



INTRODUCTION TO GENOMIC SELECTION (GS) AND AN EXAMPLE FROM THE BREEDINGVALUE project.

Jahn Davik

Norwegian Institute for Bioeconomy Research
(NIBIO)

Crossing



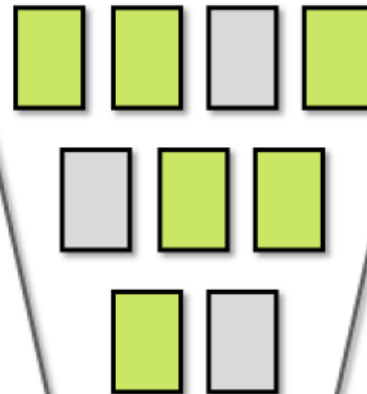
Seedling stage
unreplicated Seedlings



Clonal stage 1
unreplicated clones;
training population for
genomic selection



Clonal stage 2-5
replicated clones



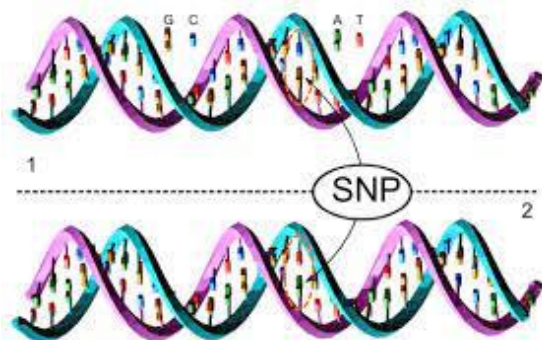
Variety release



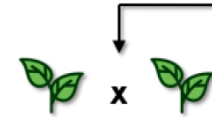
Conv

Pedigree breeding methods:

- **Traditional breeding** methods rely on pedigree, which has limitations due to:
 - Low accuracy in early generations
 - Long breeding cycles
 - High costs of phenotyping
- Advances in **molecular markers** (e.g., SNPs) and sequencing technologies have enabled high-throughput genotyping, making genomic selection feasible.



Crossing



Seedling stage
unreplicated Seedlings



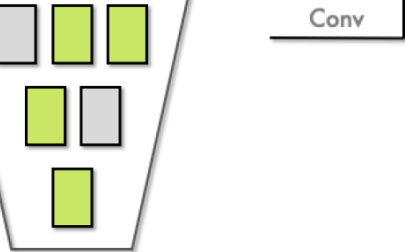
Clonal stage 1
unreplicated clones;
training population for
genomic selection



Clonal stage 2-5
replicated clones



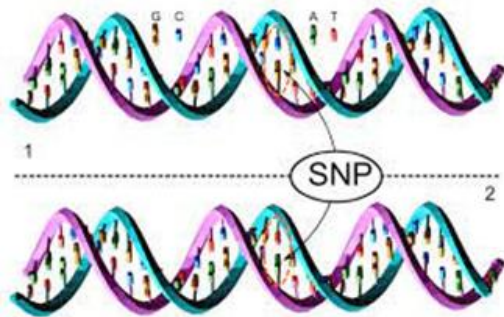
Variety release



What is Genomic Selection (GS)?

Genomic Selection (GS) is a **breeding method** that uses genome-wide molecular markers to **predict** the genetic potential of individuals for a given trait.

Unlike traditional selection methods, which rely on phenotypic evaluation or pedigree information, GS uses **genotypic data** and **advanced statistical models** to estimate **Genomic Estimated Breeding Values (GEBVs)**.





Breeding value = genetic merit of an individual

Breeding value components:

- ~ **50% due to parent average component**

This part of the breeding value comes from the additive genetic effects passed down from the parents. It is essentially the average of the parents' breeding values.

- ~ **50% due to Mendelian sampling component**

Sampling of parents' genes. It arises due to the random segregation and recombination of alleles during meiosis.

$$\text{Breeding Value} = \frac{BV_{\text{Sire}} + BV_{\text{Dam}}}{2} + \text{Mendelian Sampling}$$

Benefits of GS in Breeding Programs

GS provides multiple advantages over conventional breeding methods:

- Increases Selection Accuracy
- Reduces Breeding Cycle Time (L)
- Enhances Genetic Gain
- Cost-Effective
- Captures Mendelian Sampling Variance

$$\Delta G = \frac{\overset{\text{(Selection intensity)}}{i} \times \overset{\text{(Accuracy)}}{r_{TI}} \times \overset{\text{(Genetic variation)}}{\sigma_A}}{\underset{\text{(Generation interval)}}{L}} - dF \quad \text{(Inbreeding depression)}$$

Genetic gain is about the population not the individual.

Even if **you can't reliably pick the best individual every time**, if you are consistently selecting individuals that are on the average better than the rest, you can still move the population forward.

Think like this:

- If you are picking from a pool of thousands, and your predictions allow you to enrich the top 10% with ***somewhat better individuals***, the ***average of your selected group*** will be better than the overall population, even if some individuals in that group aren't the best.
- Repeating this over multiple generations accumulate **genetic gain** – small, consistent improvements over time.

Here is a key point of genomic selection: even if r is small, you can compensate with:

- High selection intensity (i.e., selecting from a larger group)
- Shortening generation interval (L) due to early-stage genotyping and genomic selection
- And by doing this every generation genetic gain keeps adding up

$$\Delta G = \frac{\overset{\text{(Selection intensity)}}{i} \times \overset{\text{(Accuracy)}}{r_{TI}} \times \overset{\text{(Genetic variation)}}{\sigma_A}}{L_{\text{(Generation interval)}}} - dF_{\text{(Inbreeding depression)}}$$

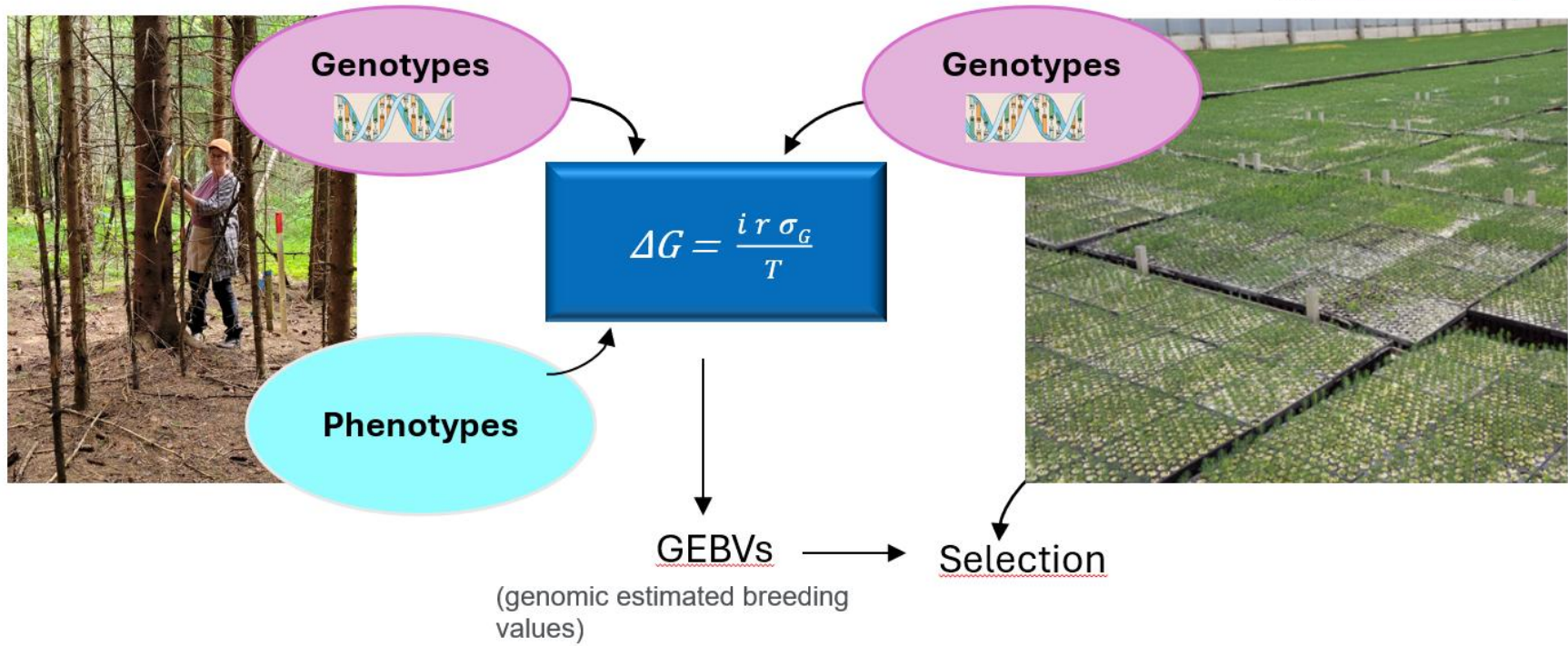
How is Genomic Selection Done?

- 1. Training Population Development:** A set of individuals with known genotypic and phenotypic data is used to build the predictive model.
- 2. Genotyping:** High-throughput sequencing or SNP arrays are used to determine genetic markers across the genome.
- 3. Model Training:** Statistical or machine learning models (e.g., GBLUP, Bayesian methods, machine learning approaches) are applied to associate genotypic data with phenotypic performance.
- 4. Prediction of GEBVs:** The trained model is used to estimate breeding values for new individuals without phenotypic evaluation.
- 5. Selection Decisions:** The best-performing individuals (based on GEBVs) are selected for further breeding, reducing reliance on time-consuming and expensive phenotyping.

Genomic selection overview

Training population

Breeding population



Genomic selection, the prediction part

$$y_{ijk} = \mu + \sum_{h=1}^5 PC_{hi} \gamma_h + b_j + g_i + ge_{ij} + \varepsilon_{ijk}$$

Note: Some of the phenotypes (y_{ijk}) can be missing!

Statistical Methods for Genomic Prediction

Model	Assumptions	Pros & Cons
GBLUP (Genomic Best Linear Unbiased Prediction)	Assumes equal effects of all markers (ridge regression)	Simple, robust, but may not capture large-effect QTLs well.
■ BayesA/B/C	Prior distributions on marker effects	Captures large-effect loci but computationally intensive.
■ BayesR	Mixture model with multiple effect sizes	More flexible but computationally demanding.
LASSO	Shrinks some marker effects to zero (feature selection)	Useful for sparse models but struggles with polygenic traits.
Random Forest & Machine Learning Methods	Non-parametric and capture interactions	Can handle complex traits but may overfit small datasets.



An example from the BreedingValue project:

Raspberry breeding material from Sant'Orsola

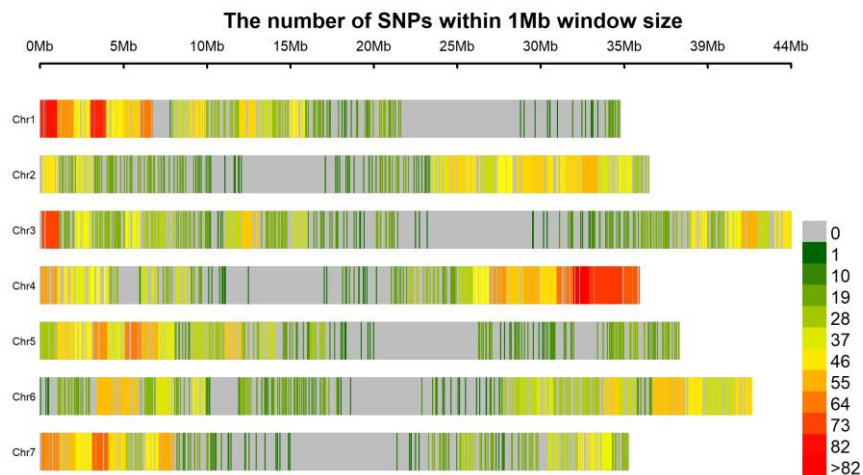


Pics.: Paolo Zucchi, Sant'Orsola

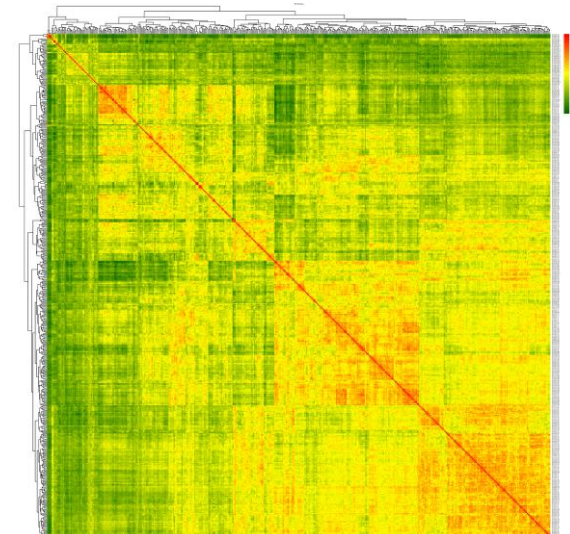


A raspberry genotyping assay was established in collaboration between JHI, NIAB, and NIBIO.

Distribution and density of RNAseq derived SNPs in two raspberry panels (NJOS & S'O).



We identified 7574 biallelic SNPs and 1149 biallelic indels in ~500 samples using targeted sequencing data.

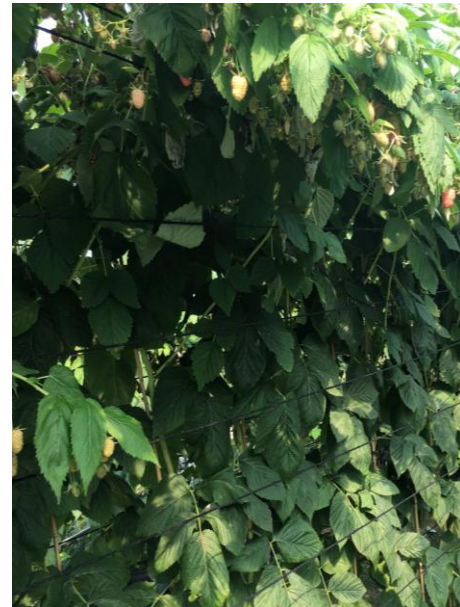


Analysed traits

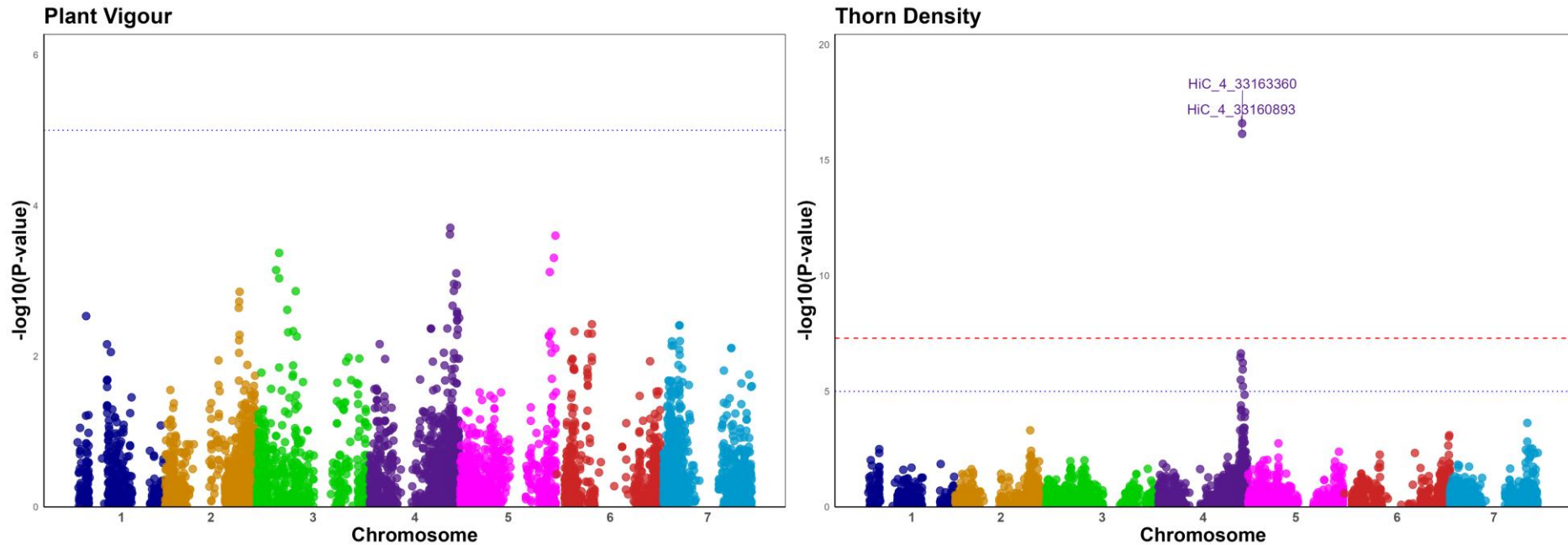
Thorne density:



Plant vigour:



Manhattan plots of analysed traits



Manhattan plots using the Sant'Orsola phenotype scores (plant vigour and thorne density BLUEs) over four experimental years and 5369 useful SNPs. The statistical method used was BLINK implemented in the R-based GAPIT. The blue dotted horizontal line indicates a suggestive significance ($-\log_{10}(p) = 5$) while the red dashed horizontal line indicates a genome-wide significance ($-\log_{10}(p) \approx 7.3$). Markers exceeding this latter threshold are named in the plots.

Genomic selection, the prediction part

$$y_{ijk} = \mu + \sum_{h=1}^5 PC_{hi} \gamma_h + b_j + g_i + ge_{ij} + \varepsilon_{ijk}$$

Note: Some of the phenotypes (y_{ijk}) can be missing!



Variance components by two Bayes estimation methods:

- Variance component estimates and confidence intervals derived from Bayes Ridge Regression (BayesRR) and BayesB (BB) models for plant vigour and thorn density.
- Phenotype scores were standardized to unit variance and hence the estimates can be interpreted as the proportion of the variance explained by each component.

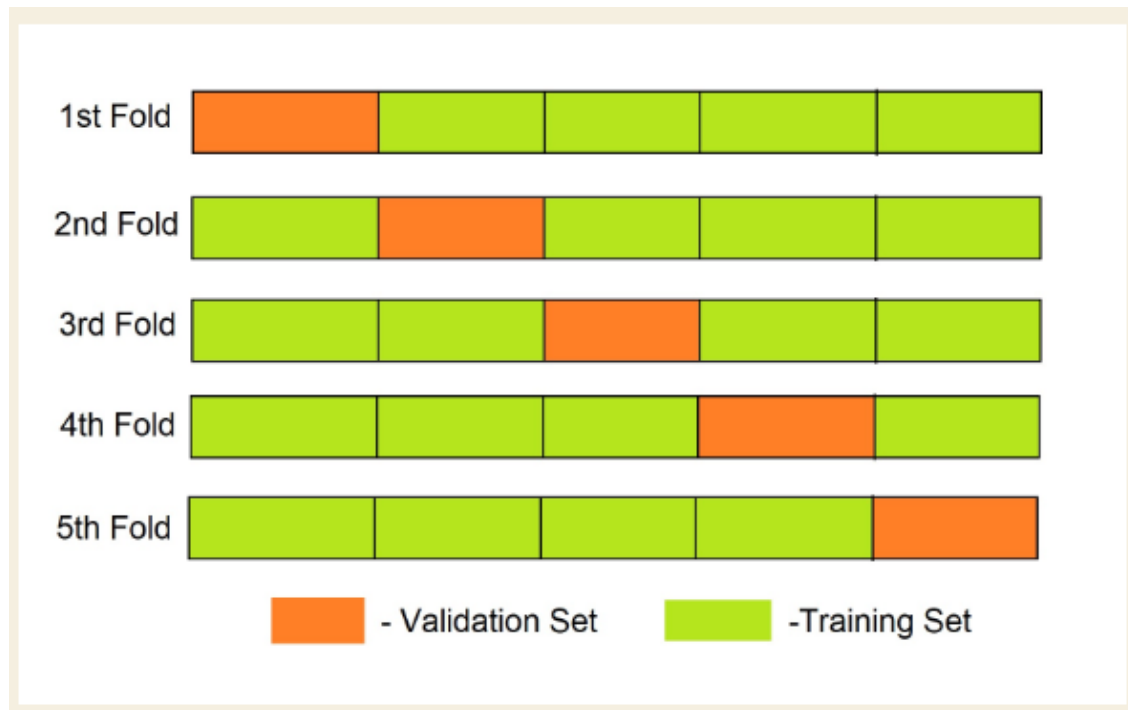
			Traits	
Method	Model	Genetic model	Plant vigour	Thorn density
<u>BayesB</u>	M1	Entry	0.28 (0.02)	0.46 (0.02)
	M2	Additive	0.49 (0.15)	0.85 (0.29)
	M3	Add + Dom	0.26 (0.08) + 0.16 (0.04)	0.65 (0.26) + 0.12 (0.03)
<u>BayesRR</u>	M1	Entry	0.28 (0.02)	0.47 (0.02)
	M2	Additive	0.46 (0.15)	0.76 (0.21)
	M3	Add + Dom	0.25 (0.08) + 0.16 (0.04)	0.66 (0.23) + 0.15 (0.04)

Genomic selection, the prediction part

$$y_{ijk} = \mu + \sum_{h=1}^5 PC_{hi} \gamma_h + b_j + g_i + ge_{ij} + \varepsilon_{ijk}$$

Note: Some of the phenotypes (y_{ijk}) can be missing!

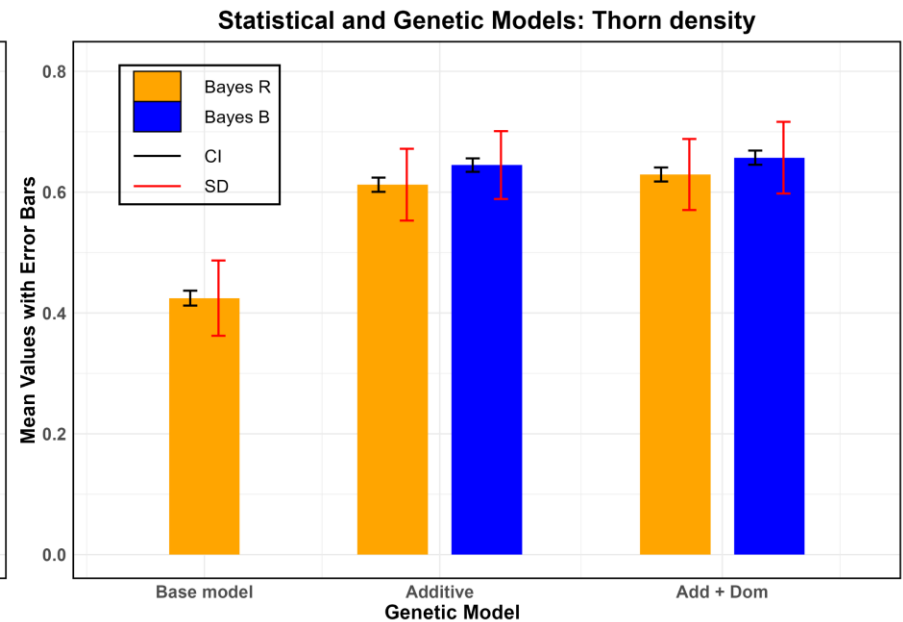
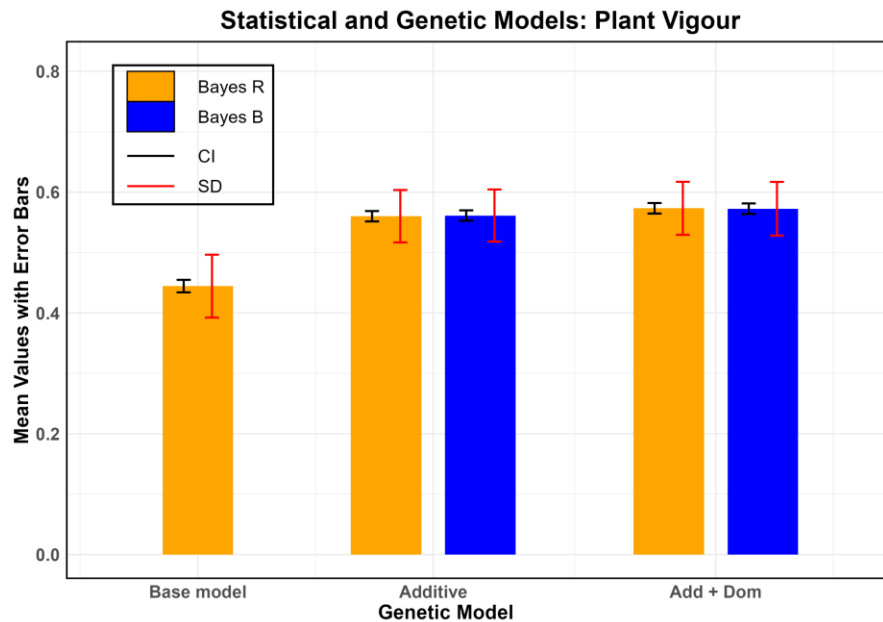
Cross-validation of the Sant'Orsola data set 2020 - 2023



driguez et al 2018



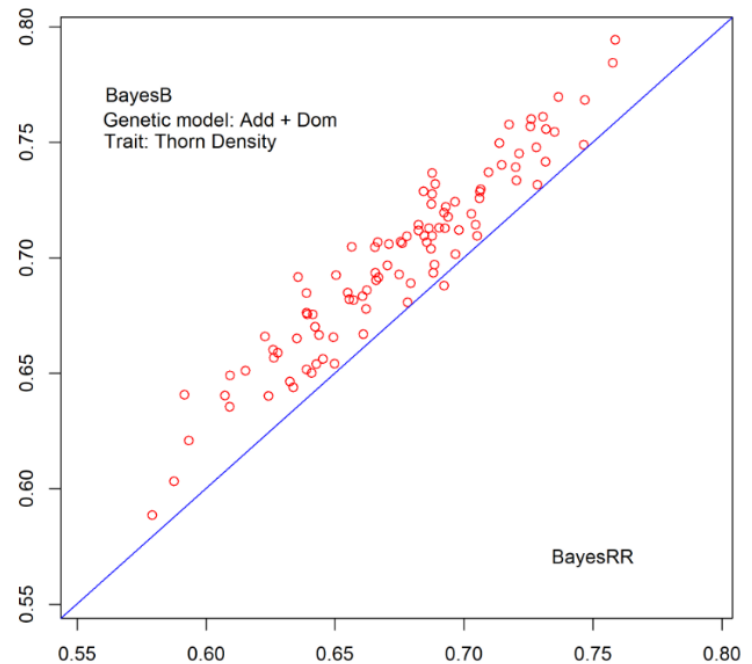
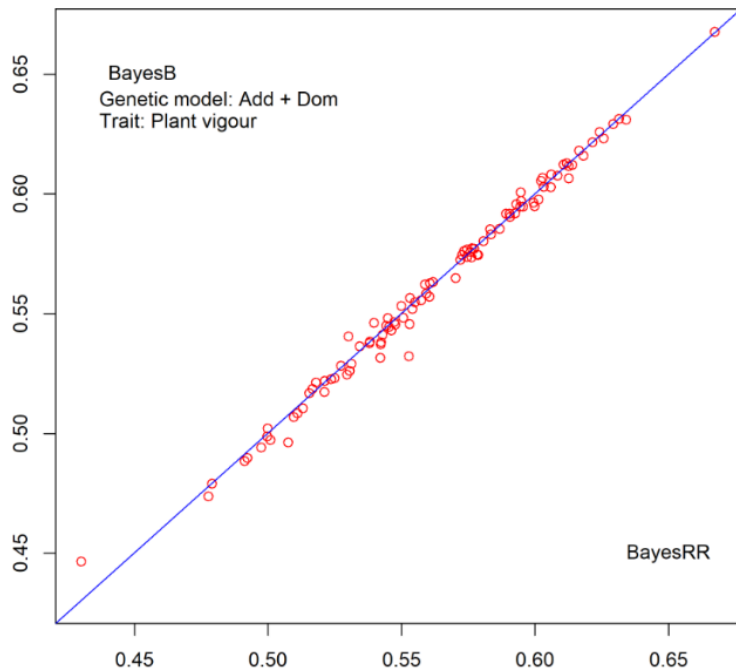
Prediction accuracies



- Prediction accuracies observed in fivefold cross-validation of plant vigour (left) and thorn density (right) in Sant'Orsola's raspberry breeding population.
- The prediction accuracies are quantified as the correlation between the actual score and the predicted score in 100 permutations of the phenotypic data set over four years of phenotypic evaluation.
- Red bars are standard deviations (SDs), and black bars are the confidence intervals (CIs).



Prediction accuracies



- Phenotype-predictions correlations from 100 training-testing partitions with plant vigour (left) and thorn density (right) using the additive + dominance allele specifications.
- For plant vigour the statistical method appears to have no impact on the prediction accuracies, while for thorn density the BayesB improved the prediction accuracies relative to Bayes Ridge Regression.

What does this increase in prediction accuracies imply?

Since the genetic gain (ΔG) is proportional to accuracy (r_{TI}), the relative increase in genetic gain should be:

$$\frac{\Delta G_{\text{with SNPs}}}{\Delta G_{\text{entries only}}} = \frac{0.6}{0.4} = 1.5$$

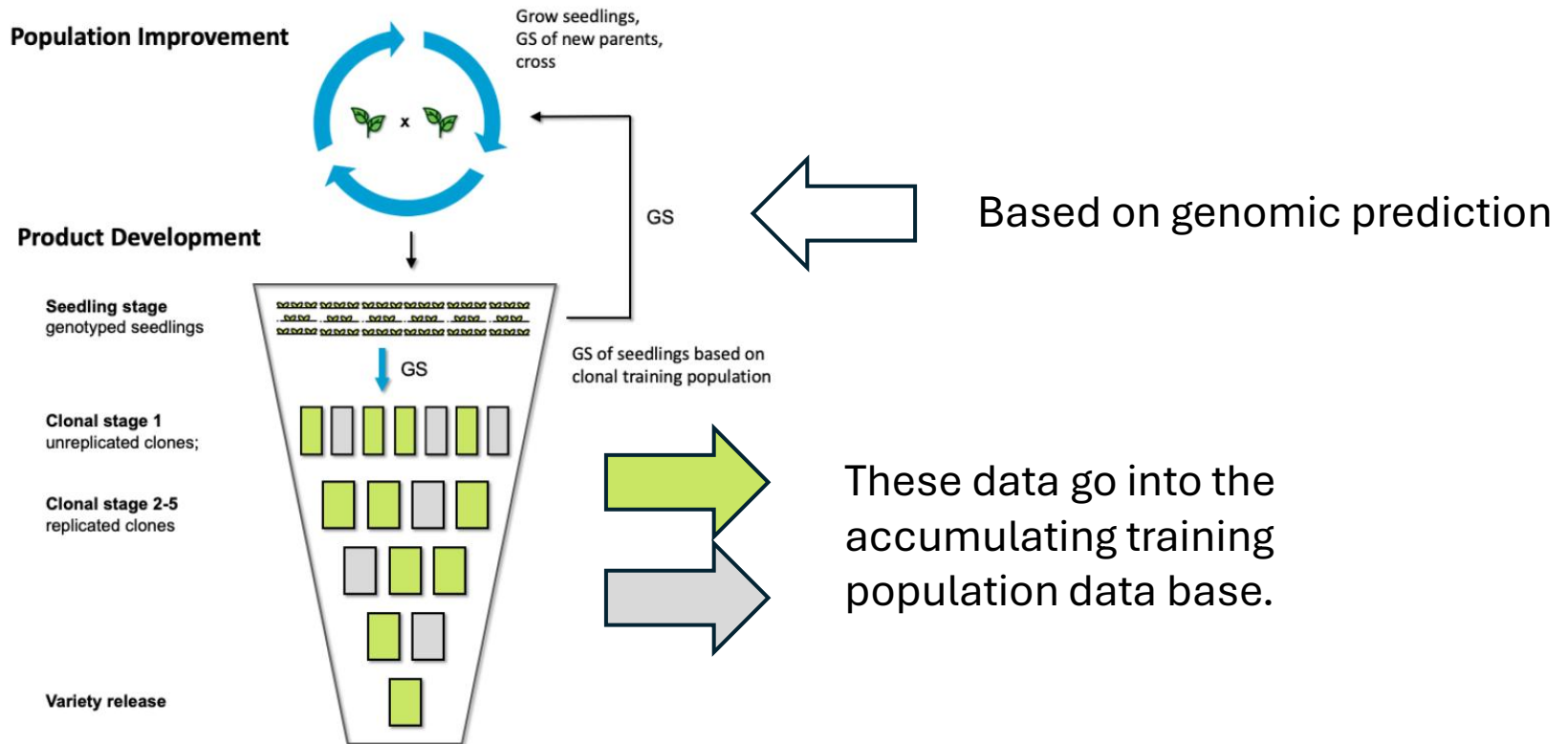
$$\Delta G = \frac{\overset{\text{(Selection intensity)}}{i} \times \overset{\text{(Accuracy)}}{r_{TI}} \times \overset{\text{(Genetic variation)}}{\sigma_A}}{\underset{\text{(Generation interval)}}{L}} - dF \underset{\text{(Inbreeding depression)}}{}$$

So, a 50% increase in genetic gain due to inclusion of SNP data is indicated.

In addition, genomic selection is particularly useful in reducing the generation (L) interval and selection intensity (i) which will further boost the genetic gain per time unit.



If I were a small fruit breeder - or any clonal species breeder:



Modified from Werner et al. 2023

Contributors:

The genotyping assay: NIAB, NIBIO, JHI
The phenotype data: Sant'Orsola
The sequencing: LGC (Germany)
SNP calling: NIBIO
Genomic predictions: NIBIO

Funding: The European Union's Horizon 2020 research and innovation program (grant agreement 101000747)



NIBIO

What does an r-value of 0.5 mean?

An $r = 0.5$ means there's a **moderate** correlation between the proxy (e.g. genomic estimated breeding value, GEBV) and the true trait (TBV). But here's the nuance:

- $r = 0.5$ implies that the **explained variance** is:

$$r^2 = 0.25$$

This means **25% of the variation** in the true breeding value is explained by your genomic prediction.

- Conversely, **75% of the variance remains unexplained** (due to noise, imperfect models, environment, etc.).

Fifty:fifty chance of success — what does that mean?

People sometimes say that with $r = 0.5$, there's a "50:50 chance" of selecting the better individual based on GEBVs. While not **mathematically precise**, this *rule of thumb* tries to reflect the **probability of correct ranking** between two individuals.

Let me explain:


- Suppose you have **two individuals**, and you want to choose the better one based on GEBV.
- If $r = 1$, GEBVs perfectly predict TBVs. You **always choose the better one**.
- If $r = 0$, GEBVs are just noise. Your **choice is random** — a 50% chance of picking the better one.
- At $r = 0.5$, you're **somewhere in between**: you'll pick the better one more often than not, but not reliably.

There's a statistical way to express this through the **probability of correct selection** P , which is:

$$P = \frac{1}{2} + \frac{\arcsin(r)}{2\pi}$$

Using $r = 0.5$:

$$P = \frac{1}{2} + \frac{\arcsin(0.5)}{2\pi} \approx 0.5 + \frac{30^\circ}{2\pi} \approx 0.5 + 0.083 = 0.583$$

So with $r = 0.5$, you have about a **58.3% chance** of  choosing the better individual — slightly better


What does this mean in practice?


- $r = 0.5$: your genomic predictions are *somewhat useful*, but not highly reliable.
- It **doesn't mean you'll succeed half the time**, but rather that your ability to **distinguish good from bad** is limited.
- In breeding programs, higher r values (like 0.7 or above) are typically more desirable **efficient selection**.

Genetic gain is about the **population**, not the individual

Even if **you can't reliably pick the best individual every time**, if you're consistently selecting individuals that are *on average* better than the rest, you can still move the population forward.

Think of it like this:

- If you're picking from a pool of thousands, and your predictions allow you to enrich the top 10% with *somewhat better* individuals, the **average of your selected group** will be better than the overall population, even if some individuals in that group aren't truly the best.
- Repeating this over multiple generations accumulates **genetic gain** — small, consistent improvements compound over time. 

 The breeder's equation gives us insight:

$$\Delta G = \frac{i \cdot r \cdot \sigma_A}{L}$$

Where:

- ΔG : genetic gain per year
- i : selection intensity
- r : accuracy of selection
- σ_A : additive genetic standard deviation
- L : generation interval

Here's the key part: even if r is **small**, you can compensate with:

- **High selection intensity** (i.e. selecting the best from a *very large group*),
- **Shorter generation intervals** (thanks to early genomic selection),
- And by doing this **every generation**, gain keeps adding up.

Here us a key point of genomic selection: even if r is small, you can compensate with:

- High selection intensity (i.e., selecting from a larger group)
- Shortening generation interval (L) due to early-stage genotyping and genomic selection
- And by doing this every generation genetic gain keeps adding up

Genomic Selection in Breeding Programs

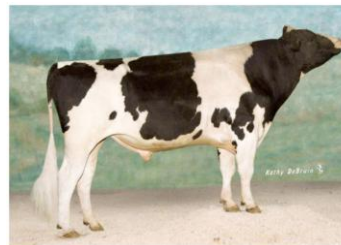
Genomic selection has been widely implemented in animal breeding and plant breeding (at least in major crops).

•Benefits:

- **Reduces generation interval** → Faster genetic gain.
- **Increases selection accuracy** → Uses thousands of markers instead of limited pedigree information.
- **Improves selection of hard-to-measure traits** (e.g., disease resistance, drought tolerance).

•Challenges:

- **Requires a large training population** to build accurate models.
- **Computational complexity** increases with marker density.
- **Decline in prediction accuracy across generations** due to recombination.



- **Genomic Prediction (GP)** refers to the use of genome-wide marker data to predict the genetic potential (breeding values) of individuals for complex traits.
- **Genomic Selection (GS)** is an application of GP in breeding programs, where individuals are selected based on genomic estimated breeding values (GEBVs) instead of phenotypic or pedigree-based selection.

$$\Delta G = \frac{\overset{\text{(Selection intensity)}}{i} \times \overset{\text{(Accuracy)}}{r_{TI}} \times \overset{\text{(Genetic variation)}}{\sigma_A}}{L \text{ (Generation interval)}} - dF \text{ (Inbreeding depression)}$$

Key **Concepts** in Genomic Prediction

3.1. Genetic Architecture of Traits

- **Quantitative traits** are controlled by multiple genes with small effects.
- **Marker effects** are estimated using a statistical model that links genotype with phenotype.

3.2. Genomic Estimated Breeding Values (GEBVs)

- The GEBV is the predicted genetic merit of an individual based on genome-wide markers.
- The accuracy of GEBVs depends on:
 - **Marker density**
 - **Training population size**
 - **Trait heritability**
 - **Relationship between training and target populations**



✓ BayesB

- **Assumption:** Only a **subset** of markers have non-zero effects; many have **no effect** at all.
 - **Modeling:** It uses a **spike-and-slab prior** — i.e., a proportion of markers (π) are assumed to have zero effect, and the rest have their effects drawn from a distribution (usually normal with marker-specific variance).
 - **When it's most suitable:**
 - Traits controlled by a **few major QTLs**
 - Traits with **significant SNPs/regions**
 - When you suspect **sparse architecture**
-

✓ BayesRR (Bayesian Ridge Regression)

- **Assumption:** **All markers** contribute small effects.
- **Modeling:** All marker effects are drawn from the same normal distribution with common variance.
- **When it's most suitable:**
 - **Polygenic traits**, controlled by many small-effect loci
 - Traits with **no clear major QTLs**