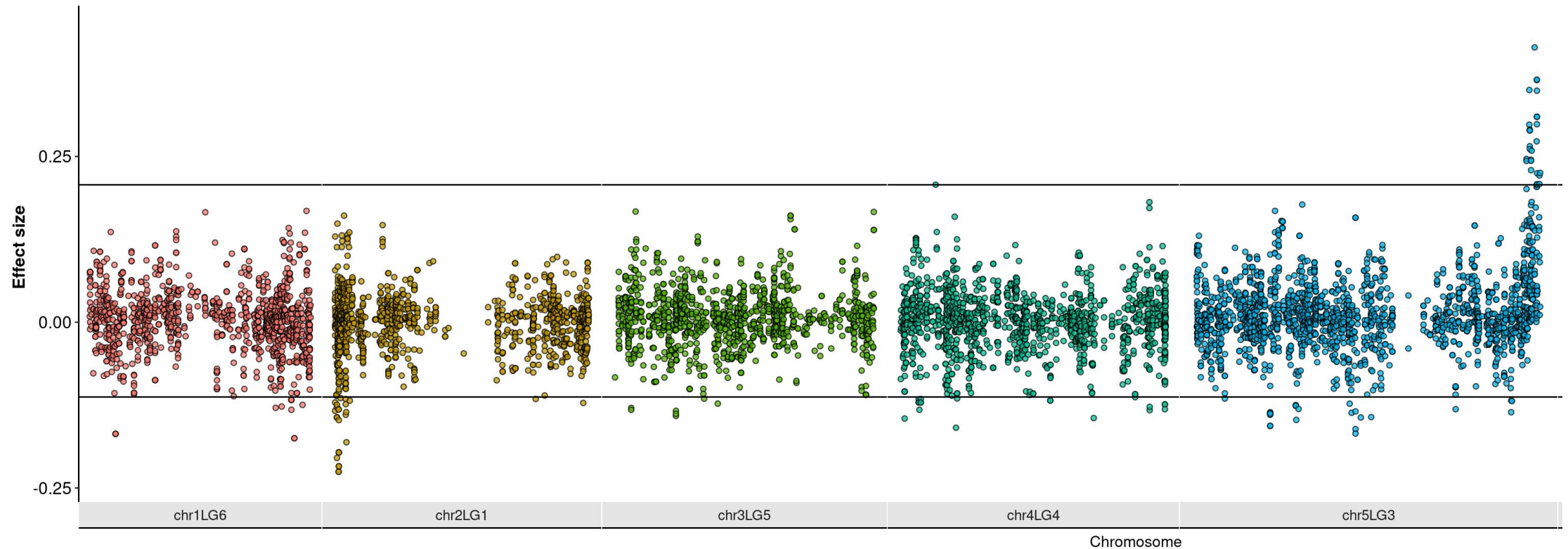
- **Down up**

- Starting with a set of QTLs and add markers until a sufficent precision is reached
 - [GMStool](#)
 - Advantages: Picks the absolute best marker combinations; very targeted
 - Disadvantages: prior knowledge of QTLs; no marker count threshold selection, models fixed

Additional – extracting the most useful markers for GP

Many markers explain very little to none of the phenotypes variation

- Reducing the complexity and costs by «selecting» markers



Additional Information

10.1534/genetics.112.147983

<http://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf>

http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=gs_lourenco.pdf

<https://www.nature.com/articles/ng.3097>

<https://www.nature.com/articles/s41598-020-76759-y#Abs1>

<https://github.com/JaeYoonKim72/GMStool>

<https://www.youtube.com/watch?v=UAj4TeAZ-AM>

<https://www.youtube.com/channel/UCtYLUtTgS3k1Fg4y5tAhLbw>

<https://www.nature.com/articles/s41598-020-76759-y#Abs1>

<https://github.com/JaeYoonKim72/GMStool>

2. Part



Get into the code – applied programming

Checking out different R packages

Quiz & Homework

- **To receive the participation certificate**
 - 1. Please perform the homework**
 - The tasks & data can be found here:
 - <https://github.com/mischn-dev/LiveSeeding-Training-T4.2/tree/homework>
 - 2. Send your results to michael.schneider@fibl.org**
 - 3. Today's documents can be found here:**
 - <https://github.com/mischn-dev/LiveSeeding-Training-T4.2/tree/main>



LiveSeeding

Thanks for your attentio

VI EUCARPIA CONFERENCE

on breeding to meet environmental &
societal challenges

26th TO 28th

MAY 2025
PORTUGAL

Instituto Politécnico
de Coimbra - ESAC



www.LiveSeeding.eu

key results, newsletter,
upcoming events, policy
briefs, videos, training,
practice abstract



Link to sister project on
organic fruit breeding
www.InnOBreed.eu